



VNIVERSITAT
DE VALÈNCIA

**Search for the Higgs boson produced in
association with a top quark using τ leptons with
ATLAS**

PhD Thesis

Pablo Martínez Agulló

Instituto de Física Corpuscular (IFIC)
Departament de Física Atòmica, Molecular i Nuclear
Programa de Doctorat en Física

Under the supervision of

**Carlos Escobar Ibáñez
and
Susana Cabrera Urbán**

València, February 2024

Carlos Escobar Ibáñez,
investigador Ramón y Cajal del Consejo Superior de Investigaciones Científicas
(CSIC), y

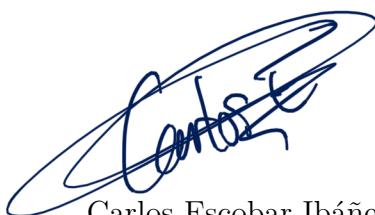
Susana Cabrera Urbán,
científica titular del Consejo Superior de Investigaciones Científicas (CSIC),

Certifican:

Que la presente memoria, **Search for the Higgs boson produced in association with a top quark using τ leptons with ATLAS** ha sido realizada bajo su dirección en el Instituto de Física Corpuscular, centro mixto de la Universitat de València y del CSIC, por **Pablo Martínez Agulló**, y constituye su Tesis para optar al título de Doctor por la Universitat de València una vez cursados los estudios en el Doctorado en Física.

Y para que así conste, en cumplimiento de la legislación vigente, presenta en el Departamento de Física Atómica, Molecular y Nuclear de la Universidad de Valencia la referida Tesis Doctoral, y firman el presente certificado.

Valencia, a 21 de diciembre de 2023,



Carlos Escobar Ibáñez



Susana Cabrera Urbán

*Ja que no s'entén el que dius,
que s'entenga la lletra.*

—ROSITA CANDELA-ESCLAPEZ

Preface

This thesis documents a physics analysis of crucial importance for the ATLAS (A Toroidal LHC ApparatuS) experiment at the Large Hadron Collider (LHC) at the CERN laboratory: the cross-section measurement of the production of the Higgs boson in association with a single top quark. This process involves the two most special elementary particles: the Higgs boson, which gives mass to all fundamental particles in the Standard Model (SM), and the top quark, which is the heaviest elementary particle and thus the one with the strongest coupling to the Higgs boson.

The coupling between the Higgs boson and the top quark is one of the most intriguing parameters of the SM. While its magnitude has already been measured, its sign is still unknown. The only way to determine its relative sign is through the production of a single top quark in association with a Higgs boson. This thesis aims to measure the rate of the tHq production targeting the final state with two light-flavoured charged-leptons and one hadronically-decaying τ -lepton (named $2\ell + 1\tau_{\text{had}}$ channel). The possible observation of an excess of signal events with respect to the SM prediction would be evidence of new physics in terms of CP-violating coupling and it would help to explain one of the current limitations of the SM, the matter–antimatter asymmetry.

The study presented uses 140 fb^{-1} of proton–proton collision data at a centre-of-mass energy of 13 TeV from the LHC Run 2 collected by the ATLAS detector. This search is exceptionally challenging due to the extremely small cross-section of the tHq process, $\mathcal{O}(10^2 \text{ fb})$, of which only 3.5% of events have a $2\ell + 1\tau_{\text{had}}$ final state. Additionally, the large presence of processes that can mimic the signal signature in the detector (known as backgrounds) makes it even more difficult to select a pure dataset for the analysis. Therefore, machine-learning techniques are used to distinguish the tHq signal events from the background as well as to distinguish the main background processes. Particularly, boosted-decision trees are employed to define the signal, control, and validation orthogonal regions of the phase space used in the analysis.

The analysis is further split into two categories depending on whether the light-flavoured charged-leptons have the same charge ($2\ell \text{SS} + 1\tau_{\text{had}}$) or opposite charge ($2\ell \text{OS} + 1\tau_{\text{had}}$). This split is motivated by the fact that the background composition is very different for the two cases, requiring basically two completely different data analyses. For the $2\ell \text{SS} + 1\tau_{\text{had}}$ sophisticated multivariate analysis is used for the light-lepton-origin assignment, which is to match to its parent which can be either the top quark or the Higgs boson.

The measured signal strength in the two sub-channels is found to be compatible with the predictions from the Standard Model: $\mu_{tHq}^{2\ell \text{OS} + 1\tau_{\text{had}}} = -22.1^{+29.2}_{-34.1}$ and $\mu_{tHq}^{2\ell \text{OS} + 1\tau_{\text{had}}} = -2.7 \pm 5.5$. These results are reproduced for the inverted Yukawa coupling hypothesis. Upper limits on the cross-sections are also reported. This result significantly contributes to getting closer to a statistical combination of all channels of the process tHq process being currently measured within the ATLAS experiment. This thesis constitutes an important step towards the first observation of this process, which will require more data (probably even more than what will be collected during LHC Run 3) and a better understanding of the main systematic uncertainties.

Chapter 1 introduces the fundamental concepts and scope of the SM, setting the stage for the in-depth discussion that follows in the subsequent chapters. The characteristics of the top quark and the Higgs boson are discussed in Chapter 2. In this chapter, the interaction between these two particles is described in detail, with special consideration to the scenario with CP-violating interaction. The experimental setup in which this work is contextualised is described in Chapter 3. The chapter provides a comprehensive overview of the LHC, the world's most powerful particle accelerator nowadays, and the ATLAS machine, its largest detector. The phenomenology of proton–proton collisions and generation of Monte Carlo simulations within the ATLAS detector are covered in Chapter 4. Chapter 5 focusses on the methodologies for the reconstruction and identification of physical objects. The strategies, tools and methods developed to set the best limits to the production cross-section of the tHq process in both the $2\ell \text{OS} + 1\tau_{\text{had}}$ and the $2\ell \text{SS} + 1\tau_{\text{had}}$ channels of this thesis are also presented in this chapter. Finally, the conclusion of the tHq search is presented in Chapter 7.

Acknowledgements

This work would not have been possible without the invaluable assistance of a large number of people whom I have been fortunate to meet. I would like to thank them all and dedicate this thesis to them.

First, thanks to my supervisors, Carlos Escobar and Susana Cabrera, who have always encouraged me and shared their wisdom. Both of you have supported and guided me through my master's thesis and subsequent PhD. I am very thankful for your efforts and help.

I also want to extend my gratitude to the rest of the ATLAS IFIC team for all their advice, help, and insightful suggestions. Carmen García, thank you for introducing me to this group. Salva Martí, thank you for your management. My thanks also go out to all the other researchers on the team. Special thanks to my fellow students; the mutual support we've shared is fundamental in overcoming the challenges of the PhD journey.

Thank you to the top-Higgs analysis group colleagues for your constructive ideas, brainstorming and discussions. Especially, I want to thank Valentina Vecchio, Carlos Escobar (again), Ian Brock and Nello Bruscino for organising this analysis and building its foundations. Also thank you for your collaboration on the tau-channel to Tanja Holm, Christian Kirfel and Florian Kirfel. We have worked together and shared the challenges of this analysis. Thank you as well to Oleh Kivernyk for your help with Section 6.5.

I want to express my esteem to all the friends who have accompanied me during this time. Sin salir del IFIC, agradecer a los pinches: Jose, Marcos, Pepe, Kevin y Stef. Juntos hemos compartido incontables momentos, tanto dentro como fuera del trabajo, brindándome el privilegio de convertir la jornada laboral en un encuentro fraternal. También gracias a Víctor y Mireia por amenizar las comidas. To Álvaro, Gustavo, Miguel, Mari Luz, Isabel, and all those whose kindness makes IFIC a warm and welcoming environment.

A heartfelt thanks to the friends who were with me at CERN. When I first arrived there, I knew almost no one, yet it didn't take long for Geneva and St. Genis to feel like my new home. Initially, Irene and Flavio took great care of me, for which I will always be grateful. At CERN, Carlos E. was always there to support me. During lunch, Esteban's humor brought me to tears of laughter, and at the coffee break, Miguel Ángel's endless anecdotes captured my interest. The companionship of Óscar, Sergi and, and especially Florencia, was crucial for me at that time.

On my second stay, I enjoyed the company of people like Carlos M., Elena, Lourdes, Dario, and others, who made my stay very pleasant. My old flatmate David B. was around, making it feel as if I had been living there forever. I will always remember my first day at the fitness club, when I convinced Jacopo of my extensive boxing experience — a claim far from the truth — to allow me to join the sparring sessions. This led to a memorable encounter with Chikuma, who swiftly taught me the reality of this game. Those afternoons at the club laid the foundation for a hobby that has since become an integral part of my life. There I met one of my dearest friends, David F., who has been supplying me with the highest-quality memes ever since. Grazas tamén aos meus dous irmáns galegos, Adrián C. e Alexandre, sodes xeniais. From that point onward, the rest of my visits to CERN were a cause for celebration because I would get to see you guys. I also extend my gratitude to those who made my third stay a delightful experience, with special thanks to Mariam and Olga. All of you transformed that cold city into one of my most cherished places in the world.

Beyond the confines of my professional sphere, there exists a constellation of individuals whose support and influence have been just as pivotal to my journey. It is with profound gratitude that I turn my attention to those figures. En primer lloc, vull expressar el meu més sincer agraiament a Cristina. Gràcies per la teua estima, per cada gest, pel teu suport incondicional, especialment durant els moments més difícils de la tesi. Has estat uns dels pilars fonamentals en aquesta etapa, i sempre t'estaré profundament agraiit. En gran mesura, monkey, aquesta tesi també és teua.

Continúo con el Mursiano, la C de SyC. Convivir estos años ha sido una de las mejores experiencias de mi vida. Un zulo en confinamiento puede ser un buen lugar si estás acompañado de alguien que hace que te duela la barriga de reír. Gracias por este tiempo.

Al resto de compañeros de Mistral, os quiero a todos. Dani, sé que es difícil, pero, por favor, lee el prospecto del Frenadol. María Betún, mardisión, mi yunta, eres una de las personas más bondadosas que he conocido, no cambies (o sí, no sé, yo tq). À Julien, mon ami, celui qui nous fournit toujours des croissants et crotolamo. Por supuesto, al Canario, a quien tanto echamos de menos, te recordamos. Estés donde estés, esto es para ti.

Mi agradecimiento a los Brandans es inmenso. Sé que nuestra complicidad no se extinguirá. A Álvaro, a pesar de que lleves mil años en la meseta, es como si nunca te hubieses ido. A Jose, crrrggg [gesto de boca de predator]. A Pablo, eres una de mis amistades más valiosas. Gracias por tu apoyo incondicional, por las infinitas risas y por todo el crecimiento compartido.

Gracias a los old brosters. Especialmente a Jesús y sus “¿Hasta cuándo te quedas?”. Contigo fue con quien decidí estudiar física y, mira, empezó como una broma pero aquí estamos. También gracias a los hermanos Mira. Manuel, infinitas gracias por acogerme tantas veces. Por vosotros, aunque lleve más de 12 años fuera, Elche sigue siendo mi casa.

Aún quedan mil personas a las que agradecer. Gracias Luna por tu amistad, por tantísimos desayunos de Obrador y tantas charlas, eres un sol. A Patri y su *ginga*. A Julis y Alice, personas bonitas que llegaron hace poquito y que se quedarán. Jose, te doy las gracias porque ya sea en Cambridge, Heidelberg, Valencia o Almería, eres una de las personas que más estimula mi cerebro y me incita a cuestionar todo lo que creo saber. A Chema y Adrián S., gracias. A Christoph. A Joel. A Julia. A Miryam. A Gema. Gracias también María Jesús y Marina, por haberme cuidado tanto.

Gracias a los colegas de Tarongers, sobre todo a Javi. A los compañeros del SFC, en particular a Ray. Las $\sim 10^7$ neuronas perdidas y estar cojeando la mitad de los días valen, sin duda, la pena. Sin el desahogo que habéis supuesto, no podría haber sido una persona estos años.

Finalment, gràcies a la meua família per haver confiat en mi i pel suport que he rebut de la vostra part. Yayo, gràcies per ensenyar-me el valor del treball i de l'esforç des de ben petit. Eres una figura a la qual sempre miraré amb admiració i estima. Yaya, eres la representació de l'amor incondicional, eres la meua llum predilecta, t'estime. Abuelito, el mejor chef que haya existido, me has enseñado que vivir es urgente. Abuelita, eres la personificación de la generosidad. Los recuerdos más dulces de mi vida son junto a vosotros, gracias, gràcies.

Gràcies a tots els Candela. A los primos. A Claudia, t'estime. A mi tanti, gracias por cuidarme desde siempre ¡Qué suerte tenerte! A ma mare. Gràcies pel teu suport. Mamà, sé que estàs orgullosa d'açò, però jo ho estic molt més de tu.

To each and every one of you, this thesis is as much yours as it is mine. Thank you. Gracias. Gràcies.

Pablo

Contents

Preface	v
Acknowledgements	vii
1 The Standard Model of particle physics	1
1.1 Introduction	2
1.2 Quantum electrodynamics	6
1.3 Electroweak interaction	7
1.3.1 Electroweak unification	8
1.4 Quantum chromodynamics	10
1.4.1 Quarks and colour	10
1.4.2 Gauge invariance for $SU(3)$	11
1.5 Particle masses	12
1.5.1 The Englert–Brout–Higgs mechanism	13
1.6 Remarks about the limitations of the Standard Model	17
1.6.1 The parameters of the Standard Model	17
1.6.2 Limitations of the Standard Model	17
2 Physics of the top quark and the Higgs boson	23
2.1 The top quark	24
2.1.1 Top-quark discovery	24
2.1.2 Top-quark production at LHC	25

2.1.3	Top-quark decay	29
2.1.4	Top-quark properties	30
2.2	The Higgs boson	33
2.2.1	Discovery of the Higgs boson	34
2.2.2	Higgs-boson production at LHC	34
2.2.3	Higgs-boson decay	36
2.3	The interplay between the top quark and the Higgs boson	39
2.3.1	CP properties in top-quark–Higgs-boson interactions	39
2.3.2	The $t\bar{t}H$ process	40
2.3.3	The tH process	42
3	The ATLAS experiment at the LHC of CERN	51
3.1	CERN	52
3.2	Large Hadron Collider	53
3.2.1	Machine design	53
3.2.2	Accelerator complex	55
3.2.3	LHC experiments	55
3.2.4	LHC computing grid	57
3.3	The ATLAS detector	59
3.3.1	Coordinate system	60
3.3.2	Inner Detector	60
3.3.3	Calorimeters	63
3.3.4	Muon Spectrometer	66
3.3.5	Magnet system	68
3.3.6	Trigger and Data Acquisition System	68
3.4	Performance of the ATLAS detector	69
3.4.1	Local coordinate frame and residuals	70
3.4.2	Track parameters and degrees of freedom	71

3.4.3	Track based alignment	71
3.4.4	ID alignment monitoring	75
4	Recording data and simulating events with the ATLAS detector	79
4.1	Phenomenology of proton–proton collisions	80
4.1.1	Proton structure and parton model	81
4.1.2	Underlying event	83
4.1.3	The pile-up effect	84
4.1.4	Luminosity	85
4.2	Recording data with the ATLAS detector	87
4.2.1	Cumulative luminosity	87
4.3	Simulating events within the ATLAS detector	88
4.3.1	Steps for simulating event data	91
4.3.2	MC generators	96
5	Object reconstruction and identification	99
5.1	Tracks and vertices	100
5.2	Charged leptons	102
5.2.1	Electrons	103
5.2.2	Muons	106
5.2.3	Hadronically decaying taus	108
5.3	Jets	110
5.3.1	Bottom-quark-induced jets	111
5.4	Missing transverse energy	114
5.5	Overlap removal	115
6	Search for the tHq production with a τ_{had} in the final state	117
6.1	Channels of the search of the tHq process	118
6.2	Data and simulated events	120

6.2.1	ATLAS-collected data samples	121
6.2.2	Simulated event samples	122
6.3	Object definition	128
6.3.1	Triggers	128
6.3.2	Electrons	129
6.3.3	Muons	129
6.3.4	Hadronically-decaying taus	130
6.3.5	Jets	130
6.3.6	Overlap removal	131
6.3.7	Missing transverse momentum	133
6.4	Study of the tHq signal process	133
6.4.1	Validation of parton-level simulations	133
6.4.2	Light-lepton-origin assignment	135
6.4.3	Reconstruction of the top quark and the Higgs boson	150
6.5	Background estimation	154
6.6	Event selection	159
6.6.1	Preselection	160
6.6.2	BDTs for region definition	161
6.6.3	Signal Region	174
6.6.4	Background dedicated regions	176
6.7	Systematic uncertainties	183
6.7.1	Symmetrisation of systematic uncertainties	184
6.7.2	Theoretical uncertainties	184
6.7.3	Experimental uncertainties	188
6.8	Fit results	191
6.8.1	Likelihood fit	192
6.8.2	Fit strategy	195
6.8.3	Binning optimisation and distributions for the fit	199

6.8.4	Asimov-hypothesis fit in the 2ℓ OS + $1\tau_{\text{had}}$	203
6.8.5	Asimov-hypothesis fit in the 2ℓ SS + $1\tau_{\text{had}}$	207
6.8.6	Full-data fit results in the 2ℓ OS + $1\tau_{\text{had}}$ channel	211
6.8.7	Full-data fit results in the 2ℓ SS + $1\tau_{\text{had}}$ channel	217
7	Conclusion	225
Appendices		231
A	Details about the parton-level simulations	231
A.1	Software package for parton-level information	231
A.2	BR-based calculation for the tHq fractions	232
B	Boosted Decision Trees	235
B.1	How does a BDT work?	235
B.1.1	Training	237
B.2	Feature-selection optimisation	241
B.3	Treatment of negative weights	242
B.3.1	Negative weights in $\text{BDT}^{\text{Lepton Assignment}}$	243
B.4	Cross validation and k -folding	244
B.5	Hyperparameters	245
B.5.1	Hyperparameter optimisation of a BDT with the genetic algorithm	249
B.6	Other considerations about BDTs	250
B.7	Additional plots and tables	252
B.7.1	$\text{BDT}^{\text{Lepton Assignment}}$	253
B.7.2	BDTs for region definition	253
B.7.3	Correlation between input features	253
B.7.4	Evolution of training	254

C Distribution of kinematic variables	259
C.1 BDT ^{Lepton Assignment}	259
C.2 BDT($tHq _{\text{OS}}$)	263
C.3 BDT($t\bar{t} _{\text{OS}}$)	273
C.4 BDT($tHq _{\text{SS}}$)	279
D Effect of negative weights	285
D.1 Negatively weighted events	285
D.2 Statistical uncertainty of negative weights	286
D.2.1 Errors in binned histograms	287
D.3 Negative weights in MVA methods	288
E Pruning of the non-impactful nuisance parameters	289
E.1 Asimov fit in the $2\ell \text{ OS} + 1\tau_{\text{had}}$ channel	289
E.2 Asimov fit in the $2\ell \text{ SS} + 1\tau_{\text{had}}$ channel	295
Bibliography	301
Resum de la tesi	325
1 Marc teòric	327
1.1 El Model Estàndard	327
1.2 La física del quark top	329
1.3 La física del bosó de Higgs	331
1.4 Interacció entre el quark top i el bosó de Higgs	332
2 L'experiment ATLAS de l'LHC al CERN	334
2.1 El detector ATLAS	335
2.2 Alineament del detector intern d'ATLAS	336
3 Recollida de dades, simulació i reconstrucció d'objectes	339
3.1 Simulació i adquisició d'esdeveniments en ATLAS	339

3.2	Identificació i reconstrucció d'objectes en ATLAS	341
4	Recerca de processos tHq amb un estat final $2\ell + 1\tau_{\text{had}}$	343
4.1	Estudis sobre la senyal	344
4.2	Estimació del fons	351
4.3	Selecció d'esdeveniments	352
4.4	Fonts d'incertesa	359
4.5	Resultats	361
5	Conclusions	370

Chapter 1

The Standard Model of particle physics

*La science, mon garçon, est faite d'erreurs,
mais d'erreurs qu'il est bon de commettre,
car elles mènent peu à peu à la vérité.*

—JULES VERNE,
VOYAGE AU CENTRE DE LA TERRE (1874)

Since the very first moment of our history, humankind has pursued the knowledge of nature and has tried to understand and describe how the universe works at a fundamental level. In fact, the word *physics* comes from the Greek word “φύσικη” which means “nature” [1, 2]. Most of the enquires regarding this, can be boiled down to two basic questions: what are the ultimate building blocks of reality? and which are the rules that govern them?

In the 7th century BCE, the pre-Socratic philosopher Thales of Miletus already proclaimed that every event had a natural cause [3]. Later, to understand how the basic components of the matter were formed, the ancient Indian philosophers such as Kanāda [4] in the 6th century BCE and Greeks Democritus and Leucippus [5] in the 5th century BCE developed the atomism, which comes from “ατομον” meaning uncuttable or indivisible.

From then to our days, the search for the minute fragments that comprise the matter and its interactions has led us to the Standard Model of particle physics (SM), one of the most successful scientific theories cultivated so far. This understanding of the universe can explain phenomena from the behaviour of atoms to how stars burn.

In this chapter, the basics of the SM are described by presenting in Section 1.1 the particles that compose it. The three fundamental forces are described in Sections 1.2, 1.3 and 1.4. Section 1.5 presents the mechanism that provides masses to the fundamental particles. Finally, the limitations of the SM are presented in Section 1.6.

1.1 Introduction

Based on the Quantum Field Theory (QFT), the SM of particle physics provides the theoretical framework that constitutes what is currently accepted as the best description of particle physics. It aims to explain both all particles of matter and their interactions. The completion of the SM was a collaborative effort of several scientists during the second half of the 20th century, being the current formulation finalised in the decade of 1970s. A representation of the fundamental particles that compose the SM is presented in Figure 1.1. As the scheme in Figure 1.1 indicates, the 12 fermions have their corresponding 12 anti-fermions and the quarks and gluons carry colour charge.

The SM is a gauge theory based on the symmetry group $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$, which describes all fundamental interactions except the gravitational force¹. This theory provides a description for the strong, weak and electromagnetic interactions via the exchange of the corresponding vector² bosons (spin-1 gauge fields). The mediation for the electromagnetic interaction (explained in Section 1.2) is done by one massless photon (γ). This force is invariant under the $U(1)$ symmetry group. While for the weak interaction, guided by $SU(2)$, three massive bosons, W^+ , W^- and Z , act as mediators (with masses $m_{W^\pm} = 80.385 \pm 0.015$ GeV [8] and $m_Z = 91.1876 \pm 0.0021$ GeV [9]). Although the electromagnetic and weak interactions seem completely different at low energies, they are two aspects of the same force and can be described simultaneously by the $SU(2)_L \otimes U(1)_Y$ symmetry group, which represents the so called electro-weak (EW) sector (detailed in Section 1.3). The strong force, with its eight massless

¹The gravitational interaction is described by Einstein’s General Relativity (GR) [7].

²“Vector bosons” refer to all particles that have spin 1 in contrast to the “scalar bosons” which have spin 0.

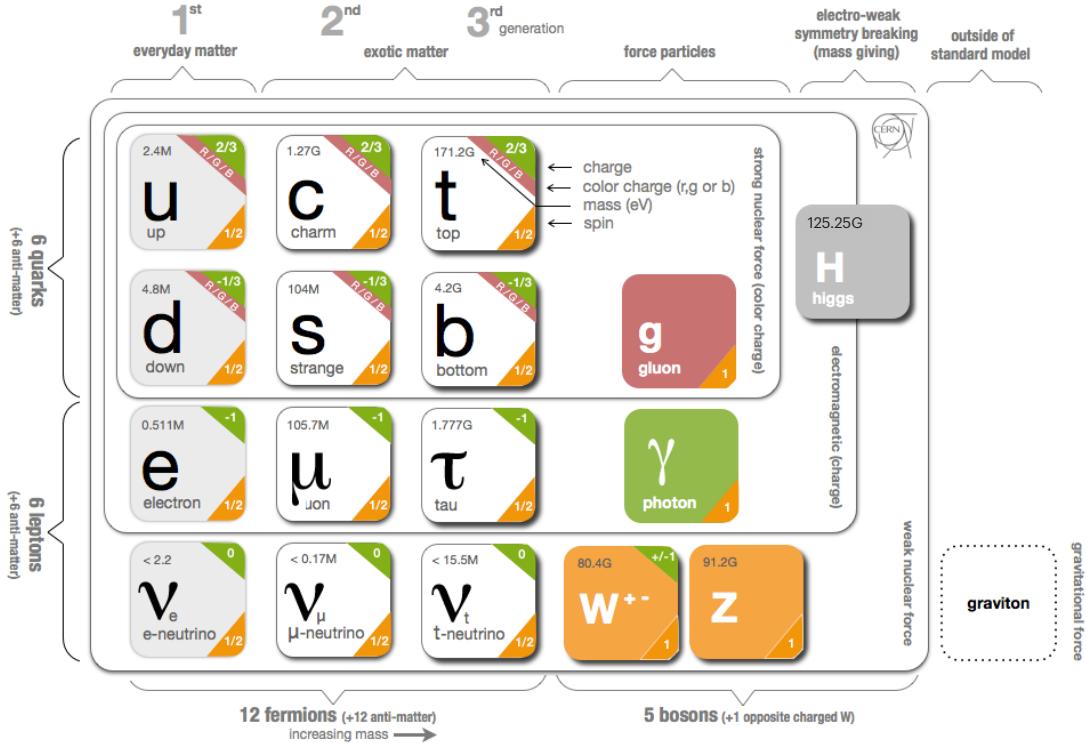


Figure 1.1: Fundamental particles of the SM (image taken and then modified from Reference [6]).

gluons (g), is described by the $SU(3)_C$ colour group (see Section 1.4). All these interactions differ in their magnitude, range and the physical phenomena that they describe. These features are summarised in Table 1.1, where not only the interactions described by the SM are included but the gravitation is shown as well for completeness.

Apart from the vector bosons, there is one massive scalar boson, the Higgs boson (with mass $m_H = 125.25 \pm 0.17$ GeV [10]). Through the interaction with this particle, all massive particles of Figure 1.1 gain their masses via the EW spontaneous symmetry breaking. This mechanism was first described by F. Englert, R. Brout [11] and P.W. Higgs [12], and it is summarised in Section 1.5.1.

Before describing the fundamental interactions of the SM in the QFT formalism, it is convenient to introduce the two main types of particles according to their spin, i.e. their intrinsic angular momentum: fermions and bosons.

Fermions

The fermions are the particles that follow the Fermi–Dirac statistics, i.e. obey the

Interaction	Theory	Mediator	Relative strength	Range (m)
Strong	QCD	g	1	10^{-15}
Electromagnetic	QED/EW	γ	1/137	∞
Weak	EW	W^\pm, Z	10^{-6}	10^{18}
Gravitational	GR	-	6×10^{-39}	∞

Table 1.1: Typical strength of the fundamental interactions with respect to the strong interaction. Here the strength is understood as the coupling constant or gauge coupling parameter. In GR the gravitational interaction is not a force but the effect of the four-dimensional spacetime curvature and, hence, it has no mediator in this formalism.

Pauli exclusion principle [13], resulting in a distribution of particles over energy levels in which two elements with the same quantum numbers cannot occupy the same states. The fermions include all particles with half-integer spin: quarks, leptons and baryons. A baryon is a non-fundamental particle composed of an odd number of valence quarks.

The fundamental fermionic matter is organised in the three families of leptons and quarks, shown also in Table 1.2:

$$\begin{bmatrix} \nu_e & u \\ e^- & d \end{bmatrix}, \begin{bmatrix} \nu_\mu & c \\ \mu^- & s \end{bmatrix}, \begin{bmatrix} \nu_\tau & t \\ \tau^- & b \end{bmatrix} .$$

These three generations, which are defined as the columns in Figure 1.1, exhibit the same kind of gauge interactions and they only differ in their mass [14]. According to the EW symmetry, each family can be classified as:

$$\begin{bmatrix} \nu_\ell & q_u \\ \ell^- & q_d \end{bmatrix} \equiv \begin{pmatrix} \nu_\ell \\ \ell^- \end{pmatrix}_L, \begin{pmatrix} q_u \\ q_d \end{pmatrix}_L, \ell^-_R, q_{uR}, q_{dR} .$$

(plus the corresponding antiparticles) where the subindices L and R stand from left- and right-handed particles, respectively. This structure responds to the fact that left-handed particles convert differently than right-handed ones under $SU(2)$ transformations. The left-handed fields are $SU(2)_L$ doublets and the right-handed ones $SU(2)_L$ singlets.

The fundamental representation of $SU(3)$ is a triplet, this is why each quark can appear in three different colours, whereas each antiquark can exhibit one of the corresponding “anticolours”.

The SM fermions properties are summarised in Table 1.2. The neutrino flavour states do not correspond to the mass states. What happens is that each flavour

Family	Name	Mass [MeV]	Charge
Quarks	Up (u)	$2.16^{+0.49}_{-0.26}$	2/3
	Down (d)	$4.67^{+0.48}_{-0.17}$	-1/3
	Charm (c)	$(1.27 \pm 0.02) \times 10^3$	2/3
	Strange (s)	93^{+11}_{-5}	-1/3
	Top (t)	$(172.76 \pm 0.30) \times 10^3$	2/3
	Bottom (b)	$(4.18^{+0.03}_{-0.02}) \times 10^3$	-1/3
Leptons	Electron (e^-)	$0.5109989461 \pm 0.0000000031$	-1
	Muon (μ)	$105.6583745 \pm 0.0000024$	-1
	Tau (τ)	776.86 ± 0.12	-1
	Electron neutrino (ν_e)	-	0
	Muon neutrino (ν_μ)	-	0
	Tau neutrino (ν_τ)	-	0

Table 1.2: Properties of the quarks and leptons. The electric is presented in units of elementary charge (1.602×10^{-19} C). The neutrino mass eigenstates, which have a fix mass, are different from the flavour eigenstates. Therefore, the mass of ν_e , ν_μ and ν_τ is not defined.

state is a quantum mechanical combination of neutrinos of different masses and vice versa. More details about the neutrino masses can be found in a dedicated text in Section 1.6.2.

Bosons

Bosons differ from fermions by obeying the Bose–Einstein statistics, thus, bosons are not limited to single occupancy for a determined state. In other words, the Pauli exclusion principle is not applied. All particles with integer spin are bosons; from the particles shown on the right columns of Figure 1.1 to the mesons. Mesons, along with baryons, are part of the hadron family, i.e. particles composed of quarks (see Section 1.4). The particularity of mesons is that they are formed from an equal number of quarks and antiquarks (usually one of each) bound together by strong interactions. Some examples of mesons are $\pi^{\pm,0}$, $K^{\pm,0}$ and J/ψ .

The elementary vector bosons are the force carriers and are presented in Table 1.1 while the Higgs boson is a fundamental particle as well.

1.2 Quantum electrodynamics

In QFT, particles are described as excitations of quantum fields that satisfy the corresponding mechanical field equations. The Lagrangians in QFT are used analogous to those of classical mechanics, where the equation of motion can be derived from the Lagrangian density function (\mathcal{L}) and the Euler–Lagrange equations for fields:

$$\frac{\partial \mathcal{L}}{\partial \phi} - \partial_\mu \frac{\partial \mathcal{L}}{\partial (\partial_\mu \phi)} = 0,$$

where $\partial_\mu = \frac{\partial}{\partial x^\mu}$ denotes the partial derivatives with respect to the four-vector x^μ and $\phi = \phi(\vec{x}, t)$ is the quantum field of a given fermion or boson. The Lagrangian is used to express the dynamics of the quantum field. In QFT, Noether's theorem [15] relates a symmetry in the \mathcal{L} to a conserved current.

The Dirac equation, $(i\gamma^\mu \partial_\mu - m)\Psi(x) = 0$, is one of the simplest relativistic field equations. Its Lagrangian describes a free Dirac fermion:

$$\mathcal{L}_0 = i\bar{\Psi}(x)\gamma^\mu \partial_\mu \Psi(x) - m\bar{\Psi}(x)\Psi(x), \quad (1.1)$$

being Ψ and $\bar{\Psi}$ the wave function of the particle and its hermitic conjugate, γ^μ are the Dirac matrices and m the rest-mass of the fermion. The first term of \mathcal{L}_0 is the kinetic term while the second is the mass term.

The gauge principle requires that the $U(1)$ phase invariance should hold locally. To satisfy this, the Lagrangian density for quantum electrodynamics (QED) can be defined by replacing the partial derivatives in \mathcal{L}_0 (equation 1.1) with the covariant derivatives:

$$\begin{aligned} \mathcal{L}_{\text{QED}} &\equiv i\bar{\Psi}(x)\gamma^\mu D_\mu \Psi(x) - m\bar{\Psi}(x)\Psi(x) \\ &= i\bar{\Psi}(x)\gamma^\mu [\partial_\mu + ieQA_\mu(x)]\Psi(x) - m\bar{\Psi}(x)\Psi(x) \\ &= \mathcal{L}_0 - eQA_\mu \bar{\Psi}(x)\gamma^\mu \Psi(x). \end{aligned} \quad (1.2)$$

The covariant derivative $D_\mu = \partial_\mu + ieQA_\mu(x)$ is defined this way to ensure gauge invariance. Here, A_μ is a gauge vector field that transforms like this: $A_\mu(x) \xrightarrow{U(1)} A'_\mu(x) \equiv A_\mu(x) + \frac{1}{e}\partial_\mu \theta$

Therefore, the Lagrangian in equation 1.2 is invariant under $U(1)$ local transformation. Along with the original Lagrangian in equation 1.1, the \mathcal{L}_{QED} has an additional term describing the interaction between the fermion Ψ and the gauge field A_μ with a strength proportional to the charge eQ . This term, $eQA_\mu \bar{\Psi}\gamma^\mu \Psi$, which has been generated only by demanding the gauge invariance under $U(1)$, is not other than the vertex of QED.

This new A_μ term is the electromagnetic field and its quanta is the photon. A mass term containing $A^\mu A_\mu$ is forbidden because it would violate the $U(1)$ local invariance. Consequently, the mediator of the new A_μ field, the photon, is predicted to be a massless particle. To make A_μ a propagating field it is necessary to add the kinetic term of the field A_μ :

$$\mathcal{L}_{\text{kin}} \equiv -\frac{1}{4}F_{\mu\nu}(x)F^{\mu\nu}(x), \quad (1.3)$$

where $F_{\mu\nu} \equiv \partial_\mu A_\nu - \partial_\nu A_\mu$. The kinetic term $F_{\mu\nu}F^{\mu\nu}$ is already invariant under local $U(1)$ phase transformations. The \mathcal{L}_{QED} with this kinetic term is written as:

$$\mathcal{L}_{\text{QED}} = \bar{\Psi}(x)(i\gamma^\mu\partial_\mu - m)\Psi(x) - eQ\bar{\Psi}(x)\gamma^\mu A_\mu\Psi(x) - \frac{1}{4}F_{\mu\nu}(x)F^{\mu\nu}(x). \quad (1.4)$$

1.3 Electroweak interaction

1.3.0.1 Weak interactions and symmetries

The weak interaction is mediated by the W^\pm and Z massive gauge bosons. The range of the interactions is within a scale of $\sim 10^{-18}$ m. It is responsible for radioactive decays and flavour-changing³ decays of fermions such as the decay of the muon ($\mu^- \rightarrow e^-\bar{\nu}_e\nu_\mu$).

Another particularity of this interaction is that it is the only interaction that violates several fundamental symmetries. The three discrete symmetries that are fundamental for the SM formulation and always hold for the electromagnetic and strong interactions but not for the weak interactions are:

- **Charge conjugation (C):** Replace positive quantum charges by negative charges and vice versa. It does not affect mass, energy, momentum or spin. Essentially, it is a transformation that switches all particles with their corresponding antiparticles.
- **Parity (P):** Parity involves a transformation that changes the algebraic sign of the spatial coordinate system. It does not reverse time, mass, energy or other scalar quantities.
- **Time reversal (T):** Consists in flipping the sign of the time.

³The lepton charge (also called lepton number) is conserved for every leptonic family.

The simultaneous combination of these three symmetries mentioned above results in the CPT symmetry, a profound symmetry of QFT which is consistent through all experimental observations [16]. If CPT symmetry is conserved, particles and their respective antiparticles are predicted to have, for example, the same mass and lifetime. Meanwhile, the P and C symmetries can be combined to create the CP symmetry, the product of the two transformations. The weak interaction violates P and C symmetries and the combined CP symmetry. Therefore, through the CPT theorem [17], if the CP is violated, T is violated as well to preserve the CPT invariance [16]. The CP violation plays a fundamental role in explaining the dominance of matter over antimatter in the present universe. The “direct” CP violation is a phenomenon where the same decay process has a different probability for a particle than for an antiparticle. The measurement of the tHq cross-section allows us to determine the possible existence of a CP-violating phase as it is further discussed in Section 2.3.3.

CKM matrix

The eigenstates that interact through weak interactions, known as “weak eigenstates” (d' , s' , u'), are different from the physically observed mass eigenstates (d , s , u). This makes possible the charged-flavour-changing-weak decays through the Cabibbo–Kobayashi–Maskawa (CKM) matrix. The CKM matrix, V_{CKM} , describes the mixing between the three generations of quarks in the SM. The coupling of two quarks i and j to a W boson is proportional to the CKM matrix element V_{ij} .

$$\begin{pmatrix} d' \\ s' \\ u' \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ u \end{pmatrix}$$

It is a 3×3 unitary matrix described by four independent parameters: three angles and one complex phase. These angles are known as the Euler angles and the phase allows the CP violation [18]. The largest values correspond to the diagonal elements of the matrix. This implies that the processes that do not change the flavour are strongly preferred over the family-changing charged currents.

1.3.1 Electroweak unification

At energies above the scale of the mass of the weak vector bosons ($E_{\text{EW}} \sim m_Z \sim m_W \sim 100 \text{ GeV}$), the electromagnetic and weak interactions are unified into the electroweak (EW) force. In other words, electromagnetism and weak interactions are simultaneously described by the symmetry group $SU(2)_L \otimes U(1)_Y$. The subindex L refers to left-handed fields and Y to the weak hypercharge, a quantum

number conserved under the strong interaction. In contrast, at low energies, these interactions are treated as independent phenomena, the electromagnetism is described by the QED and the weak interaction proposed by E. Fermi.

In the EW model (i.e. Glashow–Salam–Weinberg model), two new quantum numbers are assigned to the particles of the SM: the weak isospin (\vec{T}) and weak hypercharge Y' . Here, the left-handed chiral states of fermions form isospin doublets (χ_L) with $T_3 = \pm 1/2$ and the right-handed form chiral states are composed of isospin singlets (χ_R) with $T_3 = 0$. For a particle, T_3 is the third component of the \vec{T} , which is related to the electric charge (Q) and the $U(1)$ weak hypercharge by Gell-Mann–Nishijima relation:

$$Q = T_3 + \frac{1}{2}Y'. \quad (1.5)$$

Through this expression, the electromagnetic coupling and the electroweak couplings are connected. Having χ_L with $T_3 = \pm 1/2$ and χ_R with $T_3 = 0$ implies that a $SU(2)$ weak interaction can rotate left-handed particles (i.e. convert a left-handed e^- into a left-handed ν_e emitting a W^-) but cannot do the same with right-handed.

Using the gauge invariance principle it is possible to find the QED Lagrangian:

$$\begin{aligned} \mathcal{L} &= i \sum_{j=1}^3 \bar{\Psi}(x) \gamma^\mu \partial_\mu \Psi(x) \\ &= i \sum_{j=1}^3 \bar{\chi}_L(x) \gamma^\mu \partial_\mu \chi_L(x) + i \sum_{k=1}^3 \bar{\chi}_R(x) \gamma^\mu \partial_\mu \chi_R(x) \end{aligned} \quad (1.6)$$

where the wave function Ψ has been split into the left isospin doublets χ_L and right isospin singlets χ_R . The indices j and k run over the three generations of the SM.

To ensure gauge invariance under $SU(2)_L \times U(1)_Y$, four different gauge fields have to be added. While three week-isospin currents couple to the triplet of vector bosons W_μ^n with $n \in \{1, 2, 3\}$, the weak hypercharge couples to an isosinglet B_μ . The fields W_μ^1 and W_μ^2 are electrically charged whereas W_μ^3 and B_μ are neutral fields. The EW covariant derivative is defined as:

$$D^\mu \chi_{L_j}(x) = [\partial_\mu - ig \frac{\tau_i}{2} W_\mu^i(x) - ig' \frac{y_j}{2} B_\mu(x)] \chi_{L_j}(x) \quad i \in [1, 2, 3] \quad (1.7)$$

$$D^\mu \chi_{R_j}(x) = [\partial_\mu - ig' \frac{y_j}{2} B_\mu(x)] \chi_{R_j}(x), \quad (1.8)$$

where g and g' are the interaction couplings to W_μ^i isotriplet and the B_μ isosinglet. Finally, if kinetic terms for the gauge bosons are included, the EW SM Lagrangian

is obtained:

$$\begin{aligned}\mathcal{L}_{\text{EW}} = & i \sum_{j=1}^3 \bar{\chi}_L^j(x) \gamma^\mu D_\mu \chi_L^j(x) + i \sum_{k=1}^3 \bar{\chi}_R^k(x) \gamma^\mu D_\mu \chi_R^k(x) \\ & - \frac{1}{4} W_{\mu\nu}^n(x) W_n^{\mu\nu}(x) - \frac{1}{4} B_{\mu\nu}(x) B^{\mu\nu}(x),\end{aligned}\quad (1.9)$$

where the addition of kinetic terms gives rise to cubic and quadratic self-interactions among the gauge fields.

The \mathcal{L}_{EW} in equation 1.9 can be divided in two different parts according to the charge of the bosons: charged currents and neutral currents. Relating the charged currents (W_μ^1 and W_μ^2) to the W^+ and W^- bosons of the SM and the neutral (W_μ^3 and B_μ) ones with the Z and γ , it is possible to build linear combinations of the original gauge fields that define the SM bosons.

There is no mass term for the bosons in the EW Lagrangian that has been obtained in equation 1.9 by demanding the $SU(2)_L \times U(1)_Y$ local invariance, which enters in contradiction with the experimental observations for the W and Z bosons that indicate that $m_{Z,W}$ is $\mathcal{O}(100)$ GeV. The introduction of such a mass term would break the symmetry, however, it is possible to add the mass for the W and Z bosons without losing the properties of the symmetry. The method to do so is known as the Englert–Brout–Higgs mechanism or, more commonly, just the Higgs mechanism. This mechanism is described in Section 1.5.

1.4 Quantum chromodynamics

The QCD theory is a QFT-based model for describing the strong interactions between quarks and gluons (partons). This type of interaction is responsible of the nuclear force, the one that acts between the protons and neutrons of atoms binding them together.

1.4.1 Quarks and colour

The QCD theory is based on the $SU(3)$ symmetry group and its name derives from the “colour” charge, an analogous to the electric charge of QED but for strong interactions. The colour charge was introduced in 1964 [19] to explain how quarks could coexist within some hadrons apparently having the same quantum state without violating the Pauli exclusion principle. To satisfy the Fermi–Dirac statistics it is necessary to add an additional quantum number, the colour, to the

theory. Each species of quark (q) may have three different colours (q^α , $\alpha = 1, 2, 3$): red, green, and blue. Baryons and mesons are described then by colour singlet combinations.

Additionally, it is postulated that all hadrons must have a global neutral colour charge, i.e. the hadrons must be “colourless”. This assumption is known as the confinement hypothesis and it is made to avoid the existence of non-observed extra states with non-zero colour. It is called colour confinement because it implies that it is not possible to observe free quarks since they carry colour charge and, hence, they have to be confined within colour-singlet combinations.

1.4.2 Gauge invariance for $SU(3)$

The dynamics of the quarks and gluons are controlled by the QCD Lagrangian. The starting Lagrangian density is

$$\mathcal{L}_0 = \sum_f \bar{q}_f^\alpha (i\gamma^\mu \partial_\mu - m_f) q_f^\alpha, \quad (1.10)$$

where q_f^α denotes a quark field of colour α and flavour f . The mass of q_f^α is m_f .

The Lagrangian in equation 1.10 has to satisfy invariance under local transformations and hence its derivatives have to be substituted by covariant objects. The $SU(N)$ with $N = 3$ algebra demands that there are eight independent gauge parameters and, hence eight different gauge bosons $G_a^\mu(x)$ are needed. With g_s being the QCD coupling, the covariant derivatives are:

$$D^\mu q_f^\alpha \equiv \left[\partial_\mu + ig_s \frac{\lambda^a}{2} G_a^\mu(x) \right] q_f^\alpha \equiv [\partial_\mu + ig_s G^\mu(x)] q_f^\alpha.$$

In contrast to the case of QED, the non-commutativity of the $SU(3)_C$ matrices give rise to an additional term involving the gluon fields themselves: $-f^{abc}\delta\theta_b G_c^\mu$, with f^{abc} being the structure constant. With this, it is necessary to introduce the corresponding fields strengths to build a gauge-invariant kinetic terms for the gluon fields:

$$G_a^{\mu\nu} \equiv \partial_\mu G_a^\nu - \partial_\nu G_a^\mu - g_s f^{abc} G_b^\mu G_c^\nu.$$

Normalising the gluon kinetic term, the $SU(3)_C$ invariant QCD Lagrangian is obtained:

$$\mathcal{L}_{\text{QCD}} \equiv -\frac{1}{4} G_a^{\mu\nu} G_{\mu\nu}^a + \sum_f \bar{q}_f^\alpha (i\gamma^\mu D_\mu - m_f) q_f^\alpha. \quad (1.11)$$

Note how the gluon-gluon vertex is found by demanding the gauge invariance under local $SU(3)_C$ transformation. A mass term is forbidden for the gluon fields

by the $SU(3)_C$ gauge symmetry because a term of the form $\frac{1}{2}m_G^2 G_a^\mu G_\mu^a$ would not be invariant. The gluons are, then, predicted by the theory to be spin-1 massless particles.

Thanks to the colour symmetry properties, this Lagrangian looks very simple and all its interactions depend on the strong coupling constant, g_s . In contrast to the Lagrangian derived for QED (see equation 1.2), in \mathcal{L}_{QCD} the boson field has a self-interacting term. These gluon self-interactions give rise to the triple and quadratic gluon vertex. The self-interactions among the gluon fields can explain features such as the asymptotic freedom and confinement. The asymptotic freedom causes interactions between particles to become asymptotically weaker as the energy scale increases and the corresponding length scale decreases. The confinement implies that the strong forces increase with the distance, therefore, as two colour charges are separated, at some point it becomes energetically favourable for a new quark-antiquark pair to appear rather than to keep getting further. These new quarks bond with the previous two, preventing single quarks from being isolated. This mechanism, depicted in Figure 1.2, explains graphically why the strong interaction is responsible for keeping the quarks together forming hadrons.

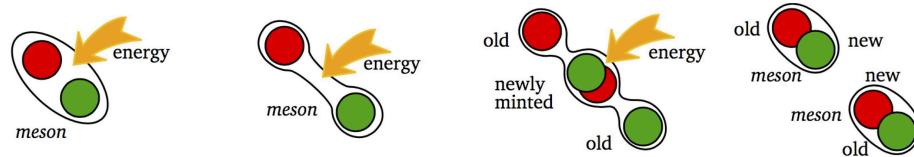


Figure 1.2: The QCD colour confinement explains the inseparability of quarks inside a hadron despite of investing ever more energy. In this example, the mechanism is shown for a meson.

1.5 Particle masses

For the QED Lagrangian, \mathcal{L}_{QED} (see equation 1.4), it is clear how the mass of the photon must be zero in order to satisfy the $U(1)$ local gauge symmetry because, if a mass term for the vector gauge field A_μ is included, the \mathcal{L}_{QED} would be non-invariant.

The Lagrangian of QCD as presented in equation 1.11 also demonstrates a similar restriction: the mass term for the gluon fields is forbidden by the $SU(3)_C$ gauge symmetry. Therefore, the mediating bosons for the strong interactions are massless as well (experimentally, a mass as large as the upper limits of a few MeV have been set, see Reference [20]).

While the prohibition of mass terms for the bosons of QED and QCD is not a problem, this requirement also applies to the $SU(2)_L$. This condition enters into open contradiction with the measurements of large masses for the W and Z bosons of weak interactions.

For weak interactions, the problem of massless particles does not only affect the bosons. Since under the $SU(2)_L$ left-handed particles transform as weak isospin doubles and right-handed particles as isospin singlets, the mass term of a spinor field Ψ written as chiral states also breaks the required gauge invariance.

The Englert–Brout–Higgs mechanism [11, 12, 21] describes how both the W and Z bosons and the fermions acquire mass without breaking the local gauge symmetry of the SM.

1.5.1 The Englert–Brout–Higgs mechanism

The Higgs mechanism in the SM for bosons

The Goldstone theorem states that “for a continuous symmetry group \mathcal{G} spontaneously broken down to a subgroup \mathcal{H} , the number of broken generators is equal to the number of massless scalars that appear in the theory” [22]. To apply this mechanism to the SM, it is necessary to generate mass for the W^+ , W^- and Z bosons while keeping the photon massless. To do so, the EW symmetry group $SU(2)_L \times U(1)_Y$ has to be broken into a $U(1)$ subgroup describing electromagnetism. A gauge-invariant interaction that gives masses to fermions without mixing chiral components is introduced by defining a $SU(2)$ isospin doublet of complex scalar field with hypercharge $Y = 1$:

$$\Phi = \begin{pmatrix} \phi^+ \\ \phi_0 \end{pmatrix} .$$

Being ϕ^+ positively charged and ϕ^0 neutral. The Lagrangian $\mathcal{L}_{\text{Higgs}}$ has to be added to the \mathcal{L}_{EW} in equation 1.9.

$$\mathcal{L}_{\text{Higgs}} = (D_\mu \Phi)^\dagger (D^\mu \Phi) - V(\Phi) \text{ where } V(\Phi) = \mu^2 \Phi^\dagger \Phi + \lambda (\Phi^\dagger \Phi)^2 ,$$

with $\lambda > 0$ required for vacuum stability. When $\mu^2 > 0$, the minimum of the potential occurs when both fields (ϕ^+ and ϕ^0) are at zero. If $\mu^2 < 0$, the minimum of the potential has an infinite number of degenerate states that satisfy $\Phi^\dagger \Phi = \mu^2 / 2\lambda$ and the physical vacuum state will correspond to any particular point on the circle of Figure 1.3. Having to choose a particular point breaks the global $U(1)$ symmetry of the Lagrangian. Without loss of generality, in this scenario, the ground state Φ_0

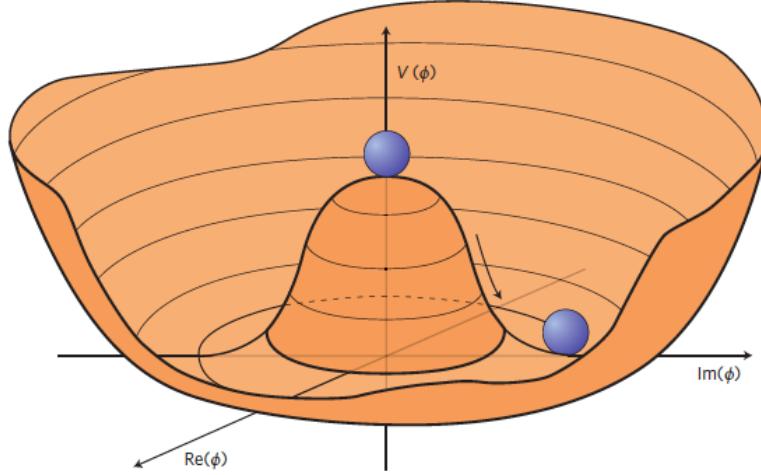


Figure 1.3: An illustration of the Higgs potential $V(\Phi)$ in the case of $\mu^2 < 0$ [23]. Choosing any particular point in the circle defined by v spontaneously breaks the $U(1)$ rotational symmetry. This type of potential is frequently called “Mexican hat”.

can be chosen to be:

$$\Phi_0 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix} \text{ where } v = 2\sqrt{\frac{\mu^2}{\lambda}}.$$

being v the vacuum expectation value. This defines the already mentioned circle in the minimum of $V(\Phi)$ in the $\mu^2 < 0$ scenario.

The Lagrangian density must be formulated in terms of deviations from one of these ground states. This can be done by introducing an excitation, $h(x)$, that can be understood as a small deviation of the field from the ground state. Accordingly, the fields can be expanded around the minimum as:

$$\Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}, \quad (1.12)$$

and the covariant derivative takes the form:

$$(D_\mu \Phi)^\dagger (D^\mu \Phi) = \frac{1}{2} + \frac{g^2 v^2}{4} W_\mu^+ W^{-\mu} + \frac{g^2 v^2}{8 \cos^2 \theta_W} Z_\mu Z^\mu,$$

where θ_W is the weak mixing angle, which is measured to be $\sin^2 \theta_W = 0.2310 \pm 0.0005$ [24]. By doing this, the W^+ , W^- and Z bosons have finally acquired mass. Through the Higgs mechanism, their masses within the SM are:

$$M_W = \frac{1}{2} g v \quad \text{and} \quad M_Z = \frac{1}{2} \frac{g v}{\cos \theta_W}.$$

Additionally, a new scalar field $h(x)$ has appeared with its correspondent mass term, the Higgs field. Note that the $h(x)$ was introduced as a perturbation from the ground state of the Higgs potential $V(\Phi)$, so the Higgs boson can be understood as an excitation of the Higgs potential. Apart from couplings to the electroweak gauge fields, the Higgs field also has self-interaction vertices. The mass of this boson is $m_H = \sqrt{2}\mu$.

With this covariant term, the Higgs Lagrangian density of the system is obtained:

$$\begin{aligned} \mathcal{L}_{\text{Higgs}} = & \frac{1}{2}(\partial_\mu h)(\partial^\mu h) - \frac{1}{2}m_H^2 h^2 + \frac{1}{2}m_W W_\mu W^\mu + \frac{1}{2}m_Z Z_\mu Z^\mu + gm_W h W_\mu W^\mu \\ & + \frac{g^2}{4}W_\mu W^\mu + g\frac{m_Z}{2\cos\theta_W}hZ_\mu Z^\mu - g^2\frac{1}{4\cos^2\theta_W}h^2Z_\mu Z^\mu - g\frac{m_H^2}{4m_W}h^3 \\ & - g^2\frac{m_H^2}{32m_W^2}h^4 + \text{const.} \end{aligned} \quad (1.13)$$

As can be seen in the Lagrangian in equation 1.13, the coupling strengths of the W and Z fields to the Higgs field are proportional to m_W and m_Z , respectively.

The Higgs mechanism in the SM for fermions

The Higgs mechanism for spontaneous symmetry breaking of the $SU(2)_L \times U(1)_Y$ gauge group of the SM generates the masses of the W^\pm and Z bosons. For originating the mass of the fermions without violating the EW gauge symmetry a similar procedure is carried but taking into account that the left-handed particles transform differently than the right-handed. To do so, additional terms including the Yukawa couplings are added into the Lagrangian. These terms are of the form:

$$\mathcal{L}_{\text{Yukawa},f} = -y_f(\bar{\chi}_L^f \Phi \chi_R^f + \bar{\chi}_R^f \Phi^\dagger \chi_L^f), \text{ with } \Phi = \begin{pmatrix} \phi^+ \\ \phi_0 \end{pmatrix}$$

where the f superindex runs over all quarks and charged leptons. The different y_f constants are known as Yukawa couplings of the particle f to the Higgs field. The Higgs doublet is denoted by Φ . For the electron $SU(2)$ doublet, the element with this coupling can be written as:

$$\mathcal{L}_e = -y_e \left[(\bar{\nu}_e \bar{e})_L \begin{pmatrix} \phi^+ \\ \phi_0 \end{pmatrix} e_R + \bar{e}_L (\phi^{+*} \phi^{0*}) \begin{pmatrix} \nu_e \\ e \end{pmatrix}_R \right]. \quad (1.14)$$

Here, y_e is the Yukawa coupling of the electron to the Higgs boson. After spontaneously breaking the symmetry as it is done in equation 1.12, and the y_e is

set to $y_e = \sqrt{2} m_e/v$ where m_e is the observed electron mass, the Lagrangian in equation 1.14 becomes:

$$\mathcal{L}_e = -m_e \bar{e} e - \frac{m_e}{v} \bar{e} e h \quad (1.15)$$

The first element of the Lagrangian in equation 1.15 gives mass to the electron and gives rise to the coupling of the electron to the Higgs fields in its non-zero vacuum expectation. The second term represents the coupling of the electron and the Higgs boson itself.

The non-zero vacuum expectation value occurs only in the neutral part of the Higgs doublet due to the form in the ground state in equation 1.12. This implies that the combination $\bar{\chi}_L^f \Phi \chi_R^f + \bar{\chi}_L^f \Phi^\dagger \chi_L^f$ can only generate masses for the fermions in the lower component of an $SU(2)$ doublet, i.e. the charged leptons and the down type quarks. Putting aside the procedure to give mass to the up-type quarks, this explains why the neutrinos do not get mass through the Higgs mechanism.

For up-type quarks, after applying the symmetry breaking and using $y_u = \sqrt{2} m_u/v$, the Lagrangian density takes the form:

$$\mathcal{L}_u = -m_u \bar{u} u - \frac{m_u}{v} \bar{u} u h.$$

These elements give rise not only to the mass of the fermions but also to the interaction strengths between these fermions and the Higgs boson. The Yukawa coupling of the fermions to the Higgs field is given by:

$$y_f = \sqrt{2} \frac{m_f}{v}, \quad (1.16)$$

where the Higgs vacuum expectation value is fixed by the Fermi coupling G_F and is measured to be $v = \sqrt{2} G_F \approx 246.22$ GeV. The G_F is measured from the antimuon (μ^+) lifetime measurement [25].

The value of fermionic masses is not predicted by the SM but obtained through experimental observations. Given the measured top-quark mass, $m_t = 172.76 \pm 0.30$ GeV [10], it is of particular interest the Yukawa coupling of the top quark to the Higgs field (y_t), which is almost exactly equal to one. It is important to verify this because deviation of the measured y_t from the SM prediction would be proof of new physics. The object of this thesis is precisely the measurement of a process whose cross-section is directly related to y_t .

1.6 Remarks about the limitations of the Standard Model

But the SM is not the ultimate theory, it is unquestionably one of the greatest successes of modern physics. Despite its achievements, many questions remain unsolved.

1.6.1 The parameters of the Standard Model

The SM contains 25 free parameters that must be determined through experimentation. These are the masses of the 12 fermions, the three coupling constants for describing the strength of the gauge interactions (g , g' and g_s) and the two parameters describing the Higgs potential (μ and λ) or, equivalently, its vacuum expectation value v , and the Higgs-boson mass m_h . Finally, there are three mixing angles and the complex phase of the CKM matrix and four Pontecorvo–Maki–Nakagawa–Sakata matrices, which mix the neutrino-mass eigenstates with neutrino-flavour eigenstates [26, 27].

From the 25 free parameters of the SM, 14 are associated with the Higgs field, eight with the flavour sector and only three with the gauge interactions. This emphasises the crucial significance of precisely measuring and comprehending the properties of the Higgs particle.

1.6.2 Limitations of the Standard Model

While the SM is an extremely successful theory that has passed rigorous testing there are several limitations of the SM and a variety of phenomena that it does not explain. In the following pages, the most relevant issues with the SM are described.

Matter–antimatter asymmetry

In principle, the Big Bang should have produced an equal amount of matter and antimatter which would all have then annihilated, leaving behind an empty universe filled with electromagnetic radiation. However, everything that is observed nowadays constituted of matter, from the tiniest life forms on Earth to the greatest celestial objects. In comparison, there is not a lot of antimatter around.

By looking at the cosmic microwave background (CMB) radiation, which contains the residual γ of the Big Bang, researchers have determined that there was a symmetry between the matter and antimatter content in the early universe. For

every 3×10^9 antimatter particles, there were 3×10^9 and 1 matter particle. The matter and antimatter annihilated and produced the CMB and the remaining 1 part turned into all the stars and galaxies that are seen.

Research carried out during the last few decades has revealed that the laws of nature do not equally apply to matter and antimatter [28]. So far, the only non-trivial difference between matter and antimatter found is the CP asymmetry or CP violation. However, the CP asymmetry included in the SM and observed in the experiments is insufficient to explain the composition of the observable universe and, hence, extensive searches for new sources of CP violation are being carried out. In this context, the study in this thesis seeks for new CP-violation sources.

Gravity

Gravity is the first force that any person learns about and the one known by humankind for longer. The SM describes all fundamental interactions but gravity. GR is a geometric theory that currently describes gravitation in modern physics. Some of the suggested solutions to integrate gravitational interactions in the SM consist of postulating a new force carrier particle, the “graviton”, that mediates this interaction in a similar way to how the gauge bosons were proposed. Other explanations state that gravity can only be described by a deeper theory in which the time-space structure is not flat but dynamic.

Neutrino masses

According to the SM, the neutrinos are massless, nevertheless, many experiments confirm that this is not true [29]. This is due to a property of neutrinos that allows them to change their flavour while travelling through space, this feature is known as “neutrino oscillations”. Each of the three neutrino flavours (ν_e , ν_μ , ν_τ) is a linear combination of three discrete neutrino-mass eigenstates (ν_i with $i \in \{1, 2, 3\}$) with mass eigenvalues (m_i). While the neutrino oscillation experiments could probe the squared neutrino-mass eigenvalues (Δm_{ij}^2), both the total scale of the masses and the sign of Δm_{ij} remain as some of the most relevant open questions in particle physics.

Non-zero neutrino masses opened an interesting portal beyond SM physics and, even though neutrinos are very elusive when it comes to detecting them, some next-generation experiments such as Dune [30] target to set competitive and model-independent limits on neutrino masses.

Regarding the nature of the neutrino mass, one could add mass terms to the SM as it is done in Section 1.5.1 for the up-type quarks but the origin of the neutrino masses is still not known. Also, if neutrinos gained mass through Yukawa

interaction, it would imply the presence of right-handed neutrinos, which has not been observed.

Dark energy

According to cosmological observations, the matter described by the SM only makes up around 5% of the universe. It turns out that roughly 68% of the universe is made of dark energy, which is not considered by the SM.

Dark energy is an unknown type of energy postulated to explain the observed accelerated expansion of the universe. This expansion is dominated by a spatially smooth component with negative pressure called dark energy. Modern cosmological measurements are based on supernovae, cosmic microwave background fluctuations, galaxy clustering and weak gravitational lensing, and methods agree with a spatially flat universe with about 30% matter (visible and dark) and 70% dark energy [31].

Dark matter

Dark matter (DM) adds up to approximately 85% of all matter content in the universe. This matter is called dark because it does not interact with the electromagnetic field. The only known way to interact with DM is via gravitational interaction, which is about 25 orders of magnitude weaker than the weak force (see Table 1.1). This is why DM is so difficult to detect. The SM does not provide a proper explanation but searches are being carried and candidates such as weakly-interacting massive-particles (WIMPs) or axions⁴.

The existence of DM is inferred through gravitational effects in astrophysical and cosmological observations. The rotational speed of the galaxies [34], the gravitational lensing [35] and the CMB angular spectrum [36] are some examples of phenomena that cannot be explained with general relativity unless there is more present matter than what is seen.

Although the vast majority of the scientific community accepts dark matter existence, alternative explanations for the observed phenomena have suggested. Most of these model consists in modifications of GR. The search of DM at particle colliders, which is focussed on large missing transverse energy signatures, has not resulted in any observation. Nevertheless, the existence of a particle is not discarded, only its presence within the detector sensitivity limits.

⁴An axion is a hypothetical elementary particle postulated to resolve the strong CP problem [32, 33].

Additional issues

The different problems mentioned so far are just some of the most relevant open questions that fundamental physics has not been able to answer yet. Nonetheless, there are many other open questions and theories. To mention a few:

- Strong CP violation: This refers to the fact that, while QCD does not explicitly prohibit CP violation in strong interactions, this has not been observed experimentally.
- Naturalness: It is the property that the dimensionless ratios between free parameters or physical constants appearing in a physical theory should take values of order unity. By looking at the parameters of the SM described in Section 1.6.1, it can be seen that the naturalness principle is not satisfied. For instance, the masses of the first generation of fermions are in the range of 1 MeV while the top quark has a mass of ~ 173 GeV. Though this is not a flaw in the theory itself, it is frequently seen as a sign of undiscovered principles hidden behind a more comprehensive theory.
- Majorana neutrinos: It is not clear yet if neutrinos are Majorana particles, i.e. they are their own antiparticles ($\nu = \bar{\nu} = \nu_M$) [37]. Current experiments trying to solve this question are focused on neutrinoless double- β decay, which can occur only if neutrinos are Majorana particles.
- Unification of EW with QCD: There are unification attempts to treat all interactions as one, with the same coupling constant and the same symmetry group⁵. In the same way the EW unifies QED and Weak forces, the grand unified theories unifies all three interactions of the SM at high energies, where the coupling constants approach each other.
- String theory: It is a theoretical framework in which fundamental point-like particles are understood as vibrational states of a more basic object, the so called “string”. A string is a one-dimensional entity that can either be opened (forming a segment with two endpoints) or closed (forming a loop) and may have other special properties [38].
- Supersymmetry: It is an extension of the SM. In supersymmetry the equations for force and the equations for matter are identical and each SM particle has its supersymmetric partner “sparticle” which differs by half spin unit [39].
- Multidimensions: As supersymmetry or string theory, it is another extension of the SM. This theory proposes the existence of additional spatial dimen-

⁵The most popular symmetry group for unification is $SU(5)$.

sions. If extra dimensions exist, they could explain why the universe is expanding faster than expected, and why gravity is weaker than the other forces of nature [40]. As well as string theory and supersymmetry, multidimension-based theories have not been accepted nor discarded yet by any experiment.

As exposed above, there are far too many unanswered questions and loose ends. The HL-LHC [41] and the next generation of experiments will look for evidence of physics outside the SM in the next years.

Among the open questions, unresolved concerns and measurements to be completed, the research carried in this thesis tries to address some. The study of the associated production of a single top quark with a Higgs boson allows us to experimentally determine possible sources of CP violation. Now that the bases of the SM are settled, in the chapters to come, the context of this particular production is discussed.

Chapter 2

Physics of the top quark and the Higgs boson

*Magisches Theater
Eintritt nicht für jedermann
Nur für Verrückte!*
—HERMAN HESSE
DER STEPPENWOLF (1927)

The central focus of this dissertation revolves around the top quark and the Higgs boson, two particles whose interaction can help to answer some of the SM open questions. In this chapter, I delve into their properties and explore their interactions.

The chapter is structured into three sections. Section 2.1 is dedicated to the top quark, where its primary characteristics are discussed, including its production, decay modes, and other properties. In Section 2.2, the Higgs boson is discussed and its fundamental attributes are reviewed.

Finally, in Section 2.3, the main characteristics of the associated production of the top quark with a Higgs boson are presented. A particular emphasis on the single-top-quark scenario and its influence on possible contribution to CP violation is given.

2.1 The top quark

The top quark (t) is the up-type quark of the third generation of fermions. Its most distinctive feature is its mass, which is the largest among all fundamental particles. The left-handed top quark is the $Q = 2/3$ and $T_3 = +1/2$ member of the weak isospin doublet that also contains the bottom quark. The right-handed top quark is the $SU(2)_L$ weak isospin singlet ($Q = 2/3$ and $T_3 = 0$). The top quark is often regarded as a window for new physics since it provides a unique laboratory to test the understanding of the SM.

Due to its large mass, the top quark decays with a lifetime of $\tau_t = 5 \times 10^{-25}$ s [35]. Actually, it is shorter than its hadronisation timescale ($1/\Lambda_{\text{QCD}} \sim 10^{-24}$ s). This represents an exceptional opportunity to study quarks in a free state, something that is quite exceptional due to colour confinement, as explained in Section 1.4. In fact, the top quark is the only quark that can be investigated unbounded. Its lifetime is also smaller than the spin decorrelation timescale ($m_t/\Lambda_{\text{QCD}}^2 \sim 10^{-21}$ s [42], being m_t the mass of the top quark), implying that the top-quark states conserve their spin state from its production to its decay. Therefore, the top-quark properties, such as the spin information and the top-quark polarisation, can also be accessed through its decay products and, consequently, be measured.

Another consequence of its large mass is that the top quark is the only quark with a Yukawa coupling to the Higgs boson (y_t) of the order of one; hence, a thorough understanding of its properties (mass, couplings, decay branching ratios, production cross-section, etc) can reveal crucial information on basic interactions at the electroweak symmetry-breaking scale and beyond. The main objective of this thesis is the study of the interplay between the top quark and the Higgs boson to help to determine if the y_t is that predicted by the SM or if there is some CP-violating phase that would affect the sign of y_t . The theoretical base for understanding the associated production of a top quark and a Higgs boson is given in Section 2.3.

2.1.1 Top-quark discovery

In 1973, T. Kobayashi and M. Maskawa postulated the possibility of a third generation of quarks to explain the CP violation in kaon decays [43]. To match the names of the up and down quarks, the new generation of quarks was given the names top and bottom. The GIM¹ mechanism [44], which predicted the existence

¹Standing for Glashow–Iliopoulos–Maiani (GIM), it is the mechanism to describe the flavour-changing neutral currents.

of the yet-to-be-discovered charm quark, was used to make this prediction. When the charm was observed [45], the GIM mechanism was integrated into the SM, and the postulation of the third family, and thus the top quark, gained acceptance. Shortly after the charm, the bottom quark was discovered in the E288 experiment at Fermilab [46], reinforcing the idea of the existence of the top quark. However, due to its large mass, it took 18 years to confirm the existence of the 6th quark.

The top quark was observed for the first time by the CDF [47] and D \emptyset [48] collaborations at Tevatron (Fermilab) in 1995. Back then and until the start of LHC, Tevatron was the only accelerator powerful enough to produce top quarks.

Top-quark mass

As discussed in Section 1.6.1, m_t is a free parameter of the SM. The theory does not predict its value, hence it must be determined experimentally [49]. Note that what experiments measure is not the SM m_t but either the pole mass m_t^{pole} or the MC top-quark mass m_t^{MC} . It is expected that the difference between the m_t^{MC} definition and the m_t^{pole} is of the order of 1 GeV.

Among the properties of the top quark, its mass is the one that has received the most attention so far. Accurate measurements of m_t are vital for determining the parameters of the EW global fits, which are essential for evaluating the internal consistency of the SM and exploring its potential extensions [50, 51]. Moreover, the m_t value influences the stability of the Higgs-boson potential, carrying which has cosmological implications [52–54]. The most recent studies for the top-quark mass measurements result in $m_t = 172.76 \pm 0.30$ GeV [10]. This number is an average of the measurements at the Large Hadron (LHC) with ATLAS (172.69 ± 0.66 GeV [55]) and CMS (172.6 ± 2.5 GeV at CMS [56]) and at Tevatron with CDF and D \emptyset (combined result: 174.30 ± 0.89 GeV [57]). These values are measured from the kinematics of $t\bar{t}$ events and refer to the m_t^{pole} .

2.1.2 Top-quark production at LHC

The LHC is sometimes referred to as a top-quark factory due to its capacity to produce such particles in large quantities. In this collider, at proton–proton (pp) collisions, the top quark is mainly produced via two mechanisms: through the strong interaction in top-quark–antiquark pairs ($t\bar{t}$), and by means of the tWb vertex of EW interaction in single top quarks. Apart from the $t\bar{t}$ (see Section 2.1.2.1) and single top quark (see Section 2.1.2.2) productions, the associated $t\bar{t}X$ and tX productions play a role on producing top quarks (see Section 2.1.2.3 and Section 2.1.2.4, respectively).

Since top quarks often constitute a main background in other physics analyses, a better understanding of the properties of this particle will directly translate into improvements in those studies.

2.1.2.1 Top-quark–antiquark pairs production

The production of $t\bar{t}$ events is the largest source of production of top quarks in hadron collisions. This process is one of the most important at LHC because it allows us to precisely study the properties of the top quark. Additionally, due to the dominance of this production mode, the $t\bar{t}$ production is also a major background in many measurements and searches for rare processes as mentioned above. The physics analysis carried out in this thesis is a good example, where the $t\bar{t}$ process is the main background in both of the analysed decay channels.

For proton–antiproton ($p\bar{p}$) collisions at Tevatron or pp at LHC, the $t\bar{t}$ production is described by perturbative QCD. In this approach, a hard scattering process between the two hadrons is the result of an interaction between the quarks and gluons that constitute these hadrons. This model is described in detail in Section 4.1.

At LHC, gluon fusion dominates with 90% of the $t\bar{t}$ production. It is followed by the quark–antiquark annihilation, which accounts for 10% of the total $t\bar{t}$ production. The theoretical calculations for the $t\bar{t}$ production at $\sqrt{s} = 13$ TeV are done to an accuracy of next-to-next-to-leading order (NNLO) in QCD and complemented with next-to-next-to-leading logarithmic (NNLL) resummation [58, 59]:

$$\sigma_{\text{NLO+NNLL}}^{\text{pred}}(t\bar{t}) = 833.9^{+20.5}_{-30.0}(\text{scale}) \pm 21.0(\text{PDF} + \alpha_s)^{+22.5}_{-23.2}(\text{mass}) \text{ pb}.$$

Here, PDF stands for parton distribution function and α_s is the strong coupling constant. The scale uncertainty refers to the sensitivity to the choice of factorisation and renormalisation scales² [60]. These scales are further discussed in Section 4.1.1. The ATLAS and CMS collaborations have measured its cross-section through different final state channels and the most precise result obtained so far at $\sqrt{s} = 13$ TeV is [61]

$$\sigma^{\text{obs}}(t\bar{t}) = 829 \pm 1(\text{stat.}) \pm 13(\text{syst.}) \pm 8(\text{lumi.}) \text{ pb}.$$

Thanks to this large cross-section, the statistical uncertainty is not constraining the measurements of the $t\bar{t}$ production, which is performed with good precision ($\mathcal{O}(2\%)$). This large cross-section also makes $t\bar{t}$ the most relevant background in this analysis.

²In QCD and QED, the renormalisation scale is a parameter that helps to eliminate infinities arising in the calculations of interactions. It is associated with the energy scale at which a physical process occurs.

2.1.2.2 Single-top-quark production

In addition to the $t\bar{t}$ production, the single-top-quark production is of great importance to the study of the top-quark properties at the LHC. This mechanism has a cross-section about three times smaller than that of $t\bar{t}$ processes and it is almost exclusively produced through the EW interaction. This is precisely the reason why single-top-quark production is essential to gather information about the tWb interaction and to directly measure the CKM matrix element $|V_{tb}|$ at hadron colliders. The reason why the single top quark is produced and decays via a b -quark and not via strange or down quarks is because the CKM elements V_{ts} and V_{td} are smaller than V_{tb} by several orders of magnitude.

At leading order (LO), there are three production modes for single-top-quark events, being the t -channel the dominant mechanism at the LHC with, approximately 70% of the single-top-quark cross-section at $\sqrt{s} = 13$ TeV. At this energy its cross-section for the t -channel is calculated to be [62]

$$\sigma_{\text{NLO+NNLL}}^{\text{pred}}(\text{t-channel}) = 214.2^{+2.4}_{-1.7}(\text{scale})^{+3.3}_{-2.0}(\text{PDF} + \alpha_s) \text{ pb}$$

and the most precise measurement is [63]

$$\sigma^{\text{obs}}(\text{t-channel}) = 221 \pm 1(\text{stat.}) \pm 13(\text{syst.}) \pm 2(\text{lumi.}) \text{ pb.}$$

The other processes are the s -channel and the associated tW production. For the latter, the predicted cross-section is [64]

$$\sigma_{\text{NLO+NNLL}}^{\text{pred}}(tW) = 71.7 \pm 1.8(\text{scale}) \pm 3.4(\text{PDF}) \text{ pb}$$

and it is measured to be [65]

$$\sigma^{\text{obs}}(tW) = 79.2 \pm 0.9(\text{stat.}) \pm 7.7(\text{syst.}) \pm 1.2(\text{lumi.}) \text{ pb.}$$

Only t -channel and tW productions are relevant to the EW single-top-quark production at the LHC, since the s -channel has not been observed at $\sqrt{s} = 13$ TeV. At this energy, its predicted cross-section is [66]

$$\sigma_{\text{NLO+NNLL}}^{\text{pred}}(\text{s-channel}) = 10.32^{+0.29}_{-0.24}(\text{scale}) \pm 0.27(\text{PDF} + \alpha_s) \text{ pb.}$$

2.1.2.3 Associated $t\bar{t}X$ production

The associated-top-quark productions are important processes to measure the coupling of the top quark to other particles of the SM. When a pair of top quarks is produced along with another particle the process is referred to as $t\bar{t}X$. The most

relevant $t\bar{t}X$ productions are those in which the pair is produced together with W , Z or γ bosons. From these, $t\bar{t}W$ and $t\bar{t}Z$ play a role in the analysis carried in this thesis. These two processes along with $t\bar{t}$ are the three most relevant backgrounds in one of the channels of the analysis performed in this thesis.

The cross-sections for the $t\bar{t}W$, $t\bar{t}\gamma$ and $t\bar{t}Z$ productions are measured by the ATLAS and CMS collaborations. These are presented in Figure 2.1. For the three $t\bar{t}X$, the cross-sections are small in comparison to $t\bar{t}$ and single top-quark productions. However, with the complete LHC Run 2 data event sample it is possible to explore these productions, which are also sensitive to new physics [67]. The associated production of a $t\bar{t}$ pair with a Higgs boson ($t\bar{t}H$) is another important type of $t\bar{t}X$ production and it is described in Section 2.3.2.

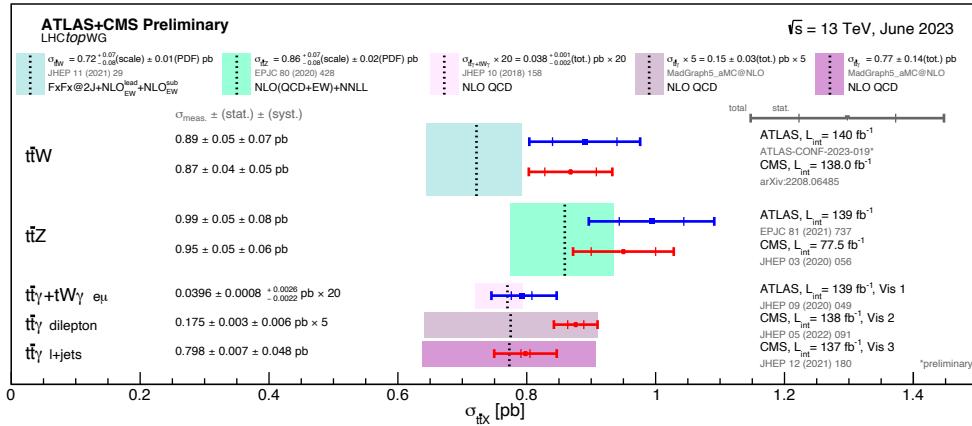


Figure 2.1: Summary of the ATLAS and CMS measurements of the $t\bar{t}X$ production cross-sections at $\sqrt{s} = 13$ TeV. Here X means either W , Z or γ .

2.1.2.4 Associated tX production

Not only the top-quark pairs but also the single top quark can be produced in association with other particles (tX). This thesis focusses precisely on the search of a tX production, the one in which the single top quark is produced along a Higgs boson and an additional parton (tHq). The features of the single-top-quark production in association with a Higgs boson are discussed in more detail in Section 2.3.3.

Apart from being the signal process, the tX production also plays a role in the background since the tZq process is one of the backgrounds more difficult to separate from the tHq signal.

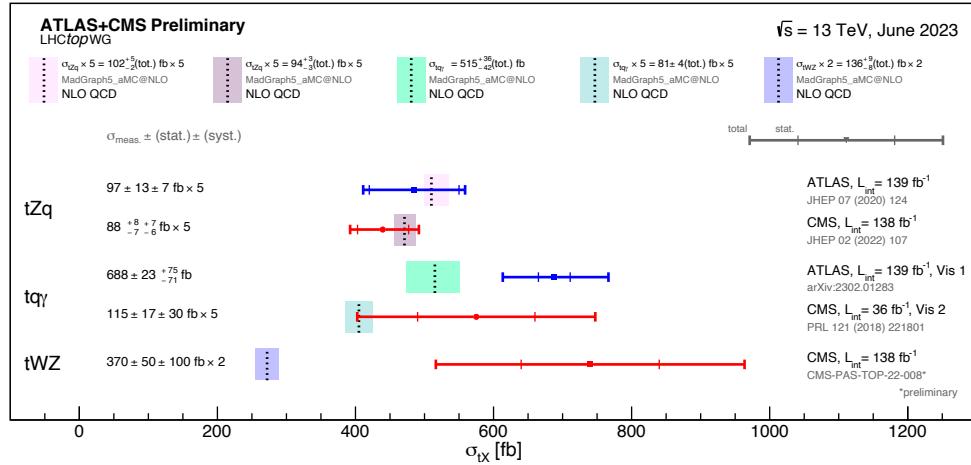


Figure 2.2: Summary of the ATLAS and CMS measurements of the tX production cross-sections at $\sqrt{s} = 13$ TeV. Here, $X = Z, \gamma$ and W . The tHq process is not included in this summary plot but the work developed in this thesis aims to help to provide better limits on its cross-section.

The other associated tX production of the EW type is the $tq\gamma$, in which the top quark is produced along with a photon. Both tZq and $tq\gamma$ are sensitive to beyond the SM (BSM) physics like flavour-changing neutral currents or vector-like quarks. ATLAS and CMS have measured both processes and have found an agreement with the SM model, as can be seen in Figure 2.2.

2.1.3 Top-quark decay

As anticipated in Section 2.1.2.2, due to the large value of the V_{tb} element of the CKM matrix, the top quark is expected to decay almost entirely ($\sim 99.8\%$) to a b quark and a W boson ($t \rightarrow Wb$). The final-state decay is classified according to the subsequent decay of the W boson. Since the W bosons are massive vector bosons, their lifetime is very short ($\tau_W \approx 3 \times 10^{-25}$ s) and hence they will rapidly decay to leptons or quarks that will form hadrons. Due to its large mass, the W boson can decay to any quark except the top quark. For the W^+ , the branching

ratios³ (BRs) for the different decay modes are [10]:

$W^+ \rightarrow e^+ \nu_e$	$(10.71 \pm 0.16)\%$,
$W^+ \rightarrow \mu^+ \nu_\mu$	$(10.63 \pm 0.15)\%$,
$W^+ \rightarrow \tau^+ \nu_\tau$	$(11.38 \pm 0.21)\%$,
$W^+ \rightarrow q\bar{q}$ (hadrons)	$(67.41 \pm 0.27)\%$,
$W^+ \rightarrow \text{invisible}$	$(1.4 \pm 2.9)\%$.

For the conjugate processes involving the W^- , the BRs are the same. Therefore, the W boson decay and consequently the top-quark decay can be classified either as leptonic or hadronic.

2.1.4 Top-quark properties

As commented, the top quark exhibits unique properties and plays a pivotal role in many analyses conducted at the LHC. Some of the key investigations related to the top quark are:

- **Charge asymmetry in $t\bar{t}$ production:** This refers to the subtle deviation observed between the rapidity distributions of top quarks and top antiquarks when produced in pairs [68].
- **Polarisation of W bosons in top-quark decays:** A measurement of the polarisation of the W bosons is conducted by quantifying the fractions of longitudinally, left-handed, and right-handed polarised W bosons [69].
- **Top–antitop-quark energy asymmetry:** Analogous to the charge asymmetry, this research focuses on scenarios where the top-quark pair is produced in association with a high- p_T jet [70].
- **Top-quark-pair spin correlations:** Due to the short lifetime of the top quark, its spin information can be extracted from its decay products. Nonetheless, not all decay particles carry the spin information equivalently. Charged leptons emanating from leptonically decaying W bosons almost wholly encapsulate the spin information of the top quark [71].
- **Flavour changing neutral currents (FCNCs):** A FCNC is a process where a quark of one flavour changes into a quark of another flavour, without

³For each decay mode, the branching ratio (BR) is defined as the fraction times that the particle decays in that particular mode with respect to total possible decays.

a change in its electric charge. Top quarks, due to their large mass, are particularly interesting when it comes to FCNCs. The top quark can decay through FCNC processes to lighter quarks in association with a photon, Z boson, or gluon.

- **Top-quark polarisation:** In t -channel single-top-quark production processes, there is a notable production of a highly-polarised-top quarks. These single top quarks manifest with spins completely aligned along the direction of the down-type quarks. Contrary, in the scenario of a single top antiquark, the alignment direction is inverted. A more comprehensive examination of polarisation can be found in Section 2.1.4.1, to which I have contributed during the tenure of this thesis.

2.1.4.1 Top-quark polarisation

The lifetime of the top quark is shorter than the depolarisation scale and, hence, the top-quark-spin information can be transferred into its decay products. This allows to measure the top-quark polarisation from its child particles. The polarisation refers to the alignment between the momentum and the spin of the top quark and antiquarks. The polarisations of the top and antitop quarks are important quantities because they are sensitive to many BSM effects and can also provide useful input for the MC generators which are described in Section 4.3.

At the LHC, the single-top-quark production is the only source of highly-polarised top quarks⁴. In the t -channel (see Section 2.1.2.2) the top quark is created with a high degree of polarisation in the direction of the spectator-quark momentum [73]. As a consequence of the vector and axial-vector form of the coupling of the top quark to the W boson and bottom quark in the t -channel (tWb vertex), specific values of the polarisation vectors $\{P_x', P_y', P_z'\}$ of top quarks/antiquarks are expected in the SM.

Even though it is not described in detail in this manuscript, during the development of my thesis I have also been involved in the first measurement of the top-(anti)quark-polarisation vectors [73]. My contribution is an extension of the work done in Reference [74]. In this work, the three components of the polarisation vector for the top quark and antiquark are measured in the single-top-quark t -channel production. Using the entire Run 2 dataset recorded by ATLAS and demanding events with exactly one light lepton, I defined a set of stringent selection requirements to discriminate the t -channel signal from the background contribu-

⁴The top quarks from the $t\bar{t}$ production are unpolarised at LO but the spins of the top quarks and antiquarks are strongly correlated [72].

tions. This signal-region⁵ definition used specific cuts⁶ in several variables such as the lepton transverse momentum (p_T) or the invariant masses of several particles. I have also developed the so called trapezoidal cut, which is described in the published paper [73].

The polarisation vectors are later obtained from the distributions of the direction cosines of the charged-lepton momentum in the top-quark rest frame: $\cos(\theta_{lx'})$, $\cos(\theta_{ly'})$ and $\cos(\theta_{lz'})$. Figure 2.3 shows the distributions for one of these angular variables.

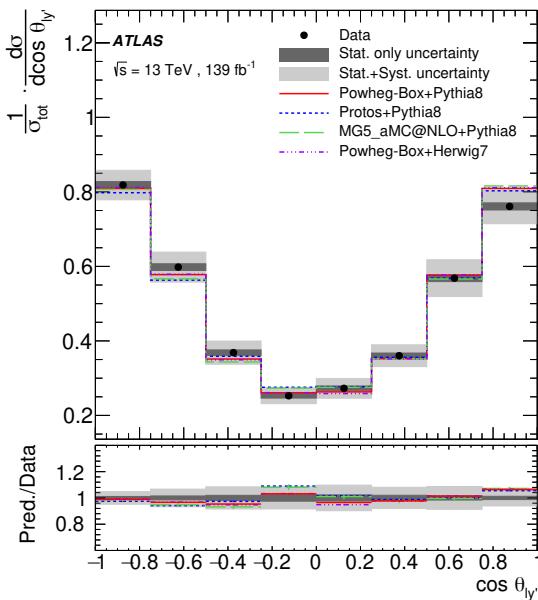


Figure 2.3: Normalised differential cross-sections as a function of $\cos(\theta_{ly'})$ [73]. The data is shown as black points with statistical uncertainties compared to the predictions of the MC generators, which are shown as lines. The ratio between the predictions and data is shown on the lower panel. These plots are inclusive for top quark and top antiquark.

Limits on the two of the components of the polarisation vector of the top quark and antiquark are set and Figure 2.4 presents the observed best-fit polarisation measurements for P_x' and P_z' in the two dimensional parameter space.

Data measurements of the polarisation-vector components and differential cross-sections show good agreement with SM predictions. The results are consistent with

⁵The signal region is a region of the phase space enriched with events of the signal process.

⁶To “cut” on a variable is to apply a threshold on this variable and keep only events satisfying this condition. A cut-based analysis is applying such thresholds on several variables to select events.

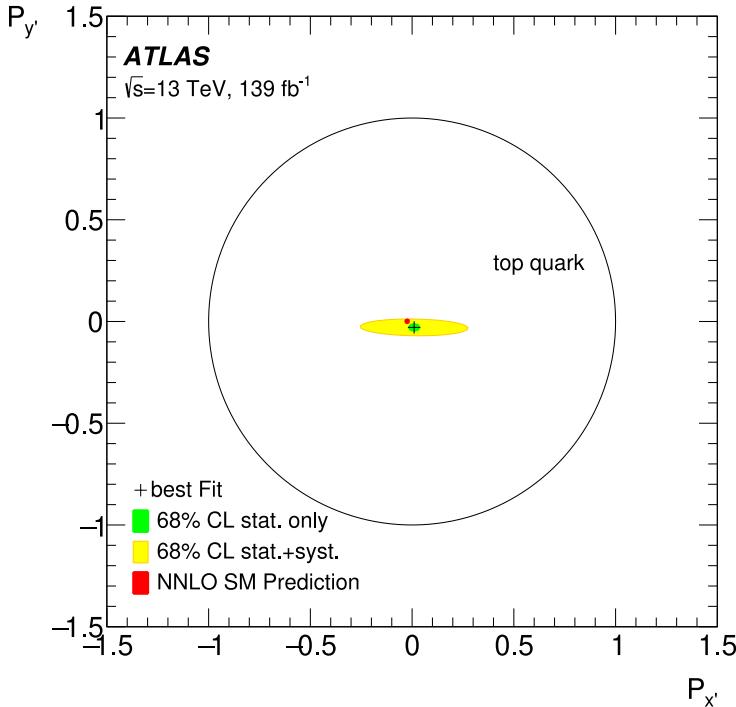


Figure 2.4: Observed best-fit limit on two-dimensional top quark polarisation parameter space $\{P_{x'}, P_{y'}\}$. The statistical-only (green) and the statistical+systematic uncertainty contours have a 68% CL [73]. The physically allowed values for $P_{x'}$ and $P_{y'}$ lay inside the black circle. The red point indicates the parton-level prediction at NNLO.

NNLO QCD predictions and expectation of $P_{y'}^t = P_{y'}^{\bar{t}} = 0$ from the hypothesis of CP symmetry in the top-quark and top-antiquark decay.

2.2 The Higgs boson

Following the top quark, the Higgs boson (H) is the second most massive particle in the SM with a mass of $m_H = 125.25 \pm 0.17$ GeV [10]. The value provided by Reference [10] is an average of the ATLAS combined measurement ($m_H = 124.86 \pm 0.27$ GeV [75]) and the CMS results ($m_H = 125.46 \pm 0.16$ GeV [76]). The existence of the Higgs boson was theorised in 1964 by three independent groups [11, 12, 21], and its discovery meant one of the greatest successes of the SM.

2.2.1 Discovery of the Higgs boson

Any particle physicist enthusiast remembers July 4th of 2012 pretty well, it was the day when the ATLAS [77] and CMS [78] collaborations announced the discovery of a massive state H with the properties expected for the Higgs boson within the SM.

Both the ATLAS and CMS collaborations reported excesses of events for 2011 ($\sqrt{s} = 7 \text{ TeV}$ and $\mathcal{L} = 4.8 \text{ fb}^{-1}$) and 2012 ($\sqrt{s} = 8 \text{ TeV}$ and $\mathcal{L} = 5.8 \text{ fb}^{-1}$) datasets of pp collisions at the LHC. This surplus of events was compatible with its production and decay with the SM Higgs boson in the mass region $m_H \in [124, 135] \text{ GeV}$. Before that, the CDF [79] and D \emptyset [80] collaboration at Tevatron also reported an excess in the mass region $m_H \in [120, 135] \text{ GeV}$ both of them with a significance lower than 3σ and therefore, not enough to claim a discovery⁷.

2.2.2 Higgs-boson production at LHC

One of the reasons the Higgs boson was the last of the SM's fundamental particles to be discovered is its relatively high mass, which required significant energy for its production. Even in high-energy collisions, the production of a Higgs boson is a rare event. Only a tiny fraction of collisions at the LHC produce a Higgs boson, so vast numbers of collisions had to be analysed to find the Higgs boson. This is the reason why it was not found at Tevatron. Colliders such as SLAC's Linear Collider [81] or CERN's LEP [82] had enough energy to produce the Higgs boson but they were colliding electrons and positrons. Since the coupling of the Higgs boson to fermions is proportional to the fermion's mass, the $e^-e^+ \rightarrow H$ process is highly suppressed⁸ and, for this reason, there were not enough statistics of events with a Higgs boson at SLAC and LEP. The most favoured way of producing a Higgs boson is through the mediation of the heaviest fundamental particles in the SM because these have the strongest couplings with the Higgs boson and, consequently, the greater cross-sections.

The four most dominant processes for Higgs-boson production at LHC are summarised in Figure 2.5. These processes are:

⁷The 3σ expresses that there is a 99.7% probability that a given result is not a random fluctuation, meaning there is roughly a 0.3% chance that the observed effect is due to random chance. A 3σ level is often considered evidence for a potential discovery.

⁸The dominant Higgs-boson production in e^-e^+ annihilation is the so called Higgsstrahlung, a process in which the H is produced in association with a Z boson similar to Figure 2.5c. Due to the small mass of the electrons, the electron–Higgs-boson coupling does not favour the $e^-e^+ \rightarrow H$ process.

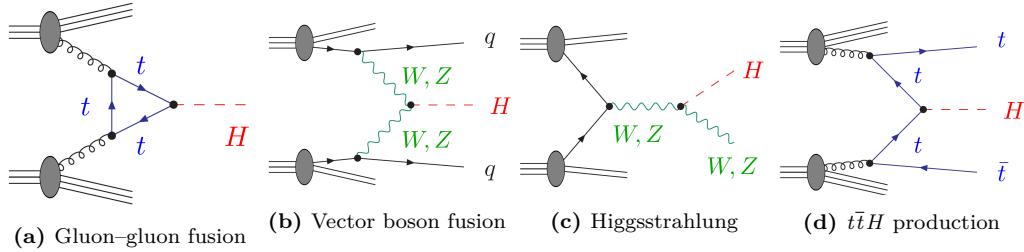


Figure 2.5: LO Feynman diagrams for the dominant production mechanisms of a Higgs boson at hadron colliders.

- **Gluon–gluon Fusion (ggF):** The process $gg \rightarrow H$ has to be mediated by a massive fermion loop. This is due to the fact that there is no direct gluon–Higgs-boson coupling within the SM. Although in principle all quarks should be included in the loop, in practice it is the top quark the one doing so because its coupling to the Higgs boson is 35 times stronger than the next-heaviest fermion, the bottom quark. Due to the abundance of gluons in pp collisions, the ggF is very favoured at the LHC.
- **Vector boson fusion (VBF):** The second most important mode is the radiation by the incoming quarks of a pair of W or Z bosons that fuse to form a Higgs boson. The vector bosons ($V = W$ or Z) of the process $V\bar{V} \rightarrow H$ are originated from initial state quarks which scatter through the final state (changing its flavours in the case of W fusion) producing two forward jets.
- **Higgsstrahlung (VH):** There is another significant contribution involving the W or Z bosons, the Higgsstrahlung or associated WH or ZH production. Here, an off-shell W or Z boson (formed from the annihilation of two quarks) radiates a Higgs boson via $V^* \rightarrow VH$.
- **Quark-pair associated production ($q\bar{q}H$):** In this mode, the Higgs is produced from a $q\bar{q}$ pair via $q\bar{q} \rightarrow H$ with a $q\bar{q}H$ final state. Typically, the involved quark pair is either a $b\bar{b}$ or $t\bar{t}$. In the case of $t\bar{t}$, the top quarks decay before hadronising, leading to final states with a high number of physics objects.
- **Single top quark associated production (tHX):** This sub-dominant contribution can be either a tHq , a tWH or a s -channel. The former process constitutes the central topic developed in this thesis. Details about these production modes are further discussed in Section 2.3.3.

The cross-section of the different mechanisms for single-Higgs-boson production at $\sqrt{s} = 13$ TeV as a function of m_H are shown in Figure 2.6. Assuming a $m_H =$

125.2 GeV, the Higgs-boson production cross-sections (σ_{tH}) for the different modes (see relative fraction in Figure 2.7) are [83]:

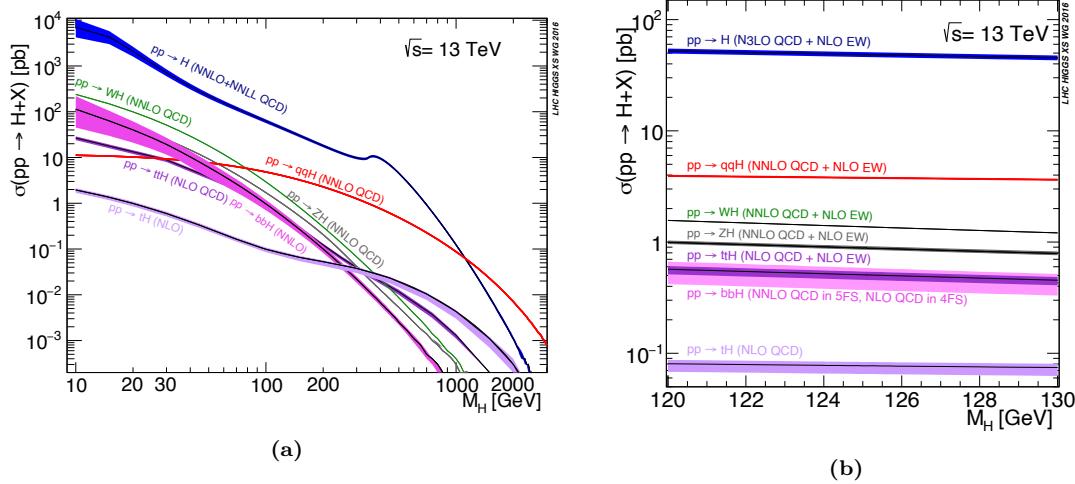


Figure 2.6: Higgs-boson-production cross-sections as function of m_H at $\sqrt{s} = 13$ TeV [83]. The σ_{tH} presented in these plots accounts for the t -channel and s -channel but not the tW process. A wide range of m_H values is showed in (a). In (b) is shown the result zooming around the measured Higgs-boson mass value.

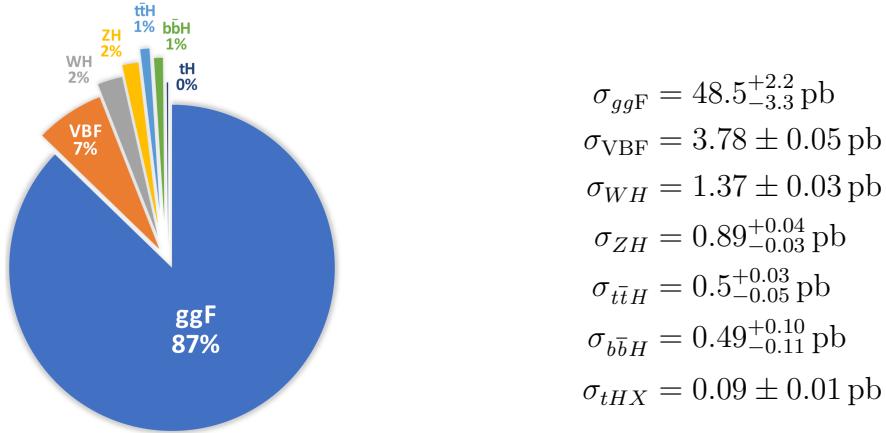


Figure 2.7: Higgs-boson-production modes.

2.2.3 Higgs-boson decay

The Higgs boson has a very short lifetime ($\tau_H = 1.6 \times 10^{-22} \text{ s}$ [83]) and, hence, is always detected through its decay products. Figure 2.9 shows the BR for the different Higgs-boson decay modes.

Despite the expected large Yukawa coupling between the Higgs boson and the top quark, the $H \rightarrow t\bar{t}$ is forbidden since $m_H < 2m_t$. Consequently, the most prominent decay mode is $H \rightarrow b\bar{b}$ followed by $H \rightarrow W^+W^-$. This is why for the tHq searches, the channel in which the Higgs decay to $b\bar{b}$ is the one with higher statistics. For the other fermionic decays, the decay rates are ordered by the fermion masses, being the $\tau^-\tau^+$ decay mode (see Figure 2.10b) the biggest among the leptonic ones. Regardless of the expected large coupling between the weak-force bosons and the Higgs boson, the $H \rightarrow VV^*$ is suppressed due to the requirement that one vector boson has to be produced off-shell⁹.

For the analysis carried out in this thesis, three particular Higgs-boson-decay modes are considered: $H \rightarrow W^+W^-$ (see Figure 2.10a), $H \rightarrow ZZ$ (see Figure 2.10c) and $H \rightarrow \tau^-\tau^+$ (see Figure 2.10b). Sorted by its importance and assuming a $m_H = 125.2$ GeV, the BRs for the Higgs boson are presented in Figure 2.8 and listed below [83].

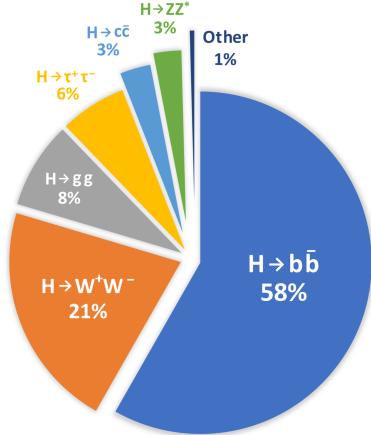


Figure 2.8: Higgs-boson-decay modes.

$H \rightarrow b\bar{b}$	$(57.92 \pm 0.29)\%$
$H \rightarrow W^+W^-$	$(21.70 \pm 0.11)\%$
$H \rightarrow gg$	$(8.17 \pm 0.26)\%$
$H \rightarrow \tau^-\tau^+$	$(6.24 \pm 0.03)\%$
$H \rightarrow c\bar{c}$	$(2.888 \pm 0.014)\%$
$H \rightarrow ZZ$	$(2.667 \pm 0.013)\%$
$H \rightarrow \gamma\gamma$	$(2.270 \pm 0.023)\%$
$H \rightarrow \mu^-\mu^+$	$(2.165 \pm 0.011)\%$
$H \rightarrow Z\gamma$	$(0.155 \pm 0.008)\%$
$H \rightarrow$ Others	$< 0.2\%$

⁹Off-shell means that the particle is produced virtually and it does not satisfy the energy-momentum relation: $E^2 = p^2 + m^2$.

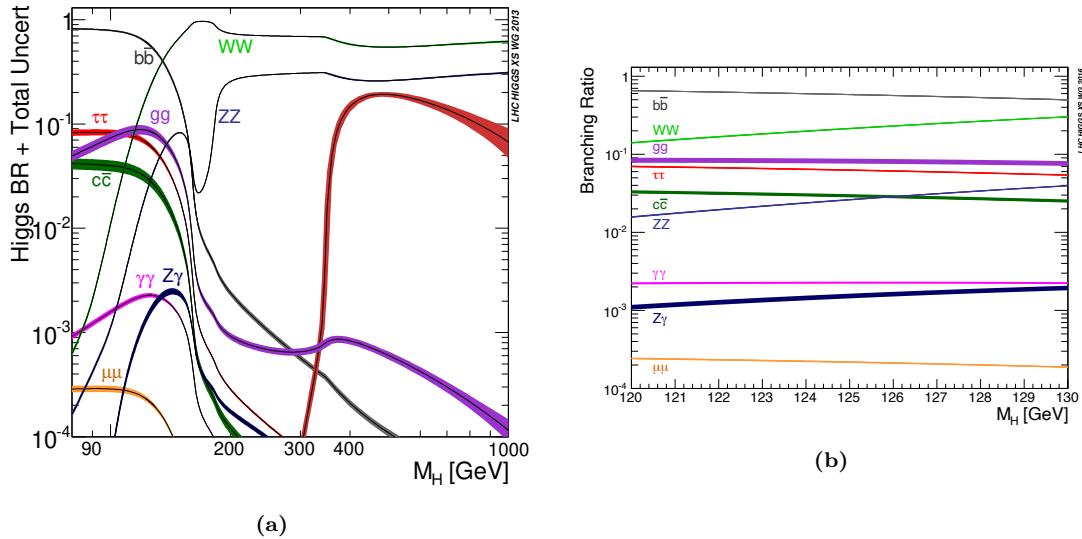


Figure 2.9: SM Higgs-boson-decay branching ratios as a function of m_H at $\sqrt{s} = 13$ TeV [83]. In (a) the BRs are shown in a Higgs-boson mass range $m_H \in (90, 10^3)$ GeV. In (b) only values of m_H around the measured one are shown. Looking at (a) it can be seen that if the Higgs boson weighed just about 60 GeV more there would have been only two relevant decay modes, $H \rightarrow WW^*$ and $H \rightarrow ZZ^*$. On the other hand, if Higgs had been just 20 GeV lighter, these two channels would have been very difficult to observe.

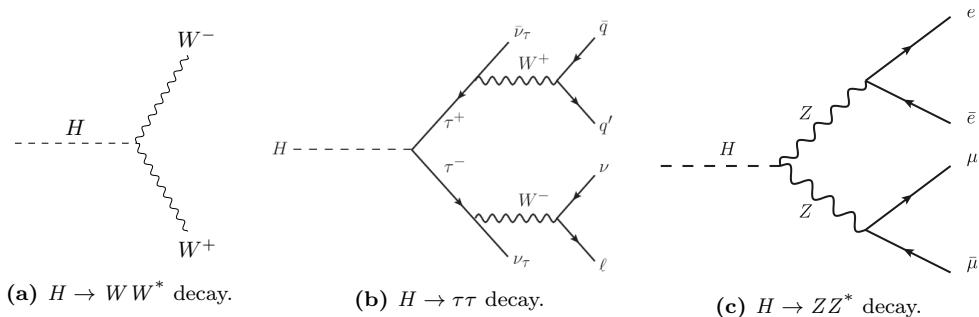


Figure 2.10: Representative LO Feynman diagrams for the Higgs-boson decay into a pair of (a) W bosons (b) tau leptons and (c) Z bosons. The three decay modes are taken into account for the associated tHq production described in this thesis. In (b), the τ^+ is decaying to quarks which form hadrons, therefore, it is referred to as hadronic tau. In contrast, the τ^- in (b) is a leptonically-decaying tau. For the diagram in (c), the Higgs boson decays into a four light-flavoured-leptons final state.

2.3 The interplay between the top quark and the Higgs boson

The couplings of the Higgs boson to the SM particles are found to be uniquely determined by the masses of these particles. Being this strength proportional to the mass in the case of fermions and the squared mass for the bosons. Figure 2.11 presents the coupling–mass relationship of the Higgs boson with other SM particles. Since the top quark is the most massive particle, the Yukawa coupling between the top quark and the Higgs boson, y_t is expected to be the strongest among all fermions and, hence, its study is of crucial importance, as it is discussed in References [84, 85] and developed in the succeeding sections. The Yukawa coupling is expected to be of the order of the unity:

$$y_t = \frac{\sqrt{2}m_t}{v} = 2^{3/4}G_F^{1/2}m_t = 0.995 \simeq 1.$$

This value is larger than the value of the couplings of the other quarks. For comparison $y_b \simeq 0.025$ and $y_c \simeq 0.007 \gg y_{s,d,u}$.

The production of a pair of top quarks along with a Higgs boson ($t\bar{t}H$) allows to measure the absolute value of y_t . This process has the advantage of being the leading mechanism to produce the Higgs together with the quark top. Section 2.3.2 discusses the $t\bar{t}H$ production.

Having a much lower cross-section than $t\bar{t}H$, the Higgs-boson production alongside a single top quark (tH) brings valuable information, especially regarding the sign of the Yukawa coupling. Note that the sign of the Yukawa coupling is not a physical property by itself, but the relative sign compared to the coupling of the Higgs boson to gauge bosons¹⁰ is indeed physical [84]. This production mechanism is discussed in more detail in Section 2.3.3.

2.3.1 CP properties in top-quark–Higgs-boson interactions

The CP properties of the Yukawa coupling of the Higgs boson to the top quark can be probed through the associated production of these two particles. While SM predicts the Higgs boson to be a scalar boson ($J^{CP} = 0^{++}$), the presence of a $J^{CP} = 0^{+-}$ pseudoscalar admixture has not been excluded yet. This pseudoscalar would introduce a second coupling to the top quark. Finding a CP-odd contribution would be a sign of physics beyond the SM and could account for the imbalance between matter and antimatter in the universe [88].

¹⁰The coupling of the Higgs boson to the gauge bosons is taken as positive.

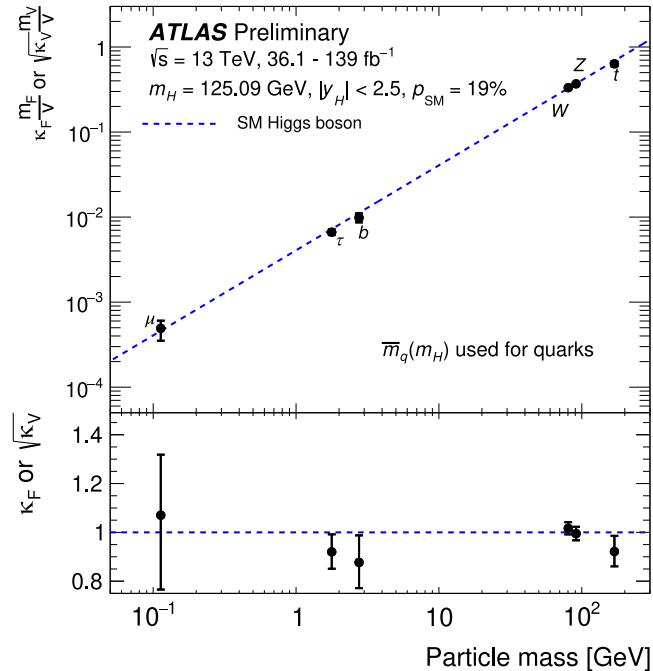


Figure 2.11: Best fit values of the coupling of the Higgs boson to fermions (μ , τ , b , t) and bosons (W , Z) as a function of the mass of the particles [86, 87]. The dotted line indicates the SM prediction and the uncertainty bars represent 68% confidence level (CL) intervals for the measured parameters. For the κ_μ , the light uncertainty bars indicate a 95% CL interval. The lower panel shows the ratio between the measured values and their SM predictions.

The production rates of $t\bar{t}H$ and tH depend on the y_t coupling. The latter is especially sensitive to y_t deviations from the SM as it is described in Section 2.3.3. As already mentioned, the presence of a CP-mixing in y_t would also affect the ggF production and $H \rightarrow \gamma\gamma$ decay rates. As it is explained in Section 1.6.2, the existence of CP violation is one of the conditions needed to explain the matter-antimatter asymmetry [28].

2.3.2 The $t\bar{t}H$ process

The production of $t\bar{t}$ in association with a Higgs boson is one of the most important process to measure the strength of the Yukawa coupling ($|y_t|$), which is crucial to understanding the origin of the fermion masses. Detecting a deviation from the SM prediction for $\sigma(t\bar{t}H)$ could indicate the presence of new physics that violates the CP symmetry. But, as Figure 2.14 illustrates, the $\sigma(t\bar{t}H)$ has a non-injective dependence with the Yukawa mixing angle.

From the phenomenology point of view, the calculations for the $t\bar{t}H$ production cross-section at $\sqrt{s} = 13$ TeV can be calculated at next-to-leading order (NLO) + NNLL accuracy [89]:

$$\sigma_{\text{NLO+NNLL}}^{\text{pred}}(t\bar{t}H) = 486.4^{+29.9}_{-24.5} \text{ fb.}$$

This calculation depends on the chosen scales for the soft and hard processes but it gives an idea of the order of magnitude for this process. The uncertainties are estimated through variations of the factorisation and renormalisation scales. This process constitutes a relevant background in the analysis.

The first associated production of a Higgs boson with a pair of top quarks was observed in 2018 by ATLAS [90] and CMS [91] collaborations. This process marked a significant milestone for the field of high-energy physics because it helped establish the first direct measurement of the tree-level coupling of the Higgs boson to the top quark, which is in agreement with the SM expectation.

The $t\bar{t}H$ process has been studied by ATLAS and CMS previously not only with Run 1 dataset at $\sqrt{s} = 7$ TeV and 8 TeV [92, 93] but also with Run 2 dataset at $\sqrt{s} = 13$ TeV [90, 91], where the cross-section was expected to be increased by a factor of four.

The ATLAS combined measurement for the $t\bar{t}H$ process using 79.8 fb^{-1} and considering $b\bar{b}$, W^+W^- , $\tau^-\tau^+$, $\gamma\gamma$ and ZZ Higgs-boson decay channels present the signal strength [90]:

$$\mu_{t\bar{t}H} = \frac{\sigma(t\bar{t}H)}{\sigma^{\text{SM}}(t\bar{t}H)} = 1.32 \pm 0.13 \text{ (stat.)} {}^{+0.17}_{-0.15} \text{ (syst.)}.$$

For CMS, the $t\bar{t}H$ signal strength has been measured with a dataset corresponding to a luminosity of 137 fb^{-1} at $\sqrt{s} = 13$ TeV. For the $\gamma\gamma$ Higgs-boson-decay channel [94]

$$\mu_{t\bar{t}H}^{\gamma\gamma} = 1.14^{+0.36}_{-0.29}$$

and for the W^+W^- , $\tau^-\tau^+$, $\gamma\gamma$ and ZZ Higgs-boson-decay channels with multileptonic final states [95]

$$\mu_{t\bar{t}H}^{\text{MultiLep.}} = 0.92 \pm 0.19 \text{ (stat.)} {}^{+0.17}_{-0.13} \text{ (syst.)}.$$

As can be seen, the statistical uncertainty of these results is sometimes dominating the measurements. Therefore, an analysis with more data would be very beneficial to have more precise results.

2.3.3 The tH process

The associated tH production takes place via three different types of processes. Firstly, the t -channel, where the Higgs boson couples to a top quark or W boson. The search presented in this thesis is focused on the tH t -channel production, which is usually referred to as tHq production. Figures 2.12a and 2.12b, present the LO diagrams for the tHq production, in the former the Higgs boson couples to the top quark and in the latter to the W boson.

The second most important production mode is via the tW process, in which the Higgs boson couples to a top quark and is produced alongside a W boson. This production mechanism is presented in Figure 2.12c.

Finally, the third tH production mechanism corresponds to the single-top-quark production in the s -channel but with a Higgs boson being radiated either from the top quark or the W boson.

All three processes have a much smaller cross-section than the main Higgs production channels that were discussed in Section 2.2.2. However, the tH modes yield a unique feature that makes them fascinating: they are simultaneously sensitive to the sign and magnitude of the Higgs-boson coupling to both the top quark, y_t , and the weak bosons, g_{HVV} . In Section 2.3.3.1 the three mentioned tH production mechanisms are discussed in the following order: tHq , tWH and s -channel.

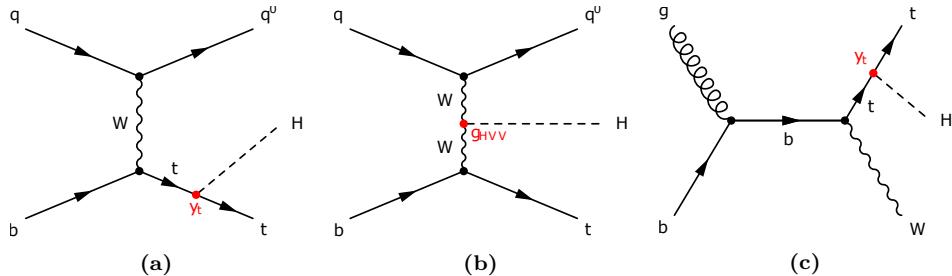


Figure 2.12: Representative LO Feynman diagrams for the t -channel tH associated productions. In Figures (a) and (b) the tHq process is presented. Here the Higgs boson couples either to the top quark (a) or the W boson (b). On (b), g_{HVV} is the coupling of the Higgs boson to the vector bosons. On (c), the tWH production is shown.

2.3.3.1 The tH production modes

In this section, the three tH production modes are discussed. This categorisation in tHq , tWH and s -channel tH is the same as for the single-top-quark processes (see Section 2.1.2.2), since the tH production is, basically, a single-top-

quark process in which a Higgs boson is radiated either from the W boson or the top quark.

The tHq production

The t -channel tH production mode resembles the ones described in Section 2.1.2.2. These are classified in 4FS and 5FS as it is done for the single-top-quark case. In the 4FS the initial partons are the spectator quark and the gluon that decays into $b\bar{b}$ or $t\bar{t}$. Meanwhile, in the 5FS there are no gluons involved and the spectator quark interacts with an incoming b quark [96].

The different Feynman diagrams in Figure 2.13 correspond to the 4FS while Figures 2.12a and 2.12b represent the tH production in the 5FS. For the 4FS modes, the diagrams in which the gluon decays to a top-quark pair (2.13c, 2.13d and 2.13e) contribute less than the ones in which it does to a $b\bar{b}$ (2.13a and 2.13b) because at LHC energies it is easier for the gluon to decay into $b\bar{b}$ than into $t\bar{t}$ since it is very unlikely that a gluon has enough energy to produce two particles with such a large mass. The NLO cross-section for the tHq process at $\sqrt{s} = 13$ TeV is [97]:

$$\begin{aligned}\sigma_{\text{NLO}}^{\text{pred}}(tH, t\text{-channel}) &= 47.64 \pm 9.7(\text{scale + FS})^{+2.9\%}_{-3.1\%} (\text{PDF} + \alpha_s + m_b) \text{ fb} \\ \sigma_{\text{NLO}}^{\text{pred}}(\bar{t}H, t\text{-channel}) &= 24.88 \pm 10.2(\text{scale + FS})^{+3.5\%}_{-2.6\%} (\text{PDF} + \alpha_s + m_b) \text{ fb}.\end{aligned}$$

The calculation of the $\sigma_{\text{NLO}}^{t\text{-channel}}$ depends on the choice of normalisation (μ_R) and factorisation (μ_F) scales. The numbers given here correspond to $\mu_R = \mu_F = (m_H + m_t)/4$. Regarding the uncertainties, these correspond to the parton-distribution-function (PDF) uncertainty and the μ_R and μ_F dependence, the 4FS and 5FS dependence, and m_b uncertainty. Combining the tH and $\bar{t}H$ contributions results in [97]:

$$\sigma_{\text{NLO}}^{\text{pred}}(tH + \bar{t}H, t\text{-channel}) = 72.55 \pm 10.1(\text{scale + FS})^{+3.1\%}_{-2.4\%} (\text{PDF} + \alpha_s + m_b) \text{ fb}.$$

For tHq and single-top-quark production at colliders, the 5FS calculations are easier than the 4FS due to the lesser final state-multiplicity and smaller phase space. This is why in the 5FS the single-top-quark production is known at NNLO while the 4FS is done only for NLO [97]. Another advantage of the 5FS is that the t -channel, s -channel and associated tWH productions do not interfere until NNLO. Contrary, in the 4FS, the t -channel at NLO and the s -channel at NNLO can interfere. Nevertheless, these interferences are very small and can be neglected if the aim is to evaluate the dominant t -channel cross-section [97].

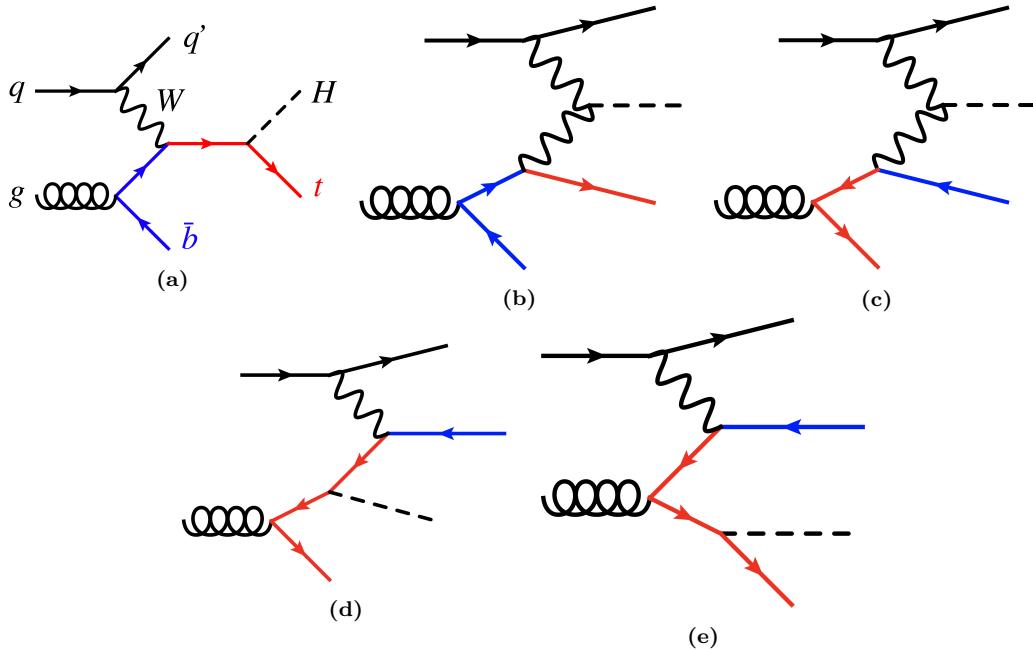


Figure 2.13: LO Feynman diagrams for t -channel tH production in the 4FS. The red line represents the top quark while the blue is the b quark.

Another feature of the 4FS is that it is assumed that the energy scale of the hard process (Q) is not much higher than the bottom-quark mass, which is also significantly larger than the QCD scale (Λ_{QCD}). Therefore, the model is limited to $Q \geq m_b \gg \Lambda_{\text{QCD}}$. When $Q \gg m_b$ inaccuracies appear. In contrast, the 5FS assumes $Q \gg m_b$. In practice, the bottom mass is set to zero in 5FS to simplify calculations [97].

The work developed in this thesis is focussing on this production mode. The associated production of a Higgs boson and top quark with an additional light-flavour quark (q) and b quark. This light-flavour quark is usually referred to as the spectator quark and it is expected to be produced preferably in the forward direction. The reason is that the q was contained within the initial parton and, therefore, it continues in the direction of the beam.

The characterisation of the tHq process using the Lagrangian formalism is presented in Section 2.3.3.2. The sensitivity of this production mode to determine the presence of a CP-violating phase in the Yukawa interaction between the Higgs boson and the top quark is discussed in Section 2.3.3.3.

The tWH production

The production of the Higgs boson in association with a top quark and W boson (tWH) is a process that can be easily defined at LO accuracy in QCD and in the 5FS, i.e. through the partonic process $gb \rightarrow tW(H)$ [98]. At NLO in QCD, the tWH process interferes with the LO $t\bar{t}H$ production. This arises from the $gg \rightarrow tWb(H)$ with a resonant \bar{t} interfering with $gg \rightarrow t\bar{t}(H)$. This makes the tWH process difficult to distinguish from the $t\bar{t}H$, which has a cross-section of one order of magnitude larger. At $\sqrt{s} = 13$ TeV, its cross-section is predicted to be

$$\sigma_{\text{NLO}}^{\text{pred}}(tWH + \bar{t}WH) = 15.2^{+4.9\%}_{-6.7\%}(\text{scale}) \pm 6.3(\text{PDF} + \alpha_s) \text{ fb}.$$

As well as for the tHq processes, the cross-section of the tWH production is very sensitive to departures from the SM in terms of CP violation in y_t [98] since the total rate can increase by more than an order of magnitude due to constructive interference effects [84, 97].

In the tHq search, this process constitutes a small source of background. In the scenario of a more inclusive tH search, the tWH processes could be included as a signal alongside tHq . This would also be beneficial to set more precise constraints in the measurement of y_t because the tWH process is also sensible to this magnitude, as it is presented in Section 2.3.3.2. By using tHq and tWH as signal, we would have a more comprehensive view of the associated top-quark–Higgs-boson production.

The tH production in the s -channel

The s -channel contribution to the total cross-section of the tH processReference is very small. Additionally, this channel contributes at low p_T and, since a minimum p_T is required, the s -channel events are suppressed. For these two reasons, this channel plays a less important role in the associated top-quark–Higgs-boson production and its inclusion on a more inclusive tH search would not increase the precision of the analysis.

The NLO total cross-section for the tH process via the s -channel at $\sqrt{s} = 13$ TeV is [97]:

$$\sigma_{\text{NLO}}^{\text{pred}}(tH + \bar{t}H, s\text{-channel}) = 2.812^{+1.6\%}_{-1.2\%}(\text{scale}) \pm 1.4(\text{PDF})^{+0.3\%}_{-0.5\%}(\alpha_s) \text{ fb}.$$

2.3.3.2 Characterisation of the Higgs boson in the tHq production

The characterisation model used in this thesis to describe the associated tH production is the one described in Reference [97]. It considers a spin-0 particle

with a CP-violating Yukawa interaction with the top quark, X_0 . This X_0 particle couples to both scalar and pseudoscalar fermionic densities, and its interaction with the W boson is the one described by the SM. The reason to call this particle X_0 instead of H is because its description does not correspond to the typical realisation of the Higgs boson but, in practice, we are referring to the Higgs boson. Within this model, the term in the effective Lagrangian that describes the Higgs-boson-top-quark Yukawa coupling below the EW SB scale is:

$$\mathcal{L} = -\bar{\psi}_t [\cos(\alpha)\kappa_{Htt}g_{Htt} + i\sin(\alpha)\kappa_{Att}g_{Att}\gamma^5]\psi_t X_0 ,$$

where ψ_t and X_0 represent the top quark and the Higgs boson fields, respectively and α is the CP mixing phase. The κ_{Htt} and κ_{Att} are real-dimensionless-rescaling parameters. Finally, $g_{Htt} = g_{Att} = \frac{m_t}{v} = \frac{y_t}{\sqrt{2}}$. Thus, the previous Lagrangian can be rewritten as:

$$\mathcal{L} = -\frac{y_t}{\sqrt{2}}\bar{\psi}_t [\cos(\alpha)\kappa_{Htt} + i\sin(\alpha)\kappa_{Att}\gamma^5]\psi_t X_0 . \quad (2.1)$$

The advantage of this Higgs-boson-top-quark parametrisation is that it is simple to interpolate between the CP-even ($\cos(\alpha) = 1$ and $\sin(\alpha) = 0$) and the CP-odd (i.e. $\cos(\alpha) = 0$ and $\sin(\alpha) = 1$) scenarios. The SM coupling corresponds to the CP-even: $\mathcal{L} = -\frac{y_t}{\sqrt{2}}\bar{\psi}_t\psi_t X_0$.

The proposed Lagrangian for the interaction of the Higgs boson with a top quark is based on considering the SM an effective field theory (EFT) applicable only up to energies not exceeding a certain scale Λ [99].

Figure 2.14 shows the cross-section for the tX_0 production in the t -channel as a function of the CP-mixing angle. For comparison, the $t\bar{t}X_0$ is also included. In the same way that tX_0 models the tHq process, $t\bar{t}X_0$ models the $t\bar{t}H$ process. The uncertainty band is derived from the choice of scale and the FS dependence. The values of κ_{Htt} and κ_{Att} in Figure 2.14 are set to reproduce the SM expectation for the gluon fusion cross-section.

Upon examining Figure 2.14, it becomes immediately apparent that the $t\bar{t}H$ cross-section exhibits symmetry around a CP angle of $\alpha = \pi/2$. This implies that by measuring $\sigma(t\bar{t}H)$ it would not be possible to discriminate between the CP-odd and CP-even scenarios. However, for the tHq production, this degeneracy is removed by the interference of its LO diagrams as described in Section 2.3.3.3.

2.3.3.3 tHq sensibility to y_t

As already mentioned, the tHq production is among the few LHC processes that are sensible to the relative size and phase between the couplings of the top

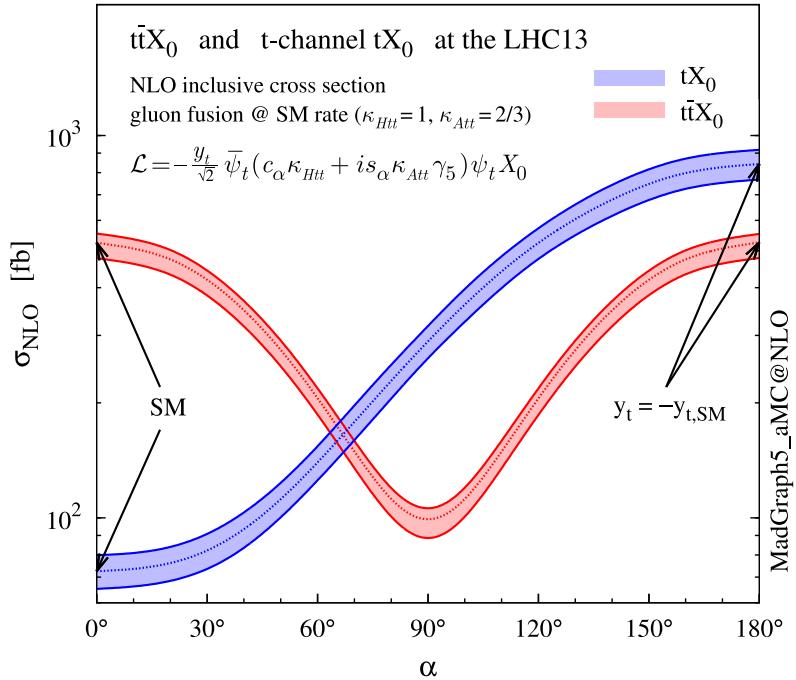


Figure 2.14: NLO cross-section as a function of the CP-mixing angle for t -channel tX_0 and $t\bar{t}X_0$ at $\sqrt{s} = 13$ TeV. The X_0 represents a general CP-violating Higgs boson. Note that while the $t\bar{t}H$ cross-section degenerates under the transformation $y_t \rightarrow -y_t^{SM}$, this is not the case for $\sigma(tH)$, which is sensible to α in an injective way.

quark to the Higgs boson, and the Higgs boson to the gauge bosons. The other mechanisms capable of determining this relative sign are $H \rightarrow \gamma\gamma$ and $gg \rightarrow ZZ$.

For the tHq , this ability is due to the fact that in the SM the tHq production where the H couples to the W (Figure 2.12 (b)) interferes destructively with those in which the H is radiated from the top quark (Figure 2.12 (a)). The cross-section is proportional to the square of the matrix element, \mathcal{M} , and if there are several diagrams for the same process, the matrix elements have to be summed before squaring leading to interference terms. For the tHq production:

$$\sigma_{tHq} \propto |\mathcal{M}_{qq \rightarrow tHq}|^2 = |\mathcal{M}_{qq \rightarrow tHq_{WH}} + \mathcal{M}_{qq \rightarrow tHq_{tH}}|^2. \quad (2.2)$$

When squaring the scattering amplitude, the destructive interference¹¹ term decreases the σ_{tHq} . This behaviour makes the tHq cross-section exceptionally sensitive to the departures of y_t from the SM predictions. Typically, the destructive interference yields a reduction in the rate as compared to the contribution from each diagram by about an order of magnitude [100]. Therefore, in the presence of

¹¹Destructive interference refers to the situation where the relative sign between $\mathcal{M}_{qq \rightarrow tHq_{WH}}$ and $\mathcal{M}_{qq \rightarrow tHq_{tH}}$ is negative.

non-SM new physics, a positive relative sign between the y_t and the g_{HVV} couplings would imply that the amount of tHq events recorded should increase a factor of ~ 13 over the SM expectations, surpassing the number of expected events from $t\bar{t}H$ production [101].

This can be seen in Figure 2.14. In contrast to the cross-section for $t\bar{t}H$, which degenerates ($\sigma(t\bar{t}H, y_t) = \sigma(t\bar{t}H, -y_t)$), the $\sigma(tHq)$ increases with the CP-mixing angle.

2.3.3.4 Previous results from ATLAS and CMS Collaborations

To gather the necessary information, the widest campaign of measurements has to be undertaken, including as many possible decay modes. In this context, the scope of this thesis is the study of the production tHq with a final state characterised by two light leptons (ℓ), i.e. electrons (e^\pm) or muons (μ^\pm), and one hadronically decaying tau lepton (τ_{had}). This signature is denoted by $2\ell + 1\tau_{\text{had}}$.

The tH production has already been searched for at LHC by both ATLAS and CMS Collaborations. A summary of the results obtained so far is presented in Table 2.1. Regarding the measure of y_t , analyses on the $t\bar{t}H$ processes already observed the Yukawa coupling between the Higgs boson and the top quark [90] but this process is only sensitive to the absolute value of the coupling. The current constraints on the value of y_t are $-0.9 < y_t < -0.7$ or $0.7 < y_t < 1.1$ at 95% CL [102].

Analysis	Luminosity (fb $^{-1}$)	Collaboration	μ_{tH}
tH (2018)[103]	35.9	CMS	limit ~ 25 (obs.) 12(exp.)
$t\bar{t}H/tH$ multilepton (2019)[102]	137	CMS	$5.7^{+2.8}_{-2.7}$ (stat.) ± 3.0 (syst.)
$t\bar{t}H/tH, H \rightarrow \gamma\gamma$ (2020)[104]	139	ATLAS	$0.85^{+3.13}_{-2.21}$ (stat.) $^{+0.97}_{-0.98}$ (stat.)
$t\bar{t}H/tH$ (2023)[105]	138	CMS	limit _{95CL} = 14.6(obs.) 19.3 $^{+9.2}_{-6.0}$ (exp.)

Table 2.1: Current results for the tHq process by the ATLAS and CMS Collaborations at $\sqrt{s} = 13$ TeV. The μ_{tH} is the ratio between the cross-section measured in the experiment and the reference cross-section given by the SM prediction. These three analyses target the tH process as the combination of tHq and tWH . The s -channel tH is neglected because of its small cross-section.

The analysis in first row in Table 2.1 uses 35.9 fb $^{-1}$ and targets the $H \rightarrow WW^*$, $H \rightarrow \tau\tau$, $H \rightarrow ZZ^*$, $H \rightarrow \bar{b}$ and $H \rightarrow \gamma\gamma$ decay modes [103] and assumes $y_t = y_t^{\text{SM}}$. The fiducial¹² cross-section is limited to 1.9 pb, corresponding to 25 times the expectation of the SM. The result presented in the second row of Table 2.1 is

¹²Refers to a specific portion of the detector where the instrument's response is known and trustworthy.

obtained by targeting multileptonic final states in the $H \rightarrow WW^*$, $H \rightarrow \tau\tau$ and $H \rightarrow ZZ^*$ Higgs-boson-decay channels [102]. The measured cross-section for the tH production is $\sigma_{tH} = 510 \pm 200$ (stat.) ± 220 (syst.) fb. For the third result shown in Table 2.1, the analysis measures the signal strength for that tH . Finally, the last study presents an upper limit at 95% CL, which has been obtained assuming $\mu_{t\bar{t}H} = 1$. The data in all the studies favour a positive value of y_t .

The primary objective of this thesis is to directly search for the tHq production. If this objective remains unfulfilled, the analysis will establish an upper boundary for the production cross-section. Notably, this analysis is among the first studies of the tHq process within the ATLAS Collaboration. Although it is not encompassed within this thesis, the analysis conducted herein possesses the potential to establish novel boundaries on y_t .

Chapter 3

The ATLAS experiment at the LHC of CERN

*Las cebollas tienen capas,
los ogros ATLAS tiene capas.*
—SHREK (2001)

The work developed in this thesis is framed in the context of the ATLAS detector [106], a general-purpose particle physics detector that records events arising from collisions within the LHC, the most powerful particle accelerator built to date. This experimental setup is located at CERN, one of the world’s premier centres for scientific inquiry.

This chapter is devoted to the introduction of the CERN laboratory and the description of the technical design of the LHC, its experiments and the ATLAS detector. In Section 3.1, the CERN organisation is presented through an overview of its history, its achievements and some of the most relevant research projects carried out currently. Through Section 3.2, the essential technical aspects of the LHC machine design are covered. The distribution and functioning of the accelerator complex and the main experiments conducted at LHC are summarised as well. Finally, in Section 3.3, a full overview of the different components of the ATLAS detector is provided, presenting the specific features of each part.

3.1 CERN

The European Organization for Nuclear Research, known as CERN, is the largest particle physics laboratory in the world. The convention establishing CERN was signed in 1953 and ratified a year later. Its name is derived from the French acronym *Conseil Européen pour la Recherche Nucléaire*, which was the provisional body designated in 1952 to foster fundamental physics research in Europe, and the acronym has been maintained until the foundation of CERN. Initially formed by 12 member states, now it has 23 member states¹ and many non-European countries involved in different ways such as associate members, partners and observers [109].

The main site of the laboratory is located at Meyrin, a municipality of the Canton of Geneva (Switzerland), at the Franco–Swiss border. There are other sites in the vicinity of the main site, being the most relevant the one located at Prévessin-Moëns (France).

Since its beginning, the objective of CERN has been helping to uncover what the universe is made of and how it works. CERN started its first accelerator, the Synchrocyclotron, in 1957 and observed the pion decay into an electron and antineutrino for the very first time [110]. Thereafter, one of the most significant achievements made by the CERN experiments was the discovery in 1973 of neutral currents in the Gargamelle bubble chamber located in the Proton Synchrotron (PS) [111]. This milestone provided indirect evidence of the existence of the Z boson and in fact, a decade later, in 1983, CERN announced the discovery of the Z and W bosons [112] by the UA1 and UA2 experiments within the Super Proton Synchrotron (SPS). Thanks to this achievement, the Nobel Prize in Physics in 1984 was awarded to C. Rubbia and S. van der Meer for their contribution to the discovery of subatomic particles. This significant accomplishment marked the first recognition of the scientific program conducted at CERN.

Other major successes of CERN were the determination of the number of light-flavour neutrino families at the Large Electron–Positron (LEP) Collider in 1995 [113] and the creation for the very first time of antihydrogen atoms in 1995 at the PS210 experiment [114]. More crucial accomplishment followed such as the discovery during the 1990’s of CP violation by the NA31 [115] and NA48 experiments [116]. Later on, in 2012, the discovery of the Higgs boson was achieved by the ATLAS and CMS experiments [77, 78], which was a fundamental probe of the robustness of the SM as described in Section 2.2.1. More recently, in 2015, a state consistent with a pentaquark particle was observed at the LHCb experiment [117].

¹Spain joined CERN in 1961 and presented its withdrawal eight years later. In 1983 Spain re-joined the organisation [107, 108].

The collider-based experiments of LHC are described in more detail in Section 3.2. Apart from these, fixed-target experiments, antimatter experiments and experimental facilities make use of the LHC injector chain. The main fixed-target experiments at CERN are the Antiproton Decelerator (AD) [118] for slowing anti-protons for the antimatter factory [119] and the On-Line Isotope Mass Separator (ISOLDE) facility for short-lived ions [120]. Other relevant experiments carried out at CERN are the Advanced Proton Driven Plasma Wakefield Acceleration Experiment (AWAKE) [121] or the Alpha Magnetic Spectrometer (AMS) [122].

3.2 Large Hadron Collider

The LHC is the largest² and most powerful particle accelerator ever built [123]. In 1991, the LHC was proposed with the purpose of searching for the elusive Higgs boson [124, 125]. Finally, after several years of planning and construction, on September 10 2008, a beam of protons was successfully directed into the LHC pipes for the first time.

The LHC is a circular hadron accelerator with a circumference of 27 km that is located where once was the LEP³ collider tunnel. Circular accelerators are more space-efficient than linear ones due to their ability to speed up particles with less physical space. They simultaneously ramp up opposite charge beams with a single magnetic field, with bending power given by $p = 0.3qBr$, where p is momentum, q is particle charge ($q = 1$ for protons), B is the magnetic field, and r is accelerator radius [127].

3.2.1 Machine design

A summary of the main parameters of LHC for pp collisions is presented in Table 3.1. These parameters are shown for Run 1 (2011–2012), Run 2 (2015–2018), and Run 3 (2022–2025). The forecasted values for Run 4 after the High Luminosity (HL) LHC upgrade, scheduled to start in 2029, are provided too. The work developed in this thesis uses the data collected during Run 2 of LHC.

The LHC has two rings with ultra-high vacuum (to prevent collisions with gas molecules while moving through the accelerator) in which particle beams travel in opposite directions. As well as protons, it can collide heavy ions, in particular lead nuclei.

²Alongside LEP, which was equally large.

³LEP is the accelerator used by CERN from 1989 to 2000 [126]

Parameter	Design	Run 1	Run 2	Run 3	HL-LHC
Beam energy [TeV]	7	3.5 - 4	6.5	6.8	7
Centre-of-mass energy (\sqrt{s}) [TeV]	14	7 - 8	13	13.6	14
Bunch spacing [ns]	25	50	25	25	25
Bunch Intensity [10^{11} ppb]	1.15	1.6	1.2	up to 1.8	2.2
Number of bunches (n_b)	2808	1400	2500	2800	2800
Transverse emittance (ϵ) [μm]	3.5	2.2	2.2	2.5	2.5
Amplitude function at the interaction point (β^*)[cm]	55	80	30→25	30→25	down to 15
Crossing angle [μrad]	285	-	300→260	300→260	TBD
Peak Luminosity [10^{34} cm 2 s $^{-1}$]	2.1	0.8	2.0	2.0	5.0
Peak pile-up	25	45	60	55	150
Nominal magnetic field (B) [T]	8.73	4.16 - 7.76	7.73	8.73	8.73 / 12
Injection energy [GeV]			450		
Circumference length [km]			26.7		
Radius [km]			4.24		
Number of dipole magnets			1232		
Length of dipole magnets [m]			14.3		
Number of quadrupole magnets			395		
Total mass [tons]			27.5		

Table 3.1: Summary of main accelerator parameters for the LHC, showing the design values, and those used during Run 1, Run 2 and Run 3, as well as the expected parameters for Run 4 at the HL-LHC [41, 128, 129].

The beams in the LHC are made up of bunches of protons that collide every 25 ns. Each bunch contains approximately 1.1×10^{11} hadrons, being about 2500–2800 the maximum possible number of bunches that can be reached with the beam preparation method currently used [130]. The size of each bunch is approximately 25 cm [123].

The LHC tunnel lies between 45 m and 170 m below the surface on a plane inclined at 1.4% sloping towards the Léman Lake. The underground construction adds some shielding from outside interferences that could interact with the detectors and cause anomalous readings. Even 100 m underground, the cosmic rays can reach the detectors, so these are used to help calibrate them. The two rings are built under the *two-in-one* twin-bore superconducting magnet design [123] .

The LHC contains 1232 twin-bore dipole magnets to curve the trajectory of the particle beams. Dipoles are also equipped with additional multipole-lattice magnets (sextupole, octupole and decapole), which correct for small imperfections in the magnetic field at the extremities of the dipoles.

The Radio Frequency (RF) cavities (also known as resonators) allow radio waves to interact with passing particle bunches. The main role of the RF cavities is to keep the proton bunches tightly packed to ensure the required luminosity at the interaction point (IP). They also transfer RF power to the beam to accelerate it to the target energy.

At the insertion of the arc and straight sections, quadrupole magnets are installed to suppress the dispersion of particles. Acting as focal lenses, quadrupole magnets gather the particles together.

In total, there are more than 9000 magnets all over the LHC and more than 50 types of magnets are needed to make the particles circulate in their path without losing speed. The coils are made of niobium-titanium (NbTi) which is cooled to less than 2 K with superfluid helium to reach superconductivity.

3.2.2 Accelerator complex

To accelerate the proton beams, the existing CERN accelerator complex is used. These accelerators were, back in the day, the state of the art colliders and now they serve as injection system for the LHC. The path followed by the particle beams is presented in Figure 3.1. The accelerator complex consists of several machines interconnected to boost the beams until these reach the LHC.

The proton bunches are produced by ionising a gas of hydrogen atoms and then they are accelerated to a momentum of 50 MeV by the linear accelerator (LINAC2). After being produced, the beams enter the first circular accelerator, the Proton Synchrotron Booster (PSB), which has 630 m radius and increases the energy of the protons up to 1.4 GeV. Right after the PSB, the PS brings the particles up to 25 GeV. It is followed by the SPS, which raises the energy of the particles to 450 GeV. Then, the beam is injected into the LHC by two different transfer injection (TI) lines [131].

Heavy-ion collisions were included in the conceptual design of the LHC from an early stage and follow the same path to maximum acceleration as the protons. Lead ions are extracted from a source of vaporised lead and are initially accelerated by the Low Energy Ion Ring (LEIR).

3.2.3 LHC experiments

In the LHC four major experiments are carried out, each of them with its own detector (see Figure 3.2) and physics programme. Distributed along the collider as is shown in Figures 3.1, these highly sophisticated experiments are:

- **A Toroidal LHC ApparatuS (ATLAS)** [132]: It is a generic multi-purpose experiment for high luminosity. It studies pp collisions and investigates a wide range of physics, from the SM to the search for extra dimensions or dark matter. It has the dimensions of a cylinder, 46 m long, 25 m in diameter.

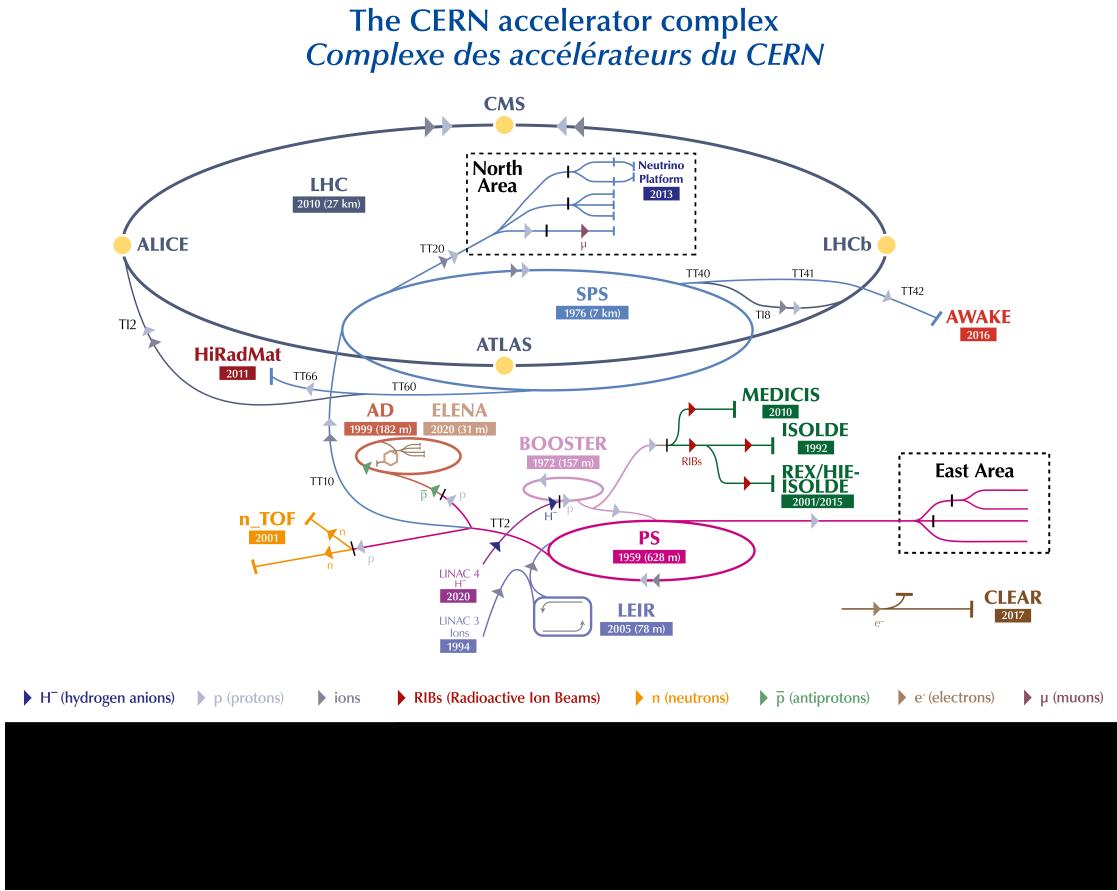


Figure 3.1: Scheme of CERN accelerator complex.

The ATLAS detector weighs 7×10^3 tonnes. Its design features excellent jet and missing transverse energy resolution, particle identification and flavour tagging and standalone muon measurements. ATLAS is covered in detail in Section 3.3 as it is the detector used to perform the analysis presented in this thesis.

- **Compact Muon Solenoid (CMS)** [133]: It is the other general-purpose experiment for high luminosity. CMS has the same objectives and goals as ATLAS but both its hardware and software designs are different. Even though CMS is smaller than ATLAS (21 m long, 15 m in diameter) it is much heavier, weighting 14×10^3 tonnes. The bulk of its weight is the steel yoke that confines the 4 T magnetic field of its superconducting solenoid. The design of CMS emphasises magnificent electron/photon energy and momentum resolution. The role of the coexistence of CMS and ATLAS is fundamental so

that the scientific community can verify and confirm the independent results found by these experiments.

- **Large Hadron Collider beauty (LHCb)** [134]: It is a lower luminosity⁴ experiment designed to study the small asymmetries between matter and antimatter through CP violation using rare decays of b -quark based hadrons. The detector is arranged as a succession of planar sub-detectors since most of the b -flavoured mesons follow the beam-pipe direction when created in the pp collision. LHCb delivers remarkable low-momentum track reconstruction and particle identification.
- **A Large Ion Collider Experiment (ALICE)** [135]: It is a low luminosity experiment in IR8 that focuses on QCD. The main feature of ALICE is a general-purpose detector that uses heavy-ion collisions to study matter interacting at extreme densities and temperatures, thus reproducing the quark-gluon plasma. This detector provides highly efficient track reconstruction in an environment full of heavy ions. Besides running with lead ions, the physics programme includes collisions with lighter ions, lower energy collisions and a dedicated proton–nucleus run.

Along the LHC machine, there are other experiments much smaller than ATLAS, CMS, LHCb and ALICE, typically sharing the cavern with the major projects. The most relevant among the minor experiments are LHCf [136], MATHUSLA [137], MilliQan [138], MoEDAL [139], TOTEM [140] and FASER [141].

3.2.4 LHC computing grid

The data collected by the different LHC experiments is stored, processed and, then, made available for all the researchers of each collaboration. This is possible thanks to the last piece of the LHC, its computing model and infrastructure: the World LHC Computing Grid (WLCG). It consists of several computing farms distributed around the world and interconnected.

Different types of computing centres are defined and these are classified into Tiers [142]:

- **Tier-0:** Facility located at CERN and responsible for providing prompt reconstruction and distributing a copy of the raw data to the Tier-1 centres.

⁴By lower luminosity it means that less collisions are delivered by the LHC the experiment.

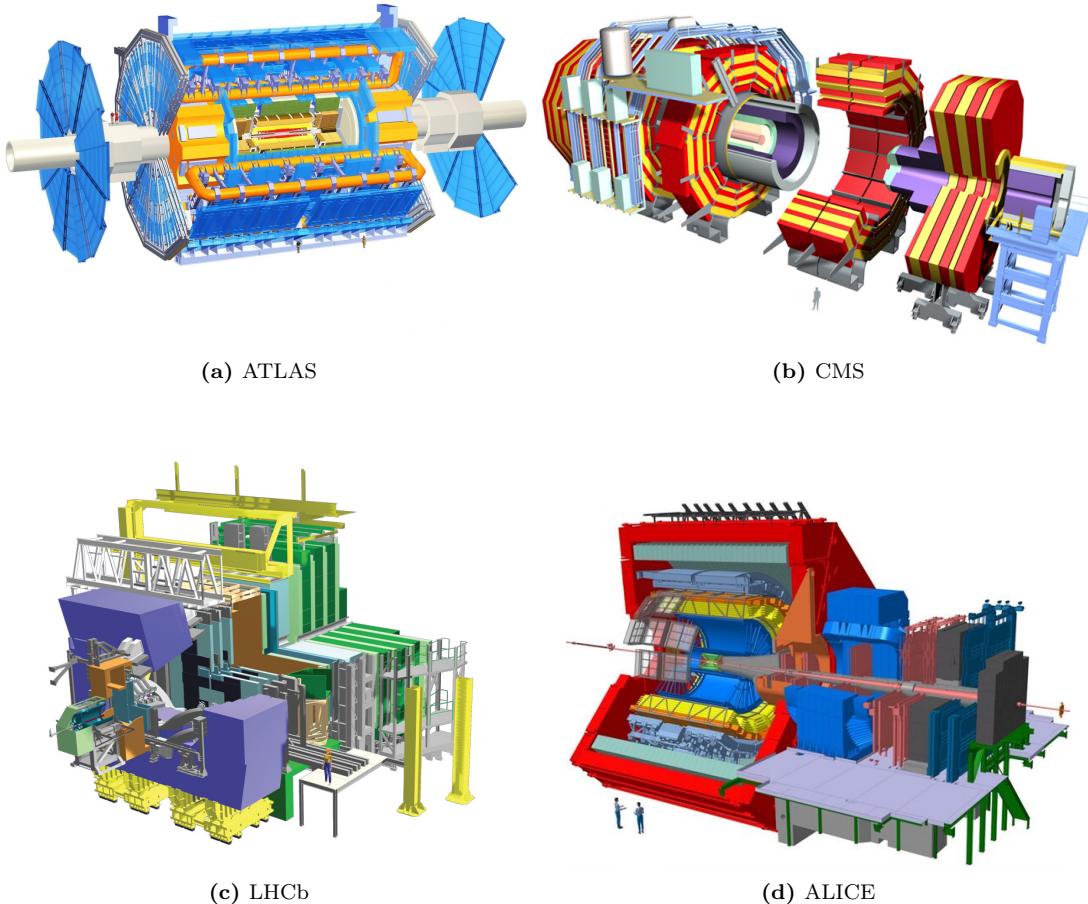


Figure 3.2: Scheme of LHC main experiments. Note that the images are not equally scaled.

- **Tier-1:** Facilities archiving the raw data permanently and providing the computational capacity for reprocessing and physical analysis. These also store the simulated and reprocessed data. Currently, 13 computer centres are serving as Tier-1, making data available to the Tier-2 centres [143].
 - **Tier-2:** Facilities typically located at universities and scientific institutes, that serve as storage and analysis facilities of processed data. The event simulations are also executed here.
 - **Tier-3:** Local computing resources.

This system provides near real-time access to LHC data worldwide. It deals with over two million tasks daily. These specifications make the LCG the most sophisticated system for data taking and analysis ever built for science.

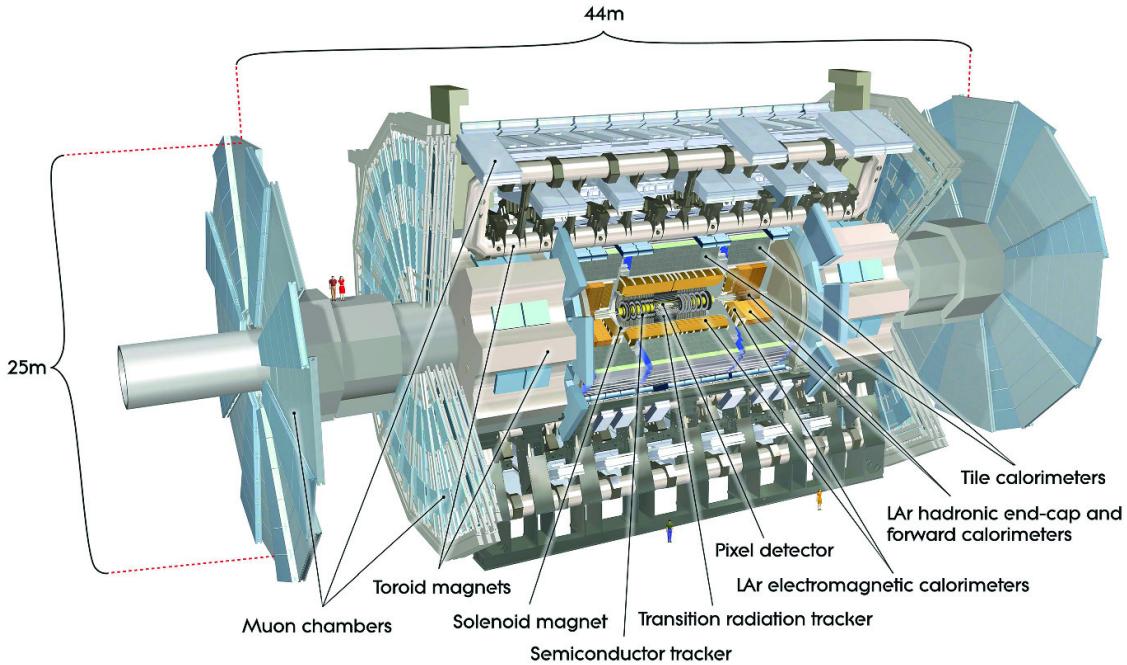


Figure 3.3: Simulated schematic view of the ATLAS detector [144].

3.3 The ATLAS detector

The ATLAS detector is the largest detector ever constructed for a particle collider. It is designed to record events of high-energy colliding particles at high luminosities. The thousands of millions of interactions that take place at the centre of the ATLAS detector are recorded and processed by the different sub-detectors, which are composed of more than 100 million electronic channels. Each ATLAS sub-detector is sensitive to a different type of particle and to different properties. Therefore, the layered structure allows for effective particle identification, as well as enables accurate measurements of energy and momentum. Figure 3.3 shows an overall layout of the ATLAS detector and identifies its different sub-detectors. In the picture, one can appreciate that the cylindrical shape of ATLAS is divided into two parts: the “barrel” and the two “end-caps”. In the barrel region, the sub-detectors are built as coaxial layers around the beam pipe. As one moves away from the axis, it finds the Inner Detector (ID), the solenoid magnet, the Electromagnetic (ECAL) and Hadronic (HCAL) calorimeters, and the Muon Spectrometer (MS) and the toroid magnet in the outermost layers. The technical details of these sub-detectors and the magnet systems are presented in Sections 3.3.2, 3.3.3, 3.3.4 and 3.3.5.

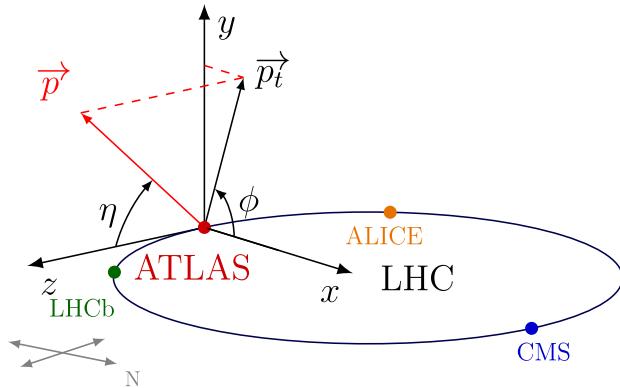


Figure 3.4: Coordinate system of the ATLAS detector in the context of LHC.

The ATLAS detector can explore a wide range of phenomena with high precision. It was designed to perform precision measurements of the SM parameters and the Higgs boson properties.

3.3.1 Coordinate system

The ATLAS detector uses a right-handed system with its origin at the IP where the collisions take place. On one side, there are the (x, y, z) cartesian coordinates. The x -axis is pointing towards the centre of the ring circumference, the y -axis is perpendicular to the plane defined by the LHC ring and it points to the surface, and the z -axis is defined by the direction of the beam. On the other side, it is more frequent to employ the cylindrical coordinates (r, ϕ, z) or the system defined by the azimuthal angle (ϕ) and the pseudorapidity, being $\eta = -\ln[\tan(\theta/2)]$ where θ is the polar angle⁵. The change in pseudorapidity $\Delta\eta$ is Lorentz invariant under boosts along the beam axis. The angular distance is measured in units of $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$, which is invariant under a boost along the z -axis. Figure 3.4 shows the coordinate system of ATLAS for both cartesian and cylindrical coordinates.

3.3.2 Inner Detector

The ID [132, 145, 146] is the closest sub-detector to the beam pipe. Its layout is shown in Figure 3.5. The charged particles follow a curved trajectory inside the ID due to the magnetic field of the bending magnet (described in Section 3.3.5).

⁵Defined as the angle between the particle three-momentum, \vec{p} and the positive direction of the beam axis.

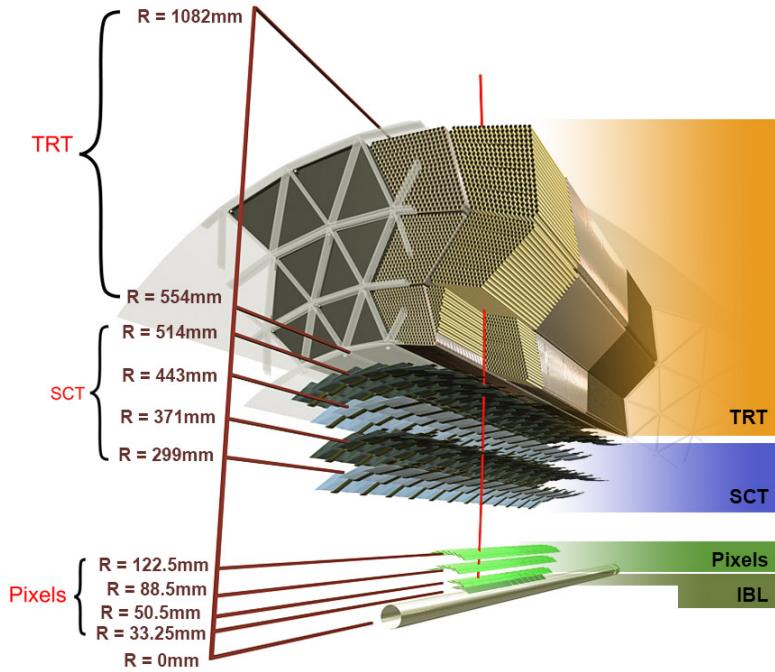


Figure 3.5: Barrel part of ID of the ATLAS experiment with the IBL, Pixel, SCT and TRT sub-detectors [147, 148].

The different pieces that comprise the ID record the hits that are later used to reconstruct the tracks of these particles with great accuracy allowing, thus, to measure its momentum. The geometric acceptance of the ID is $|\eta| < 2.5$. The ID provides p_T resolution of approximately $\sigma_{p_T}/p_T = 0.05\% \oplus 1\%$ ⁶. It is designed to provide excellent momentum resolution, pattern recognition and measurements of both primary and secondary vertex for charged particles above the p_T threshold (nominally 0.5 GeV).

The ID is composed of four sub-detectors: The Insertable B-Layer (IBL), the Pixel Detector, the Semiconductor Tracker (SCT) and the Transition Radiation Tracker (TRT).

Insertable B-Layer

The IBL [149] is the innermost component of the ID. It is located between the beam pipe and the Pixel Detector. Added after Run 1, it provides the closest-to-IP measurements. This improves the robustness and performance of the ATLAS tracking system. It plays a fundamental role in b -tagging efficiency because this

⁶The \oplus symbol means that the relative uncertainties are summed in quadrature.

tagging relies on precise vertex reconstruction. With a hit resolution of $8\text{ }\mu\text{m}$ in $r\text{-}\phi$ and $40\text{ }\mu\text{m}$ along z , the IBL covers the $|\eta| < 2.7$ and the entire ϕ range.

The IBL consists of pixel planar sensors and 3D pixel sensors [149, 150] mounted on 280 silicon pixel modules and arranged on 14 azimuthal staves.

Pixel Detector

The ATLAS Pixel Detector [151] is made up of four layers of silicon pixels of $50 \times 400\text{ }\mu\text{m}^2$ organised in modules. The barrel modules are arranged in three cylinders concentric to the beam axis and three concentric disks at both ends of the barrel. Each of these modules consists on a silicon pixel sensor bonded to the front-end electronic chips. It provides full coverage of the azimuthal angle ϕ and a pseudorapidity range of $|\eta| < 2.5$ as well as a resolution of $10\text{ }\mu\text{m}$ in $r\text{-}\phi$ and $115\text{ }\mu\text{m}$ in the z -axis.

Semiconductor Tracker

The SCT consists of 4088 modules tiling four coaxial cylindrical layers in the barrel region and two end-caps each containing nine disk layers, all of these surrounding the Pixel Detector. The SCT uses microstrip sensor technology. The reason to use microstrips instead of pixels is that the strips are more cost-effective than traditional pixels and a good spatial resolution can be obtained as well if the strips are arranged with an angular offset. Therefore, each SCT detector unit consists of two back-to-back silicon-microstrip planes with a relative angle of 40 mrad. Eight strip layers (i.e. four space points) are crossed by each track in the SCT providing valuable tracking information with a resolution of $17\text{ }\mu\text{m}$ in $r\text{-}\phi$ and $580\text{ }\mu\text{m}$ in the z coordinate. The SCT covers the entire ϕ range and up to $|\eta| < 2.5$.

Transition Radiation Tracker

The TRT is used in conjunction with the Pixel Detector and silicon micro strip (SCT). This part of the ID is formed by about 300000 straw tubes with 4 mm diameter filled with gas. The TRT relies both on the collection of both primary and secondary ionisation charges arising from the transition radiation to measure the track of charged particles. The tube surface functions as cathode while the wire in the centre as anode. When a charged particle passes through the Xe-based gas mixture in the tube, it ionises the gas and the freed electrons drift towards the anode, generating an electrical current. This detector provides a single hit resolution of $130\text{ }\mu\text{m}$ in $r\text{-}\phi$ but does not have sensitivity in z . The TRT also provides discrimination between electrons and pions since the latter generates a smaller signal than the former.

Table 3.2 summarises the main characteristics of the ID subdetectors.

Subdetector	Element size (μm)	Intrinsic resolution (μm)
IBL	50×250	8×40
Pixel	50×400	10×115
SCT	80	17×580
TRT	4000	130

Table 3.2: Overview of the main features of the ID subdetectors. The intrinsic resolution of the IBL, the Pixel and SCT is provided for both $r\text{-}\phi$ and z directions, while for TRT subdetector only along $r\text{-}\phi$.

3.3.3 Calorimeters

After the ID, the next layer of detectors in the ATLAS machine corresponds to the calorimeters (see Figure 3.6) [152]. Their purpose is to measure the energy of the particles (neutral or charged), as well as to help the ID to reconstruct the path followed by them. Most particles initiate a shower when they enter into the calorimeter. Either all or part of the energy of these particles is deposited in the device and measured. Most of calorimeters in particle physics are segmented transversely to provide information about the direction of the particles. Based on the shower shape, the longitudinal segmentation provides information for particle identification.

The ATLAS detector has two types of calorimeters depending on the type of particles that it is desired to detect: the electromagnetic calorimeter (ECAL), which measures the energy of electrons/positrons and photons, and the hadronic calorimeter (HCAL), which registers the energy of the strongly-interacting particles.

Calorimeters are typically categorised into two types: sampling and homogeneous. Sampling calorimeters are constructed from two types of materials (passive and active), while homogeneous calorimeters are made from a single type. Both the ECAL and HCAL are sampling calorimeters, which consist of alternating layers of different materials.

3.3.3.1 Electromagnetic calorimeter

The ECAL [152] absorbs the energy of the electrons (e^-), positrons (e^+) and photons (γ) covering a pseudorapidity range of $|\eta| < 1.475$ in the barrel. It is made of a lead absorber and the liquid Argon (LAr) as an active medium following

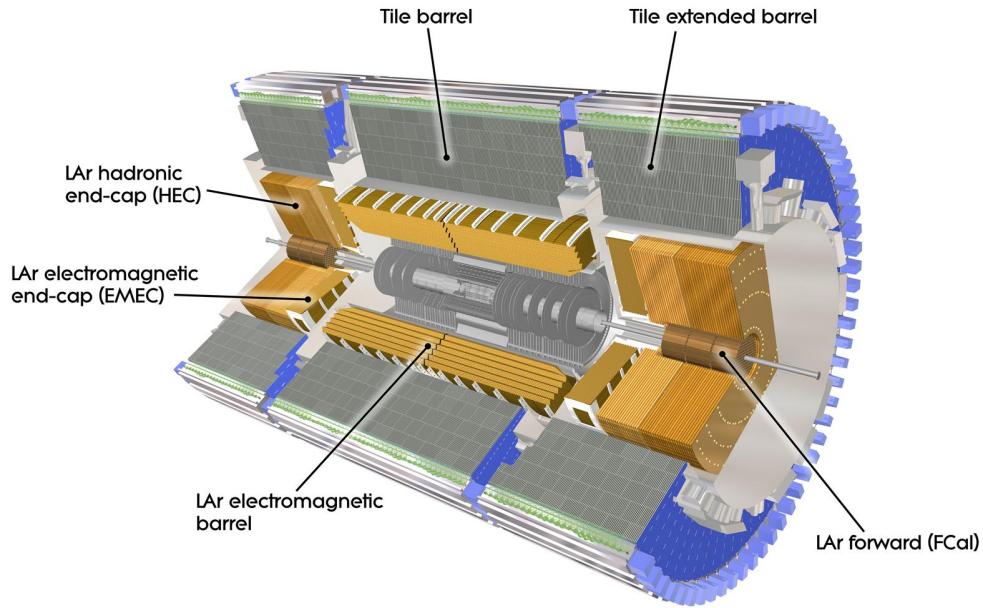


Figure 3.6: Computer generated image of the ATLAS calorimeters [153].

an accordion shape. The shower originated in the absorber layer ionises the LAr producing a measurable current proportional to the energy of the original particle. The LAr layer operates at 89.3 K to maintain a pressure of 1 bar⁷ [154]. The barrel part is split into two identical half-barrels separated by a small gap at $z = 0$. Each end-cap calorimeter is composed of two coaxial wheels that cover $1.375|\eta| < 3.2$.

The total amount of material in the ECAL corresponds to 25–35 radiation lengths⁸ over the entire η range.

The energy resolution of a calorimeter can be parametrised as:

$$\frac{\sigma_E}{E} = \frac{a}{\sqrt{E}} \oplus \frac{b}{E} \oplus c,$$

where a is the stochastic term, b the electronic noise c is a constant that includes detector instabilities and miscalibrations, and E is the energy. The stochastic term considers the statistical fluctuations in the shower detection. This term is larger for sampling calorimeters than for the homogeneous ones. The effect of a diminishes with increasing energy. The noise component b of energy resolution arises from

⁷The minimum to avoid the freezing of liquid argon in the region of the heat exchangers is 87.3 K.

⁸The radiation length is characteristic of each material and corresponds mean distance over which a high-energy electron lose its energy by a factor $e = 2.71828$ due to bremsstrahlung.

the noise of readout chain and varies with detector technique and readout circuit properties (e.g., detector capacitance, cables). The noise contribution also increases with decreasing energy of the incident particles. Finally, the constant term does not involve energy-dependent contributions and originates mainly from instrumental effects that cause variations in calorimeter response. For the ECAL, the resolution is:

$$\frac{\sigma_E}{E} = \frac{10\%}{\sqrt{E}} \oplus \frac{170 \text{ MeV}}{E} \oplus 0.7\%.$$

3.3.3.2 Hadronic calorimeter

The HCAL [152] is made of a sampling calorimeter of steel and plastic scintillator tiles covering the pseudorapidity region of $|\eta| < 1.7$ in the barrels. The end-caps are made of copper and LAr, covering $1.5 < |\eta| < 3.2$, and are embedded in the end-caps of the ECAL. This calorimeter uses 9800 electronic channels in the barrel and 5600 in the end-cap. With 2900 tones, the HCAL is the heaviest part of the ATLAS detector. It has 420000 scintillator tiles and 9500 photomultiplier tubes [153]. The scintillating-light signal produced at the tiles is read by photomultipliers tubes.

The contribution of the electronic noise is negligible for the tile calorimeter, therefore, its energy resolution only has the stochastic and constant terms [152]:

$$\frac{\sigma_E}{E} = \frac{5.9\%}{\sqrt{E}} \oplus 5.7\%.$$

3.3.3.3 Forward calorimeter

In addition to the ECAL and HCAL, a smaller calorimeter is placed in the end-caps surrounding the beam pipe to cover the forward region ($3.1 < |\eta| < 4.9$), the forward calorimeter (FCAL). This coverage is required for many physics tasks such as the reconstruction of the missing transverse energy of the forward-jet tagging.

This calorimeter is a sampling calorimeter based on LAr as active medium and copper as absorber. The thickness of the FCAL is optimised to achieve high absorption, approximately, 10 radiation lengths [154].

This detector has a resolution of:

$$\frac{\sigma_E}{E} = \frac{100\%}{\sqrt{E}} \oplus 10\%.$$

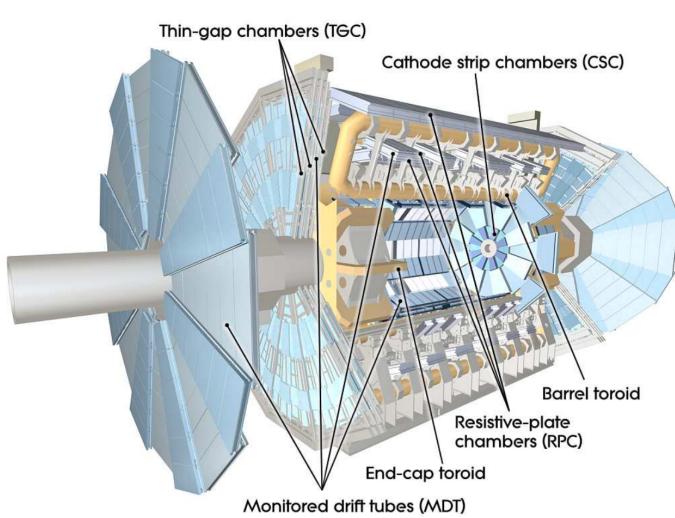


Figure 3.7: Conceptual layout of the MS (blue). The magnet system (yellow) is also shown [132].

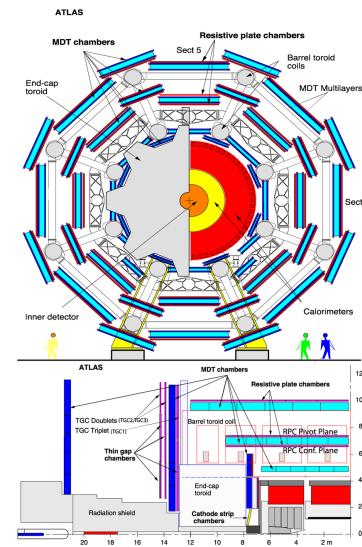


Figure 3.8: ATLAS Muon detectors [155].

3.3.4 Muon Spectrometer

The muons can penetrate through calorimeters and reach the last layer of the ATLAS detector, the MS [155]. Figure 3.7 shows a schematic cut-away view of the ATLAS MS.

The MS surrounds the calorimeters and it aims to measure the trajectories of muons with a pseudorapidity coverage of $|\eta| < 2.7$.

The MS instrumentation is based, on the one hand, on precision chambers for the coordinate measurements in the bending plane: Monitored Drift Tube chambers (MDT) and Cathode-Strip Chambers (CSC), and, on the other hand, on trigger chambers: Resistive Plate Chambers (RPC) and Thin Gap Chambers (TGC). Table 3.3 gives a summary of the MS detector components.

Type	Purpose	Location	Coverage
MDT	Tracking	Barrel + end-cap	$0.0 < \eta < 2.0$
CSC	Tracking	End-cap layer 1	$2.0 < \eta < 2.7$
RPC	Trigger	Barrel	$0.0 < \eta < 1.0$
TGC	Trigger	End-cap	$1.0 < \eta < 2.4$

Table 3.3: Summary ATLAS MS subdetectors [156].

- **Monitored Drift Tube chambers** [157]: The MDT chambers provide precise momentum measurements covering $|\eta| < 2.7$. They determine the curve of the tracks. The MDTs are designed to have stand-alone measurement capability to safeguard against any unanticipated background and to ensure good discovery potential in the scenario of unexpected topologies. To do so, the MDT uses a system of high-pressure (3 bar) drift tubes. An MDT chamber consists of six layers of drift tubes, each of them with 3 cm of diameter, filled with gas. A tube can achieve a single wire resolution of 80 μm . In the entire MDT system, there are 1 171 chambers with a total of 354 240 tubes.
- **Cathode-Strip Chambers**: It is the innermost tracking layer of the MS. The CSCs are multi-wire proportional chambers filled of a mixture of gases with cathode strip read-out. Due to its higher rate capability and time resolution, it is located close to the beam axis, where the particle fluxes are higher. They have a higher granularity than the MDT. This component of the MS covers the range $2.0 < |\eta| < 2.7$. It measures with precision the coordinates at the ends of the detector. It has 70 000 electric channels and provides a resolution of around 60 μm .
- **Resistive Plate Chambers** [158]: This is the barrel element of the trigger system. These chambers are located on both sides of the central CSC and inside the outermost CSC station, covering the $|\eta| < 1.0$ range. The RPCs are gaseous detectors used for triggering and for measuring the second coordinate in the barrel region. RCPs provide a time-space resolution of 1 cm \times 1 ns. The gas gap is of the order of 2 mm and the plate external surfaces are coated by thin layers of graphite painting that allows uniform distribution of the high voltage along the plates. This part of the MS is composed of 3 800 electric channels.
- **Thin Gap Chambers** [159]: The TGCs are multi-wire proportional chambers filled with a mixture of gases with a smaller distance between cathodes and the wire plane compared to the distance between wires. As a first-level trigger, they have to provide high efficiency and excellent time resolution for bunch-crossing tagging in a high-background environment. The TGC presents a $2.0 < |\eta| < 2.7$ coverage. The particle flux received by the TCG is higher than that of the RPC. The three TGCs are located near the middle end-cap MDT station, in the forward regions. The intrinsic spatial resolution for a single layer is 45 μm for a perpendicular incident angle, and the transition region between pads is measured to be about 4 mm. TGCs measure the second coordinate in the non-bending direction with their circa 440 000 electrical channels.

3.3.5 Magnet system

The curvature of particle tracks is crucial to determine their transverse momentum and charge. This bending is achieved through a homogeneous magnetic field produced by toroidal and solenoid magnets, with bending power proportional to $\int B dl$, where B is the magnetic field component orthogonal to the charged-particle direction.

The ATLAS magnet system consists of three subsystems: the central solenoid, barrel toroids (BT), and end-cap toroid (ECT).

- **Central solenoid magnet:** Surrounding the ID, the ATLAS solenoid provides a 2 T magnetic field at the centre of the tracking volume. With a mere 4.5 cm thickness, the particle interaction with the magnet material is minimised, ensuring optimal calorimeter performance. This thin design leverages 9 km of NbTi superconductor wires in pure aluminium strips, cooled to 4.5 K. The magnet has a cylindrical shape, 5.6 m in diameter and 2.56 m in length, weighing 5 tonnes.
- **Barrel toroid magnets:** The BT, being the largest component of the magnet system, generates a field almost perpendicular to particle tracks. Designed as a light and open structure to minimise interference with particles, its coils are placed in individual cryostats linked for stability. This magnet offers a magnetic flux density of 3.9 T on its superconductor. As the largest of its kind, it measures 25.3 m in length, weighs 830 tonnes, and utilises over 56 km of superconducting wire [153, 160]. These air-core toroids of the toroid magnets use an Al/NbTi/Cu superconductor operating at 4.5 K [160].
- **End-cap toroid magnets:** Extending the magnetic field of the BT, the ECTs ensure uniform coverage. Their 4.1 T magnetic field on the superconductor has bending power between 4 and 8 Tm in the range $1.6 < |\eta| < 2.7$ [160]. Each ECT, with a diameter of 10.7 m and weighing 240 tonnes, overlaps with barrel toroids in the range $1.4 < |\eta| < 1.6$, where bending power is reduced.

3.3.6 Trigger and Data Acquisition System

The proton bunches cross at the centre of the ATLAS detector 40 M times per second, resulting in approximately (assuming Run 2 mean pile-up⁹ $\langle\mu\rangle = 33.7$)

⁹The pile-up refers to the simultaneous occurrence of multiple pp collisions in a single bunch crossing, leading to an overlap of signals. It is described with more detail in Section 4.1.3.

1 200 million pp collisions per second. Reading out and storing all the information from these interactions is not feasible since it has a combined data volume of more than 60 million MB per second. Only some of these events are of interest to physics studies and, consequently, only this subset needs to be saved into permanent storage for later analysis. To select only interesting data, ATLAS uses a complex and highly distributed Trigger and Data Acquisition System (TDAQ) [161] that reduces the rate of recorded data from the initial 40 MHz to just an average of 1 kHz. The reduction through the trigger is carried in two steps: The custom-built electronic performs an initial selection and, afterwards, a software-based system analyses the data that passes the initial filter.

The first-level trigger (L1) is a hardware-based filter. The L1 uses the information of the partial granularity of the Calorimeters and the MS to select events up to the maximum-readout rate of the detector (100 kHz) within a latency of 2.5 μ s. Additionally, the L1 identify the regions of interest (RoI), which includes the position and the p_T of the candidate objects.

The data from filtered from the L1 is sent to the software-based trigger, the so called “High Level Trigger” (HLT) [161]. This software-based system is executed on a farm of computers, making use of fast-trigger algorithms.

An average 1.2 kHz output rare for Run 2 passes the HLT (with a latency of just 235 ms) and is sent to the Tier-0 facilities for permanent storage, for worldwide transferring and later offline physics analysis [162]. The decisions performed by trigger about whether or not to store an event are irrevocable.

3.4 Performance of the ATLAS detector

As vast as it is intricate, the performance of the detector hinges crucially on the minutiae of its components and their complex interplay. In particular, a fundamental part for the correct operation of the ATLAS detector is the alignment of its subdetectors [148]. The goal of the detector alignment is to determine the position of the detector geometry as accurately as possible in order to correct the effects of any displacement.

As commented in Section 3.3.2, the ID is used to reconstruct the trajectories of the charged particles by combining into tracks the energy deposits (hits) of the particles as well as identifying primary and secondary vertices. These functionalities are essential for some tasks such as the lepton reconstruction or the b -jet tagging (later described in Section 5.3.1). To be able to have precise and efficient tracking, the full resolution of the ID has to be exploited. The detector experiences

small movements that affect its geometry. These are due to several factors such as movements during the technical stops, thermal expansion/contractions, or any other changes in the operational conditions. With the alignment, it is possible to account for these displacements, re-calibrate and correct their effects. The accuracy of the alignment algorithm is such that the position of the various detector parts may be determined with a few microns of accuracy [106]. Since any misalignment of the different elements of the ID will degrade the quality of the track and object reconstruction, which is vital to performing any physics analysis, constant monitoring of the alignment is necessary. During the development of this thesis, I have contributed to the alignment of the ID through the development of the monitoring software of the track-based ID alignment results obtained at the calibration loop¹⁰.

3.4.1 Local coordinate frame and residuals

Local coordinate frame

In Section 3.3.1 the global (x , y , z) Cartesian coordinate system of ATLAS is introduced. The local coordinate frame of an individual sensor of the detector (x' , y' , z') is also given in the Cartesian system. The local system is a right-handed coordinate system with the origin placed at the geometrical centre of the module. According to the convention, the x' -axis and y' -axis are within the plane of the component and the z' -axis is pointing outside of this plane. The x' -axis points to the most sensitive direction of the module. For the Pixel and IBL modules this is the shorter pitch side and, for the SCT, the perpendicular to the strip orientation. In the case of the TRT the y' -axis points along the wire while the x' -axis remains perpendicular to both the wire and the radial direction. The local coordinates are represented schematically in Figure 3.9.

The hits in the different subdetectors are reconstructed in the local coordinate frame of the different modules.

Residuals

In tracking, a residual is the distance between a hit and the intersection point of the extrapolated track in the sensor plane. The residual vector (\mathbf{r}) is defined as:

$$\mathbf{r} = (\mathbf{m} - \mathbf{e}(\tau, \alpha)),$$

¹⁰The calibration loop, also called 24h loop, is the process of dynamically updating the databases with the calibration results. The calibration loop is executed every 24 hours. It is responsible for correcting the most important movements after data collection and before data reprocessing. This is done using a subset of data (*express stream*) processed at Tier-0.

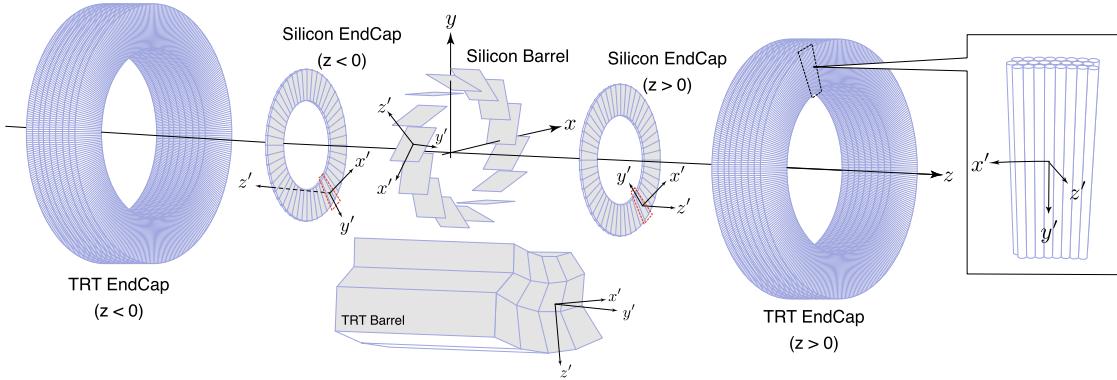


Figure 3.9: Schematic representation of the ATLAS global (x , y , z) and local (x' , y' , z') reference frames [148]. The local coordinates are shown for the Pixel, IBL, SCT and TRT.

where \mathbf{m} is the vector to centre of the module and $\mathbf{e}(\boldsymbol{\tau}, \boldsymbol{\alpha})$ is the vector to the track intersection with the surface. For every module crossed by the track there is a residual, as it is shown in Figure 3.10. The alignment algorithm presented in this section is based on the minimisation of the residuals.

3.4.2 Track parameters and degrees of freedom

The trajectory followed by a charged particle within a magnetic field B is a helix that can be fully parametrised by five track parameters: $\boldsymbol{\tau} = (d_0, z_0, \phi_0, \theta_0, q/p)$, where d_0 and z_0 are the transverse and longitudinal impact parameters; ϕ_0 and θ_0 the azimuthal and polar angles of the track. Lastly, the q/p is the ratio between the particle's charge and momentum and it measures the curvature of the tracks.

The position and orientation of a rigid body can be described by a total of six degrees of freedom. This is translated into what is known as alignment parameters $\boldsymbol{\alpha} = (T_x, T_y, T_z, R_x, R_y, R_z)$. These correspond to the three translations with respect to the origin of the local reference frame ($T_{x,y,z}$) and three rotations ($R_{x,y,z}$) around the local Cartesian axes.

3.4.3 Track based alignment

In the case of a perfectly aligned detector, the distribution of residual vectors would be centred at zero and have a width that corresponds to the module resolution. Therefore, any deviation in the residual distribution indicates a misalignment of the detector.

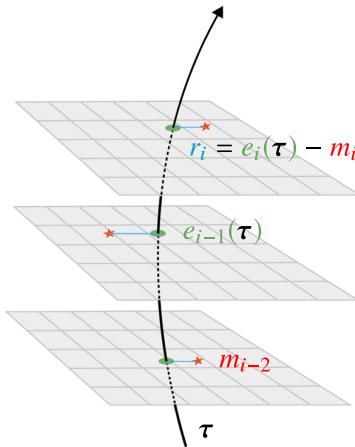


Figure 3.10: Schematic representation of a charged particle crossing detector planes [148]. The red stars represent the measurements in each plane (m_i). The black line is the fitted trajectory for a given set of track parameters. The position of the intersection of the fitted track with the plane (e_i) on which the i^{th} measurement is made is indicated with a green ellipse. The residuals (r_i) are represented by blue lines.

A schematic description of the alignment chain is illustrated in Figure 3.11. The blue rectangles on the left panel of this figure represent the true position of the detector modules. A charged particle produces hits in each module, which are marked with red dots. The track of the particle is marked with a red line. The x distance is the deviation of the module from its apparent position (considering there is no misalignment). In the middle panel, the white rectangles represent the apparent position of each module. It can be seen that the real and the apparent positions are not the same, differing by an unknown distance x . This deviation leads to a discrepancy between the reconstructed tracks and the true ones. The residuals in the middle panel are represented by a green line, which corresponds to the difference between the recorded track (red dots) and the reconstructed one (blue dots). The residual distributions in this panel are displaced from zero, indicating a misalignment. The purpose of the alignment algorithm is to centre these distributions in zero by minimising these residuals. As a result of the alignment procedure, the position of the detectors has been updated a distance x' for each module. After this, the new expected position of the modules (green rectangles) is much closer to the real one and, hence, the residuals are more centred at zero. Anyhow, this is not perfect and the different x' are not all equal to x . To improve the precision, the alignment procedure is repeated iteratively with the calibration loop until the convergence of the corrections.

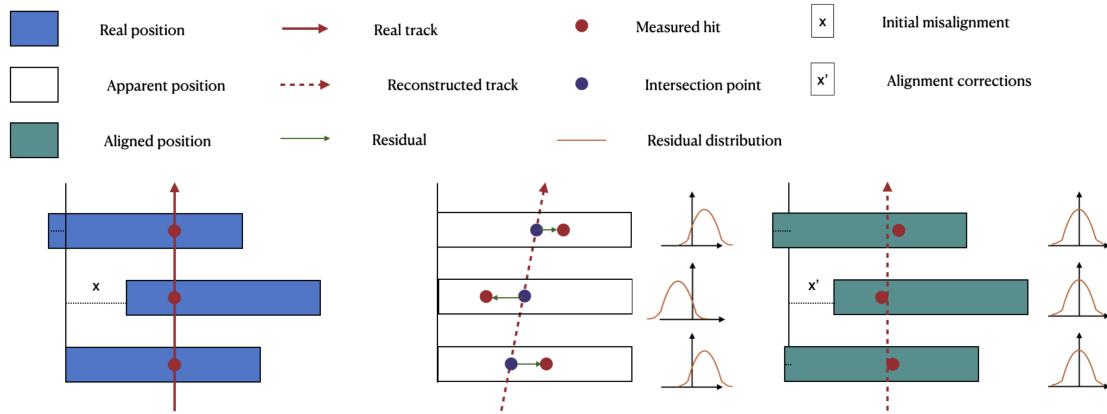


Figure 3.11: Alignment procedure scheme where each rectangle represents a detector module.

3.4.3.1 Global χ^2 algorithm

To correct the position of the ID, the alignment constants ($\boldsymbol{\alpha}$) are obtained as result of the minimisation of a χ^2 function. This function is built from the track-hit residuals:

$$\chi^2 = \sum_e \sum_t \sum_h \left(\frac{r_{t,h}(\boldsymbol{\tau}, \boldsymbol{\alpha})}{\sigma_h} \right)^2,$$

where the index e runs over the selected events, the index t runs over the reconstructed tracks with one event, and the index h is the set of hits associated to each track t in each event. The residual of each hit associated to track t is $r_{t,h}$ and σ_h is the uncertainty of the measured hit. In vector notation, the χ^2 function can be expressed as:

$$\chi^2 = \sum_e \sum_t \mathbf{r}^T \Omega^{-1} \mathbf{r},$$

where Ω is the covariance matrix of the corresponding measurements and \mathbf{r} the residual vector. The alignment constants $\boldsymbol{\alpha}$ are those that minimise the χ^2 and, therefore, first and second derivatives of χ^2 with respect $\boldsymbol{\alpha}$ are used.

$$\frac{d\chi^2}{d\boldsymbol{\alpha}} = \sum_e \sum_t \left[\left(\frac{d\mathbf{r}}{d\boldsymbol{\alpha}} \right)^T \Omega^{-1} \mathbf{r} \right]^T + \sum_e \sum_t \left[(\mathbf{r}^T \Omega^{-1} \left(\frac{d\mathbf{r}}{d\boldsymbol{\alpha}} \right)) \right] = 0$$

It is worth to remark that \mathbf{r} and Ω are defined for all events and tracks with each event, so the sum will accumulate the residuals from all considered tracks from all the events in the data sample. The last expression can be simplified taking into account that Ω^{-1} is symmetric (i.e. $(\Omega^{-1})^T = \Omega^{-1}$) and it takes the form:

$$\frac{d\chi^2}{d\boldsymbol{\alpha}} = 2 \sum_t \left(\frac{d\mathbf{r}}{d\boldsymbol{\alpha}} \right)^T \Omega^{-1} \mathbf{r} = 0. \quad (3.1)$$

Now, since $\mathbf{r} = \mathbf{r}(\boldsymbol{\tau}, \boldsymbol{\alpha})$, partial derivatives have to be taken into account:

$$\frac{d\mathbf{r}}{d\boldsymbol{\alpha}} = \frac{\partial \mathbf{r}}{\partial \boldsymbol{\tau}} \frac{d\boldsymbol{\tau}}{d\boldsymbol{\alpha}} + \frac{\partial \mathbf{r}}{\partial \boldsymbol{\alpha}}$$

Inserting this into equation 3.1, the condition for minimising the χ^2 turns to be:

$$\sum_t \left(\frac{\partial \mathbf{r}}{\partial \boldsymbol{\tau}} \frac{d\boldsymbol{\tau}}{d\boldsymbol{\alpha}} + \frac{\partial \mathbf{r}}{\partial \boldsymbol{\alpha}} \right)^T \Omega^{-1} \mathbf{r} = 0. \quad (3.2)$$

The alignment parameters that satisfy the relation in equation 3.2 are found by an iterative process consisting on evaluating the first and second derivatives of the χ^2 with respect to the current iteration alignment parameters, $\boldsymbol{\alpha}_0$. Since the derivative terms of equation 3.2 depend on $\boldsymbol{\alpha}$ itself, the procedure is repeated until a convergence criteria is achieved. To get the set of alignment parameters corrections ($\delta\boldsymbol{\alpha}$), the alignment parameters can be written as $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0 + \delta\boldsymbol{\alpha}$. Therefore, from the equation 3.1, the following relation is obtained

$$\left[\sum_e \sum_t \left(\frac{d\mathbf{r}}{d\boldsymbol{\alpha}_0} \right)^T \Omega^{-1} \left(\frac{d\mathbf{r}}{d\boldsymbol{\alpha}_0} \right) \right] \delta\boldsymbol{\alpha} + \sum_e \sum_t \left(\frac{d\mathbf{r}}{d\boldsymbol{\alpha}_0} \right)^T \Omega^{-1} \mathbf{r}(\boldsymbol{\tau}_0, \boldsymbol{\alpha}_0) = 0. \quad (3.3)$$

Hence, from this equation it is possible to define the alignment matrix (M_a) and vector (v_a) as:

$$M_a = \sum_e \sum_t \left(\frac{d\mathbf{r}}{d\boldsymbol{\alpha}_0} \right)^T \Omega^{-1} \frac{d\mathbf{r}}{d\boldsymbol{\alpha}_0},$$

$$v_a = \sum_e \sum_t \left(\frac{d\mathbf{r}}{d\boldsymbol{\alpha}_0} \right)^T \Omega^{-1} \mathbf{r}(\boldsymbol{\tau}_0, \boldsymbol{\alpha}_0).$$

And the relation in equation 3.3 can be rewritten as $M_a \delta\boldsymbol{\alpha} + v_a = 0$, implying that $\delta\boldsymbol{\alpha} = -M_a^{-1} v_a$. The alignment parameters $\boldsymbol{\alpha}$ are iteratively derived until convergence is reached. The iterative expression is $\boldsymbol{\alpha}_N = \boldsymbol{\alpha}_{N-1} + \delta\boldsymbol{\alpha}_N$, where N is the current iteration.

The Global χ^2 algorithm encompasses all alignable modules, factoring in their intercorrelations, which makes solving equation 3.3 computationally challenging. Given the intricate granularity and sophistication of the ID, alignment can be undertaken at various hierarchical levels. These levels correspond to the ID's assembly structure, with the complexity ascending progressively. This gradation is realised by mapping the residuals, initially computed at the module level, onto more expansive surfaces corresponding to broader ID structures. The subcomponents of the ID are categorised into five tiers (see Table 3.4), ranging from a count of 7 structures to 351×10^3 , based on their structural composition.

Level	Description	Number of structures
1	IBL, Pixel, SCT end-caps, TRT barrel and end-caps	7
Si2	IBL Layers, Pixel end-cap disks and barrel layers, SCT end-cap disks and barrel layers	32
Si3	IBL modules, Pixel modules and SCT modules	6112
TRT2	TRT barrel modules and end-cap wheels	176
TRT3	TRT straw tubes	351×10^3

Table 3.4: Alignment configuration split by levels used during Run 2 data-taking period.

3.4.4 ID alignment monitoring

The ID Alignment Monitoring Web Display (Figure 3.12) is an application intended for monitoring the track-based alignment results obtained at the calibration loop for the ID. It helps to evaluate the computed alignment corrections as well as many graphical distributions directly related to the performance.

The web application is structured with a server overseen by the ATLAS Distributed Computing. It incorporates a suite of backend scripts designed for generating distributions, refreshing the data, and managing HTTP requests from the users. The frontend allows to the user to interact with the application by requesting alignment-related information. This tool is integrated within Athena¹¹, the established software framework for ATLAS [142].

A segment of my responsibilities encompassed the development and implementation of both the client-side (frontend) and server-side (backend) code for the ID Alignment Monitoring Web Display.

In the enhanced version of the Alignment Monitoring Web Display, several advancements and refinements have been made compared to its predecessor:

- Integration with the standard Athena setup has been achieved.
- Conformity to ATLAS styling and debugging standards has been ensured [164]. These standards are a compendium of rules, recommendations and advice for presenting the information within the ATLAS collaboration. The web presents a professional visual identity that is both memorable and easy to recognise.

¹¹Athena represents a specific implementation of a component-based architecture, drawing inspiration from LHCb’s Gaudi [163]. It is tailored for an extensive variety of physics data-processing tasks.

- Enhanced efficiency in execution has been introduced by allowing updates for individual runs or batches of runs.
- The system now dynamically updates the year, eliminating the necessity for manual hardcoding.
- Through frontend interactions, users can now select any run for visualisation, thereby eliminating hardcoded dependencies.
- Users can now query the ATLAS Metadata Interface¹² for relevant data.
- Algorithmic optimisation has been performed, streamlining code execution through a singular iterative loop.
- The codebase has been refined to eliminate redundancies and duplicates with respect to the previous versions.

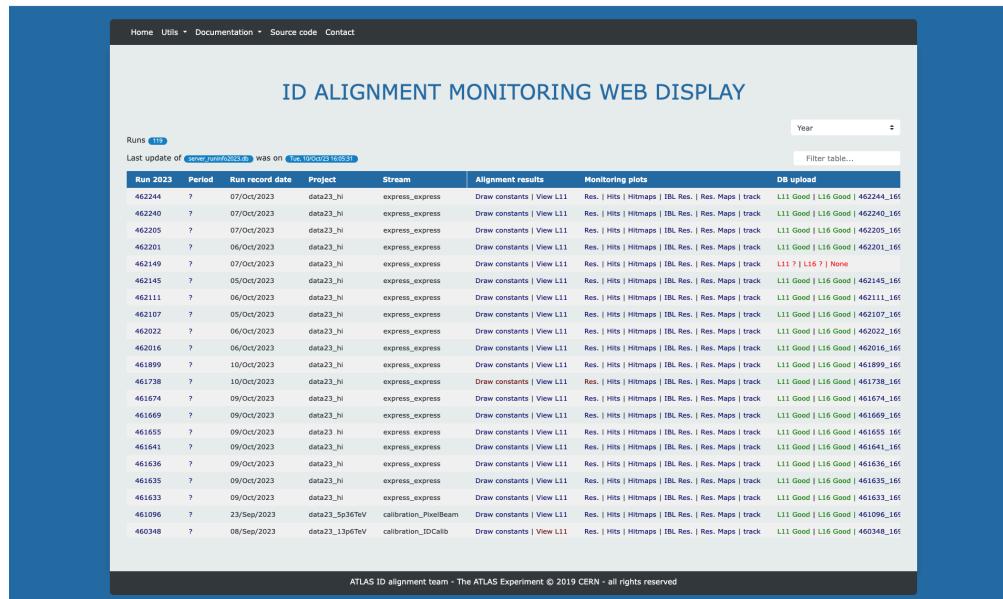


Figure 3.12: Main page of the ID Alignment Monitoring Web Display on wide resolution screen. The monitor presents the runs and allows to access the alignment information and plots. Queries can be used to filter the runs presented.

The graphical interface and layout of the application are designed to adhere to the ATLAS style guidelines. The frontend code foundation leans on the CherryPy

¹²Known as AMI, it is a generic software ecosystem for retrieving scientific data by metadata criteria. It allows to search for real and simulated data by metadata criteria as well as browse, view, compare and create ATLAS AMI-Tags

library [165], an object-oriented web framework developed in Python. This choice is motivated by its simplicity and good capabilities. However, it is pertinent to note that contemporary web development trends see an inclination towards more feature-intensive Python web frameworks, such as Django and Flask. As a result, CherryPy, while still effective, might not be the premier choice for newer undertakings.

Supplementing the core functionalities, the application seamlessly integrates CSS, HTML [166], and Java. The revamped web interface provides users with an intuitive mechanism to determine the displayed information. A pivotal enhancement is its responsive design¹³, enabling impeccable adaptability across devices, from desktops to mobiles. This feature ensures that shifters¹⁴ can effortlessly access alignment data from any device, optimising their workflow. This mobile-centric adaptability is powered by Bootstrap [167], a leading CSS framework renowned for championing responsive, mobile-first frontend web development.

Other enhancements in the frontend include:

- The navigation bar has been enriched, offering a wider array of options for seamless navigation.
- A sophisticated filter mechanism has been incorporated, enabling users to exclusively display runs that meet specific criteria.
- An intuitive selector is now integrated, showcasing the years available in the database for hassle-free selection.
- Accessibility has taken a front seat, with the introduction of features like the hover selector, a grid-based display for plots, and a modal image gallery (lightbox) that simplifies plot navigation. All these features have been meticulously crafted using CSS¹⁵.

Cumulatively, these advancements not only bolster the execution speed of the application but also provide access to more information, and allow to easily manage and access the desired data. The ID Alignment Monitoring Web Display that I engineered is currently being used actively by members of the ATLAS collaboration

¹³Responsive design in web programming refers to the approach of creating web pages that automatically adjust and optimise their layout and content based on the screen size and orientation of the device on which they are viewed.

¹⁴The term “shifters” refers to the individuals who are assigned to monitor the ATLAS detector and its data-acquisition systems during data-taking periods. In this context, I am referring to the ID alignment shifters.

¹⁵CSS was first proposed on 1994 at CERN.

to monitor the ID alignment during Run 3 as well as the reprocessing of the data of Run 2. It has significantly simplified the monitoring of alignment, making the task more efficient and user-friendly.

Chapter 4

Recording data and simulating events with the ATLAS detector

In God we trust, all others bring data.

—WILLIAM EDWARDS DEMING

In particle physics, pp collisions are studied to improve our understanding of physics of the elementary particles and their interactions. The LHC provides these collisions and the ATLAS detector records them. In order to allow a statistical interpretation of the collision data, the theoretical prediction is derived making use of Monte Carlo (MC) generators.¹. The current baseline theory to describe the interactions between particles is the SM.

For real-collision data, the detector response is evaluated via various algorithms in order to reconstruct the objects in the final state. For the simulated processes, the MC-generated particles undergo a series of steps replicating the physics of the collisions, the interaction with the detector material, the magnetic fields, and the response of the detector and its electronics. Once these steps are applied, the

¹The scientific method is to compare the prediction and the collected data. Then, if there is a disagreement, one needs to update the model that produced the prediction. This procedure is done iteratively.

simulated events are processed using the same algorithms as the ones used with real-collision data.

This chapter is divided into three sections: In Section 4.1, the phenomenology of the pp collisions is described. Here the parton model and underlying event are presented. The concepts of luminosity and pile-up are discussed as well. Section 4.2 describes the collection of the data with the ATLAS detector, while Section 4.3 discusses the generation of simulated data by presenting the steps and tools of the simulation chain.

4.1 Phenomenology of proton–proton collisions

The analysis presented in this thesis uses data obtained from pp collisions provided by the LHC. Therefore, understanding the physics behind these collisions is fundamental to understanding the physics analyses conducted at this collider. The pp collisions can be categorised into two broad classes based on how the energy and momentum are distributed among the colliding particles. Collisions can be elastic and inelastic. In the former, the kinetic energy is conserved and the collision is a scattering off two protons without changing their intrinsic properties or producing any new particles. In the latter, the protons either change their intrinsic properties or new particles are produced as the result of the collision.

At the LHC energy regime, the pp collisions are predominantly inelastic and they cannot be described as point-like interactions. Here is where the parton model and the PDFs come into play. The PDFs are functions containing the long-distance structure of the hadrons in terms of valence and sea quarks and gluons (explained in Section 4.1.1). This description is known as the “parton model” and is discussed throughout this section.

During the Run 2 data-taking period of the LHC, pp collisions took place with a centre-of-mass energy of 13 TeV. The total cross-section of pp collisions at this energy is measured to be $\sigma_{\text{tot}} = (110.6 \pm 3.4)$ mb [168]. The method used to measure σ_{tot} also made it possible to separate the cross-section into the elastic $\sigma_{\text{el}} = (31.9 \pm 1.7)$ mb and the inelastic one $\sigma_{\text{inel}} = (79.5 \pm 1.8)$ mb. However, only inelastic scattering generates particles with a sufficient angle with respect to the beam axis so that these particles enter into the geometrical acceptance of the detector. The cross-section can be computed as the convolution of PDFs with the parton scattering matrix element \mathcal{M} . In the context of scattering experiments, the \mathcal{M} refers to the amplitude for a particular scattering or decay process between initial and final states. The \mathcal{M} provides information about the probability of transitioning from an initial state to a final state through a given interaction.

4.1.1 Proton structure and parton model

Formulated by R. P. Feynman the parton model for hadrons describes these non-fundamental particles as a composite of several point-like constituents named partons [169]. In this model, the proton is not only made of its three “valence” quarks (two u -type and one d -type quarks) but also, there is a “sea” of gluons and short-lived quark–antiquark pairs created through gluon splitting. The partons in the sea are continuously interacting with each other and, depending on the energy, can have any flavour.

The distribution of the momentum of a hadron among its constituents is described by its PDFs [170]. The momentum of the partons within a proton is determined through fits to several data points obtained from experimental data such as the ones coming from deep inelastic scattering, Drell-Yan, and jet measurements. Various global fitting collaborations, including ABM [171], CT [172], MMHT [173], NNPDF [174], MSTW [175], and CTEQ [176], employ different methods to perform these fits, which are subsequently extrapolated to new energy scales.

Formally, the PDF $f_{a/A}(x, Q^2)$ is the probability of finding parton a in a hadron A carrying a fraction $x = p_a/p_A$ of its momentum at the energy scale Q^2 . At lower energies (i.e. $Q \sim 1$ GeV), the momentum of a proton is primarily shared among its valence quarks, while at higher energies (i.e. $1 < Q \lesssim 1$ GeV), the emission of gluons carrying some of the initial momentum of the quarks is more likely. As an example, PDFs for several parton flavours at two different energy scales as a function of x are presented in Figure 4.1. In QCD theory, these interactions can be divided into two categories: Hard-scattering, which can be calculated by perturbation theory due to the small α_s , and low-energy processes, which cannot be calculated because α_s is large and have a much greater impact on non-perturbative QCD.

When two protons (A and B) collide at the LHC, their partons can interact via a hard-scattering process. Each of the interactions between the parton pairs is independent of the interactions with other partons. The remaining partons also contribute to the final state as “underlying events” (see Section 4.1.2). Figure 4.2 provides a simplified representation of a pp collision. In this schema, each of the two partons that interact carries a different fraction $x_{i=a,b}$ of the total proton energy.

The total cross-section (σ) in a hadron–hadron hard-scattering process in which a parton a from hadron A interacts with a parton b from hadron B , such as a pp interaction, to produce a final-state X is:

$$\sigma_{AB \rightarrow X} = \sum_{a,b} \iint dx_a dx_b f_a(x_a, Q^2) f_b(x_b, Q^2) \hat{\sigma}_{ab \rightarrow X},$$

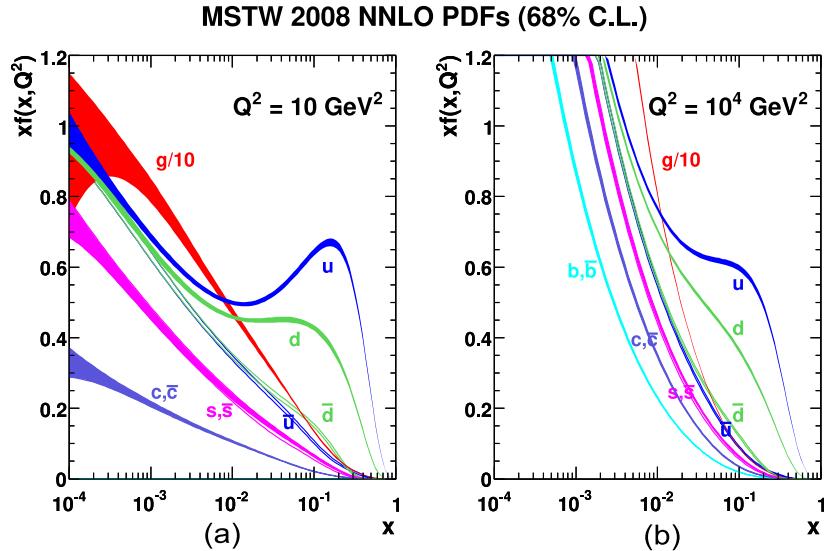


Figure 4.1: Parton distribution functions $xf(x, q^2)$ against the fraction x for gluons and different quark flavours at (a) $Q^2 = 10 \text{ GeV}^2$ and (b) $Q^2 = 10^4 \text{ GeV}^2$ energy scales using the MSTW 2008 NNLO PDF sets [177].

where $f_i(x_i, Q^2)$ is the PDF of $i = A, B$. Here, the Q is chosen to be the factorisation scale² (μ_F). The contribution of the individual partons a and b is denoted by $\hat{\sigma}_{ab \rightarrow X}$. With this equation, all processes in pp collisions can be computed.

Depending on the order achieved in perturbation theory (LO, NLO, NNLO, ...), the cross-section of the individual partons to produce the final state of interest ($ab \rightarrow X$) can be calculated as:

$$\begin{aligned}\hat{\sigma}_{ab \rightarrow X} &= \sum_{i=0}^{\infty} \alpha_s^i(\mu_R) \sigma_n(x_a, x_b, \mu_F^2) \\ &= [\sigma_{\text{LO}}(x_a, x_b, \mu_F^2) + \alpha_s(\mu_R) \sigma_{\text{NLO}}(x_a, x_b, \mu_F^2) \\ &\quad + \alpha_s(\mu_R)^2 \sigma_{\text{NNLO}}(x_a, x_b, \mu_F^2) + \dots]_{ab \rightarrow X},\end{aligned}$$

where $\alpha_s^i(\mu_R)$ is the coupling constant derived for a specific renormalisation scale³ (μ_R). In theory, if the entire perturbation series could be computed, the need for μ_F and μ_R parameters would vanish. However, this is not feasible and the series must be truncated at a specific order. Hence, it becomes crucial to establish the values of μ_F and μ_R . This results in uncertainties in the calculations which are often

²The factorisation scale μ_F establishes the boundary between low and high energy, thereby determining the scale at which perturbative calculations become valid.

³The renormalisation scale μ_R is used to address the ultraviolet divergences in QCD that occur due to high momentum in loops.

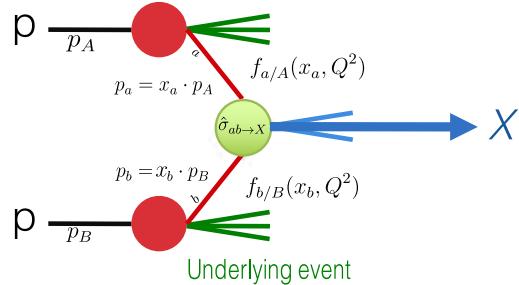


Figure 4.2: Simplified view of a pp collision. The PDFs represent the structure of the proton (red) and they are obtained using experimental data and QCD properties. The light green bubble represents the hard-scatter parton cross-section. The blue X is the product of the interaction and it can be jets, bosons, quarks or any other particles.

addressed by varying the chosen values of these parameters. A deeper discussion on the different types of uncertainties is given in Section 6.7.

4.1.2 Underlying event

Softer interactions of other partons often accompany the hard collision of two partons in the protons. These processes are referred to as multi-parton interactions. The underlying event (UE) encapsulates all that is observed from a pp collision that does not come directly from the primary hard-scattering process. This encompasses elements such as beam-beam remnants (BBR), multiple parton interactions (MPI) within a single collision, and initial and final state radiation (ISR/FSR) [178]. Since the BBR and the MPI processes are low energetic, they cannot be predicted by perturbation theory but, instead, they have to be simulated by phenomenological models containing many parameters. Typically, the UEs have lower p_T than the main process.

Precise modelling of the UE is crucial for conducting successful experimental studies because this soft interaction may affect the high- p_T measurements. This is because the UE allows for a clear differentiation between the direct products from hard scattering and the rest of the event. Therefore, without the UE one would get the wrong differential cross-sections. Through colour flow and recombination, the UE influences the hard-scattering. Figure 4.3 illustrates the UE in a pp collision at the LHC.

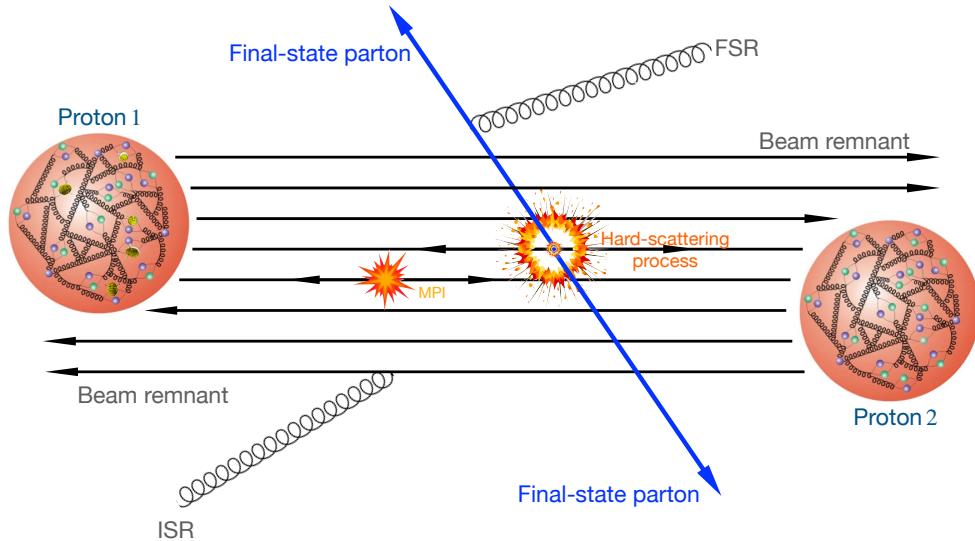


Figure 4.3: Schematic representation of a pp interaction. The collision consists of the hard-scattering interaction between two partons of the protons, resulting in two final-state partons. The UE includes additional components such as MPIs, ISR and FSR, and particles originating from the break-up of the protons (i.e. beam-beam remnants). It encompasses everything except the two outgoing hard-scattered particles.

4.1.3 The pile-up effect

Pile-up is a challenging matter in terms of designing a detector and its DAQ system as well as in the physics and performance analyses. Because the LHC collides bunches of protons instead of single protons, multiple pp interactions occur at a single bunch-crossing⁴. Even with single protons this could happen since several partons of the two protons can intervene in the event. This overlay can result in multiple collisions at the same time and several interactions with the same detector element, thereby generating overlapping signals which may be difficult to differentiate. This is what is called pile-up.

As mentioned, the bunches are composed of $\sim 10^{11}$ protons, but since the protons are so small compared to the bunch size, there are only around 30 pp collisions per bunch crossing with nominal beam at the LHC Run 2. This amount is also related to the crossing angle between the bunches (θ_c). The larger θ_c , the smaller is the area of overlap between the bunches and, hence, the smaller is the possibility of a collision. The mean number of interactions ($\langle \mu \rangle$) per bunch crossing is presented in Figure 4.4 for the three runs of the LHC. Note that for

⁴A bunch crossing is defined as the instance in which two collections of protons collide at the central region of the detector.

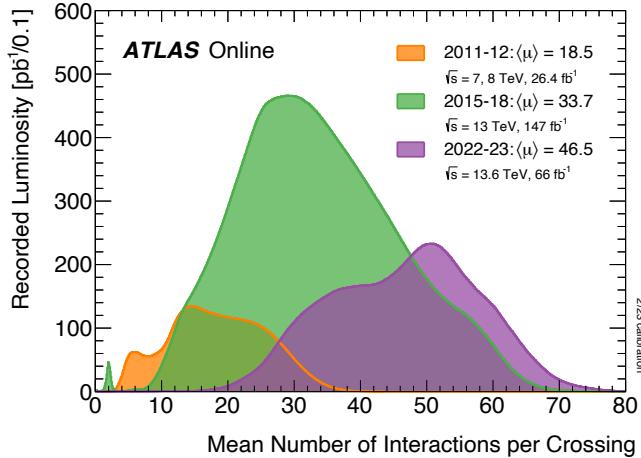


Figure 4.4: Luminosity-weighted distribution of the mean number of interactions per crossing, $\langle\mu\rangle$, in pp collisions at $\sqrt{s} = 7$ and 8 TeV during Run 1 (orange), $\sqrt{s} = 13$ TeV during Run 2 (green), and $\sqrt{s} = 13.6$ TeV at Run 3. All physics-data recorded by the ATLAS detector during stable beams in these years are shown (until July 2023) [180].

each run, since the instantaneous luminosity is increased, the pile-up is higher than for the previous run. The $\langle\mu\rangle$ corresponds to the mean of the Poisson distribution of the number of interactions per crossing calculated for each bunch. It is calculated from the instantaneous luminosity produced by a single pair of colliding bunches ($\mathcal{L}_{\text{bunch}}$), the inelastic cross-section and revolution frequency as $\langle\mu\rangle = \mathcal{L}_{\text{bunch}} \times \sigma_{\text{inel}}/f$ [179].

4.1.4 Luminosity

Luminosity is a measure of the number of collisions that take place in a time frame. It plays a pivotal role in determining the rate at which particle interactions occur, thus influencing the experimental data obtained. Besides the centre-of-mass energy, the luminosity is the most relevant parameter to characterise a collider-based experiment, and it is especially important in searches for processes with small production cross-section (known as rare processes). It measures the ability of the particle accelerator to produce enough events of a desired type.

The instantaneous luminosity, $\mathcal{L}(t)$, is the ratio of events produced in a certain period of time for a given cross-section σ :

$$\mathcal{L} = \frac{1}{\sigma} \frac{dN}{dt} = \frac{1}{\sigma} R,$$

where N is the number of the events and t the time. The symbol $R = \frac{dN}{dt}$ is known as event rate. It can be understood as the number of particle collisions per unit of area (typically expressed in cm^2) and per second, therefore it is measured in $\text{cm}^2 \text{s}^{-1}$ [181]. Alternatively, the luminosity can also be expressed in barns⁵.

The instantaneous luminosity is proportional to the number of protons in each of the bunches that collide head-on (n_1 and n_2), the revolution frequency (f) with which the bunches are crossing, and the number of proton bunches in the machine (n_b). The \mathcal{L} is inversely proportional to the effective transverse area of the beams in which the collision takes place ($4\pi\sigma_x\sigma_y$):

$$\mathcal{L} = f \cdot \frac{n_1 n_2 n_b}{4\pi\sigma_x\sigma_y} \cdot F(\theta_c, \sigma_x, \sigma_z), \quad (4.1)$$

where $F(\theta_c, \sigma_x, \sigma_z)$ is a factor accounting for the luminosity reduction due to the beam-crossing angle (θ_c). At the LHC, assuming that the particles travel at the speed of light, for its 27 km long, the bunch-crossing frequency is $f = 11.2455 \text{ kHz}$. The maximum number of proton bunches in the machine is⁶ $n_b = 2808$ (see Table 3.1). In each bunch, there are $n_1 \approx n_2 \approx 1.2 \times 10^{11}$ particles at the LHC Run 2. Finally, characterising the optics of the collision at the IP, the root mean square transverse beam width in the horizontal and vertical directions are $\sigma_x \approx \sigma_y \approx 12, \dots, 50 \mu\text{m}$. Equation 4.1 assumes that the particles in the beam follow a Gaussian distribution. According to equation 4.1 the instantaneous luminosity only depends on the machine and its beam parameters [182]. The instantaneous luminosity for the LHC is designed to achieve $\mathcal{L}_{\text{LHC}_{pp}} = 2.1 \times 10^{34} \text{ cm}^2 \text{s}^{-1}$ in pp collisions and $\mathcal{L}_{\text{LHC}_{\text{PbPb}}} = 6.1 \times 10^{27} \text{ cm}^2 \text{s}^{-1}$ for heavy ion collisions.

The integrated luminosity (L) over time is given by

$$L = \int \mathcal{L} dt,$$

and it is used to determine the number of events, N , that have taken place during that time within the detector, i.e. $N = \sigma \times L$. Therefore, the number of observed events is:

$$N_{\text{events}}^{\text{obs}} = \sigma_{\text{process}} \times \text{efficiency} \times L,$$

where the efficiency of the detection has to be optimised experimentally, the L is delivered by the LHC and the cross-section of the process (σ_{process}) is given by nature. The efficiency accounts for the geometrical acceptance of the detector, trigger efficiencies, and the reconstruction and identification of physical objects.

⁵A barn is a metric unit of area following the equivalence $1\text{b} = 10^{-28} \text{ m}^2$.

⁶The theoretical maximum of 3564 bunches cannot be reached due to space needed between bunch trains and for the beam dump kicker magnets.

Several factors can limit the maximum luminosity that can be achieved at the LHC [181] such as the beam-beam effect, crossing angle, beam offset or the hour-glass effect. On the other hand, there are diverse strategies to maximise the luminosity delivered by a machine (e.g. maximise the total beam current or compensate reduction factor).

4.2 Recording data with the ATLAS detector

The ATLAS experiment produces an enormous amount of data through high-energy pp collisions. To interpret of this vast volume of information, a well-defined data model is crucial.

The ATLAS data model serves as the foundation for organising and managing the recorded data. It provides a standardised and structured framework that allows to access and analyse data effectively. In Figure 4.5 the ATLAS data model is presented by comparing, at each level, the real-detector data with the simulated event data.

4.2.1 Cumulative luminosity

In Section 4.1.4 a description of instantaneous and integrated luminosity is given. Here the cumulative-luminosity results for the ATLAS detector during Run 2 are presented for each year. The importance of these details lies in the fact that this is the dataset used for the research presented in this thesis. The cumulative luminosity plays a crucial role in determining the statistical significance of measurements

The cumulative luminosity delivered by the LHC to the ATLAS detector is shown in Figure 4.6 for every year of LHC operation. In Figure 4.7, the total Run 2 cumulative luminosity is presented differentiating between the delivered and recorded luminosity and showing that almost all delivered events are considered to have good data quality. Note that the luminosity delivered by the LHC machine is not the same as the one registered by the ATLAS detector, although these numbers are very close. The delivered corresponds to the luminosity delivered by the LHC machine to the detector and, ideally, should be fully recorded. But in some cases, the ATLAS detector is unable to take data either because one or more subdetectors are temporally unavailable or because its DAQ chain is busy (see Section 3.3.6). The data quality algorithms check the operation of the detectors and the performance of the physical object reconstructors to decide which data should be accepted. The All Good Data Quality criteria is used and it requires

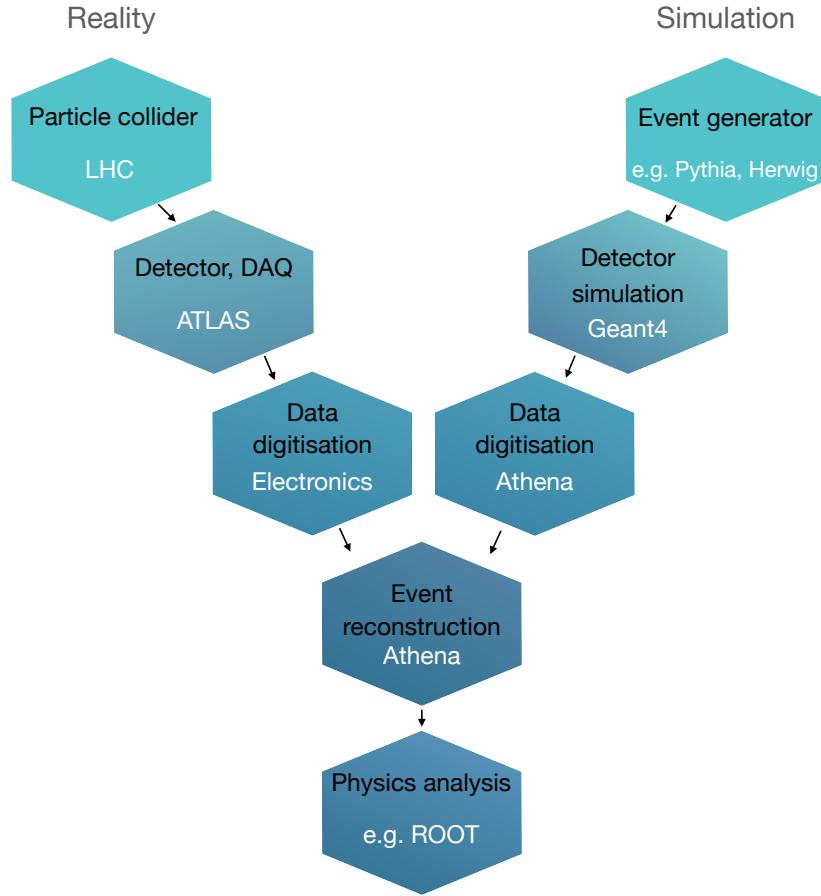


Figure 4.5: Comparison of the paths followed by data recorded by the ATLAS detector and the simulated samples. At each step, the format of the simulated data is the same as the recorded data.

all reconstructed physics objects to be of good data quality [183]. For the entire Run 2 dataset, 95.6% of events are labelled as good for physics and are included in the **Good Run Lists** (GRLs).

4.3 Simulating events within the ATLAS detector

To study the physics taking place in the ATLAS detector, the signals and backgrounds in the analyses are simulated by MC event generators according to

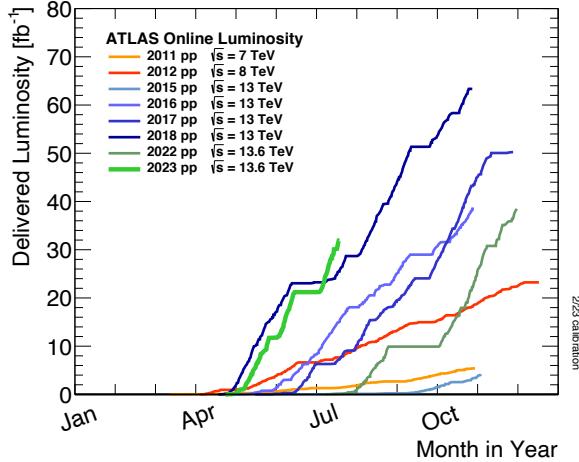


Figure 4.6: Cumulative luminosity versus day delivered to the ATLAS detector by the LHC during stable beams and for high energy pp collisions for Run 1, Run 2 [179, 183] and Run 3 [180].

the cross-sections predicted by the SM. The use of the MC simulations is extensive and there are many different model generators and techniques. As with all MC algorithms, these methods rely on repeated random sampling to obtain numerical results. Since the randomness is intrinsic to the particle-collision processes, a large number of events have to be simulated using MC techniques and such a collection of events is called an MC event sample.

Typically, the chain for simulation data within a particle detector can be divided into these three steps [184]:

1. Generation of physics events and immediate decays of the particles involved: an event generator produces the result of the collisions in terms of particles (at parton level) created through a given physics process and stores any stable particle expected to propagate through the detector. At this point, the geometry of the detector is not considered yet and only the immediate decays are taken into account. These decays include the final-state particles from the hard-scattering process, the showering, the hadronisation, and the pile-up. This is further discussed in Sections 4.3.1.1 - 4.3.1.5.
2. Simulation of the detector and physics interactions: at this point, all particles from the previous step are propagated through a simulation of the materials that form the ATLAS detector using GEANT4 [185]. This is a toolkit for the simulations of the passage of particles through matter. This part simulates all major physical components and materials as well as the interactions of particles such as ionisation in trackers, energy depositions in calorimeters,

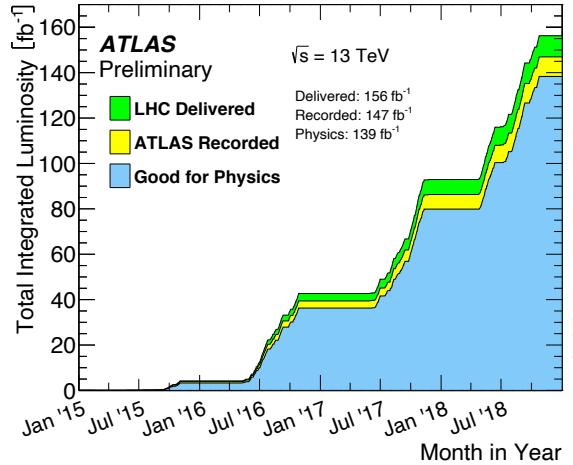


Figure 4.7: Total cumulative luminosity versus time delivered to ATLAS (green), recorded by the ATLAS detector (yellow) and passing the good-quality-data requirements for ATLAS physics analyses (blue) during stable beams for pp collisions at $\sqrt{s} = 13$ TeV during Run 2 [183].

intermediate decays, radiation and scattering. In Section 4.3.1.6, a deeper discussion on this topic is provided.

3. Digitisation of the interactions with the sensitive regions of the detector, such as energy deposits, into voltages and currents, and simulation of all the electronics. The digitisation is discussed in Section 4.3.1.6 alongside the detector interaction.

The output of the full simulation chain has the exact same format as a real event registered by the ATLAS DAQ system (see Section 3.3.6). The steps of the data–simulation model are summarised in Figure 4.5.

The so-called parton-level data contains the information of the particles from the hard-scattering process and their immediate decays. In the analysis presented in this thesis, the parton-level information has several uses, for example, the determination of misidentification rates or the lepton origin assignment. An important part of the work I carried out during this thesis is developing a software package within Athena, `TopPartons`, for obtaining, processing and analysing parton-level information for the tHq processes.

The different steps in the event generation are described through Section 4.3.1. In Section 4.3.2, the generators employed in this analysis are presented.

4.3.1 Steps for simulating event data

The generation of the simulated event samples includes the effect of multiple pp interactions per bunch crossing, as well as the effect on the detector response due to interactions from bunch crossings before or after the one containing the hard interaction.

Every different possible process that can take place during the collision has to be simulated. To ensure a proper description of the entire physics phase-space of a LHC collision, the MC event samples are not generated proportionally to the cross-section of each process because that would cause a poor characterisation of the rare processes. Instead, a sufficiently large amount of events are generated for each process and, afterwards, all events are reweighted to match their correspondent cross-section. This is the origin of the negative weights in the MC samples. The combination of all these processes provides an accurate description of the collision.

4.3.1.1 Hard-scattering process

The first element in the generation of the events is the simulation of the hard-scattering processes, as it is shown in Figures 4.5 and 4.8. Here the matrix elements (\mathcal{M}) with the hard-scattering information of the different processes are generated with a given theoretical accuracy (LO, NLO, etc). From this information, the cross-sections of the different processes are computed. The complexity determining the \mathcal{M} for a particular process scales with the accuracy order. Section 4.1 gives more details about how the pp interactions are modelled. Once the hard-scatter process is simulated, radiative corrections are applied in the form of parton shower (PS) and hadronisation.

In this analysis, the two most relevant computational frameworks for implementing calculations to obtain \mathcal{M} are MADGRAPH5_AMC@NLO [174] POWHEG BOX [186] and SHERPA [187].

4.3.1.2 Parton shower and hadronisation

Once the hard-scattering process is simulated, the second step is to incorporate corrections to account for additional radiations. Both the PS and the hadronisation are simulated by the event generator.

Parton shower

Parton showers are algorithms employed to simulate the soft radiation of gluons and

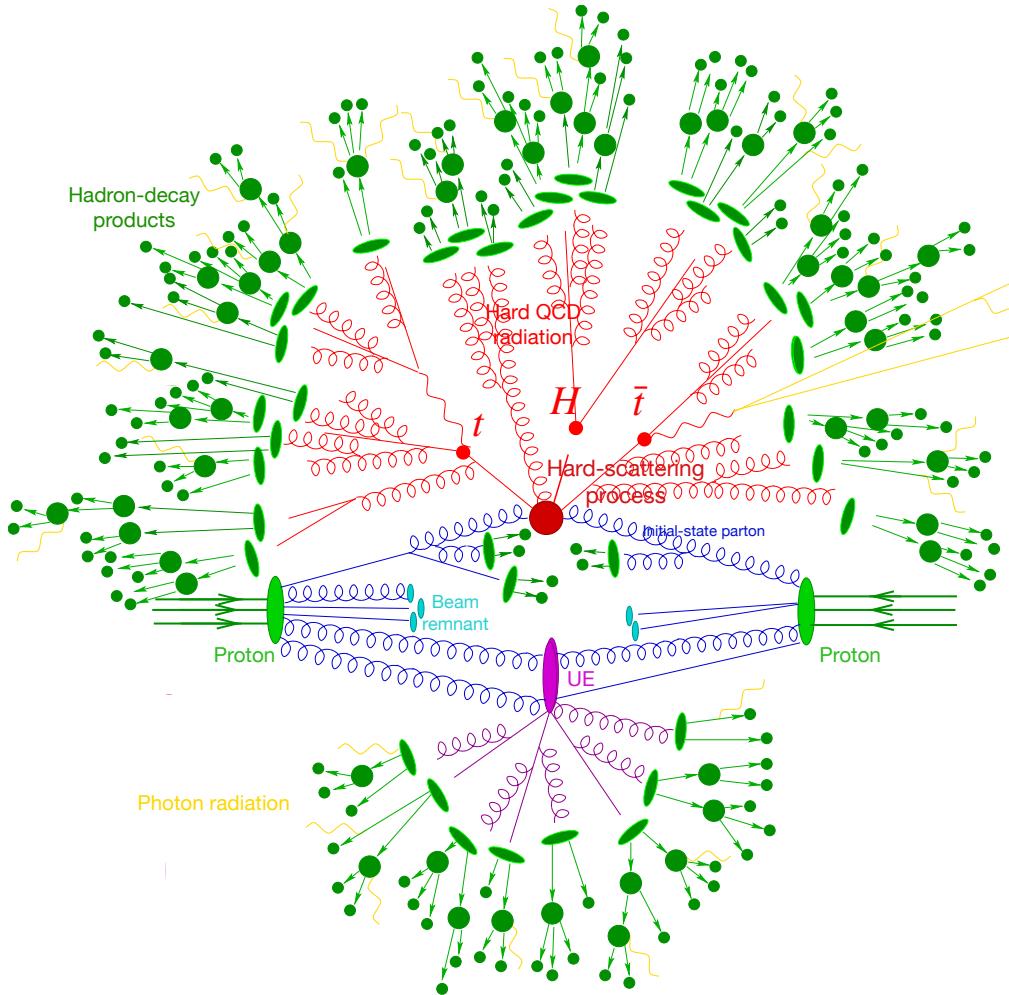


Figure 4.8: Representation of a $t\bar{t}H$ event from a $p\bar{p}$ collision as produced by an event generator [187]. The big red blob at the centre is the hard-scattering interaction, which produces the final state composed of the Higgs boson and the two top quarks, represented by the three small red blobs. The additional QCD radiation produced from these particles and the PS is also in red. The secondary interaction, in purple, occurs before the formation of hadrons (in light green). In darker green, the hadron decay is presented. The photon radiation can occur at any moment and appears in yellow.

quarks, collectively termed partons, during high-energy particle collisions. When quarks or gluons with high energy are produced post-collision, they can emit subsequent softer gluons as they traverse away from the IP.

All the incoming or outgoing partons are involved in the PS simulation. If the particles have colour charge or electric charge can emit QCD (i.e. gluons) or QED (i.e. photons) radiation, respectively. This process generates hundreds of particles with varying energies and momenta spanning multiple orders of magnitude. The PS algorithm mimics the remaining terms of the perturbative expansion in α_s by emitting gluons which will eventually split into more partons.

The PS can either be dipole-, angular- or p_T -ordered [188, 189]. For the first case, the divergent structure of QCD amplitude can be reproduced in dipole-like emissions without double counting the soft wide-angle emissions. In the second case, the angle of emission with respect to the incoming parton is decreased in each step. For the third, the ordering variable is the module of p_T of the parton. The successive branching is stopped when a certain cut-off scale is reached. While PYTHIA uses the p_T -ordered PS, HERWIG has the possibility of using the dipole- or angular-ordered shower (see Section 4.3.2 for an overview on the MC generators).

Hadronisation

As the PS evolves, the energy and momentum of the involved particles decrease until reaching a point in which the confinement occurs and hadrons are formed. The hadronisation happens around 1 GeV and combines the coloured partons into colour-neutral hadrons [178]. The hadronisation is a non-perturbative process and, hence, QCD-inspired models are used to simulate it. Hadronisation operates on the foundation of the parton–hadron duality hypothesis [190]. This entails that the exchange of momentum and quantum numbers at the hadron level should adhere to the same principles as observed at the parton level. Therefore, partons are joined together to form hadrons according to their proximity and phase spaces.

The two predominant theoretical models which are employed to calculate the hadronisation are:

- **Lund-string model** [191, 192]: This model describes the colour interaction between quarks in terms of strings. In the Lund-string model, when a quark and an antiquark are pulled apart, the interaction between them can be visualised as a string. As they move further apart, the potential energy in the string increases. When the energy of the string becomes sufficient, it becomes energetically favourable to produce a new quark–antiquark pair from the vacuum (process known as “breaking” of the string). When the string breaks, it effectively splits into two new strings, each connecting a quark to

an antiquark. In high-energy collisions such as the ones at the LHC, multiple strings can be formed, leading to the production of many hadrons.

- **Cluster model** [193]: The quarks and antiquarks in this alternative model, instead of forming strings like in the Lund-string model, group together to form colour-neutral “clusters” which are the precursor to hadrons. Once these clusters are formed, they are decayed into hadrons. Light clusters typically convert directly into mesons (like pions). Heavier clusters can undergo decays producing multiple hadrons.

The most relevant event generators for PS and hadronisation within this analysis are HERWIG [194], PYTHIA [195], and SHERPA [187]. Here, the fragmentation models are included as well with PYTHIA using the Lund-string model, HERWIG using the cluster model and SHERPA being able to use both.

4.3.1.3 Underlying event

When a hard-scattering subprocess occurs, additional production of hadrons takes place. This production cannot be attributed to the showering of the coloured partons involved in the subprocess. The concept of UE encompasses various phenomena, including pile-up reactions, MPI, and the characteristics of the soft fragments of protons. The parameters used to simulate the UE must be adjusted based on experimental data. The UE can be generated using HERWIG, PYTHIA or SHERPA.

4.3.1.4 Hadron decay

The final stage before introducing the detector geometry in event generation involves the decay of unstable hadrons. These hadrons can be produced in excited and unstable states during hadronisation and can subsequently decay into lighter, stable particles that can be detected within the range of the acceptance of the detector. A particle is considered unstable if its lifetime (τ_{Particle}) satisfies $c\tau_{\text{Particle}} < 10 \text{ mm}$.

The experimental data suggest that many of the detected final-state particles originate from the decay of excited hadronic states. Therefore, most of the decays detailed in the Review of Particle Physics [10] should be considered, along with their respective decay patterns. The high complexity of modelling and implementing this arises from the multitude of potential particles and decay chains involved.

4.3.1.5 Pile-up

The effects of the pile-up have to be simulated as well. This phenomenon is modelled by overlaying over the original hard-scattering event with inelastic $p\bar{p}$ collisions. To do so in the ATLAS experiment, PYTHIA 8 is used to generate the minimum-bias events in the pile-up, i.e. events with low-momentum transfer.

After this is done, the MC-generated events are weighted to reproduce the pile-up distribution provided by the LHC (see Figure 4.4).

4.3.1.6 Simulation of the ATLAS detector, digitisation and reconstruction

The event simulation described so far refers to the generation of physical processes based only on the models. At this stage of the simulation chain, no information about any detector is included yet.

In order to compare the simulated data events⁷ with the real data collected by ATLAS, the response of the detector has to be simulated. This includes the interaction of the many hundreds of particles present in each event with the detector material, as the electronic output of the detector.

To do so and as previously mentioned, the GEANT4 toolkit is used to simulate the passage of the generated particles through the detector. Taking into account the geometry of the ATLAS detector, GEANT4 simulates the effects of both, the magnetic fields and the detector material. Examples of these interactions are energy losses, multiple scattering or photon conversions as well as the hits in the responsive material or interactions with the non-sensitive material.

Afterwards, the electronic signal of the response of each detector is simulated too. This step is known as digitisation and it is performed within Athena framework. The ATLAS digitisation software transforms the hits generated by the primary simulation (from the hard scattering to the interaction with the detector material) into responses of the detector named “digits”. A digit is generated when the voltage or current on a specific readout channel of a detector surpasses a predetermined threshold within a certain time frame. The detector noise is added in this step. The simulated digital output is used to reconstruct the physical objects of the MC event. The reconstruction procedure is done identically for real data and generated simulation. Before performing the digitisation the triggers are simulated as well.

⁷Although in this document are used the terms “real detector data” and “MC simulated data”, it is usual to refer to these as “data” and “simulation”, respectively.

The MC events that have undergone the complete simulation chain are designated as “reconstruction-level” events. Later, for the lepton-origin assignment (Section 6.4.2) these denominations are used when comparing the information of a single event at different levels.

Full and fast simulations of the ATLAS detector

A meticulous simulation of the response of the ATLAS detector is of paramount importance to obtain accurate results. Nevertheless, given the high collision rate that takes place within the LHC experiments, conducting a comprehensive full-detector simulation considering all the events presents significant challenges and can be computationally intensive. In some instances, the use of a simplified and fast approach is enough for the analysis. Hence, there are two proposed detector MC simulation techniques: Full Simulation (FS) [184] and Atlfast-II Fast Detector Simulation (AFII) [184, 196].

The FS strategy relies on a complete detector description powered fully by the GEANT4 toolkit. This thorough simulation demands extensive CPU processing time for each event, especially within the calorimeters. Alternatively, the AFII strategy considers a parametrised cell response to simulate the particle-energy response and the energy distribution in the ATLAS calorimeter [196, 197]. This allows AFII to deliver a valid and good-performing simulation within the computing limits of the collaboration. For the ID and the MS, AFII uses the full GEANT4 simulation.

4.3.2 MC generators

The different steps employed in the event simulation chain make use of various MC event generators. The most relevant ones for this thesis are mentioned here. Later, on Section 6.2, the specific generator for each simulated sample are presented.

- **MadGraph5_aMC@NLO:** Short for “Matrix element Automatic Generator”, is a NLO generator for producing the \mathcal{M} for events following various schemes (e.g. $2 \rightarrow 1$, $2 \rightarrow 2$, $2 \rightarrow 3$ or $2 \rightarrow 7$). These schemes correspond to the multiplicity of initial and final particles in the Feynman diagram. For instance, the $2 \rightarrow 2$ and $2 \rightarrow 3$ correspond, respectively, to the 5FS and 4FS described in Section 2.3.3.1. Given the process, MADGRAPH5_AMC@NLO automatically creates the \mathcal{M} for all the subprocesses and produces the mappings for the integration over the phase space. After this step, the event can be transferred to a PS generator such as HERWIG or PYTHIA.

- **Powheg Box**: Short for “Positive Weight Hardest Emission Generator”, it is another NLO generator based on the \mathcal{M} following various schemes (e.g. 5FS, 4FS or $2 \rightarrow 4$). Here, the μ_R and μ_F are set to be equal to the p_T of the hard partons. This generator was built to implement the same physics as MADGRAPH5_AMC@NLO but outputting positively weighted events only and, hence, escaping the challenge of dealing with negatively-weighted events. Nevertheless, in some cases, the negative weights appear in the generation with (e.g. t -channel process) [198].
- **Herwig 7** [194, 199]: Its name comes from “Hadron Emission Reactions With Interfering Gluons”. This flexible generator has a huge diversity of QCD processes which includes \mathcal{M} and PS simulation at LO and NLO. Even though it can generate the hard-scattering process, in the analysis carried in this thesis it is employed for PS simulation. By default, the PS is ordered using either angular- or dipole-ordered distribution (the default is angular). HERWIG 7 makes use of the cluster model for hadronisation, and an eikonal multiple-interaction model for UE [193].
- **Pythia 8** [195, 200]: Based on \mathcal{M} at LO, this general-purpose event generator implements various calculations (e.g. $2 \rightarrow 1$ and $2 \rightarrow 2$). Contrary to HERWIG, the ISR and FSR are matched in p_T -ordered in the PS. Regarding the hadronisation, PYTHIA 8 uses the Lund-string model. It can also compute the UE using a multiple-interaction model [201]. Most of the PS and pile-up in this analysis are simulated using PYTHIA 8.
- **Sherpa** [187, 202]: Standing from “Simulation of High-Energy Reactions of Particles”. It is designed to provide a comprehensive approach to the simulation of particle collisions, like HERWIG and PYTHIA, being able to simulate from the \mathcal{M} of the hard-scattering process to the PS and hadronisation. For the hard scattering, the events can be constructed with $2 \rightarrow 2$ processes. The PS is based on the Lund model and it is interfaceable with PYTHIA 8.
- **MadSpin** [203, 204]: This generator considers the generation of NLO events involving heavy resonances. In this thesis, the processes with a single top quark are decayed at LO using this generator because it preserves spin correlations.
- **EvtGen** [205]: This MC event generator simulates the decays of heavy flavour hadrons, primarily B , and D -type mesons. It contains a range of decay models for intermediate and final states containing scalar, vector and tensor mesons or resonances, as well as leptons, photons and baryons.

Chapter 5

Object reconstruction and identification

El dos después del uno.

—ISABEL VAL

The event reconstruction consists of the local pattern recognition (i.e. the clustering and resolving of readout channels on the readout detector elements), reconstruction of tracks, segments, vertices, cells and clusters in the different subdetectors, and finally the creation of high-level physical objects, such as particles of different identification, jets including their flavour tag, or missing energy estimation.

To reconstruct the physical objects, the information of all the subdetectors and systems of the ATLAS machine is employed. A detailed description of all of these subsystems is presented in Section 3. After passing the trigger selection, the raw data are analysed to build the physics objects that constitute the basic elements of any physical analysis. The process of identifying and reconstructing these objects is described in this chapter. Figure 5.1 illustrates how each particle interacts with the different layers of the ATLAS detector. The physical objects reconstructed and identified are the particle tracks and vertices (see Section 5.1), the leptons (see Section 5.2), the photons, jets and their flavour (see Section 5.3), and the missing transverse momentum (see Section 5.4). To avoid the reconstruction of the same

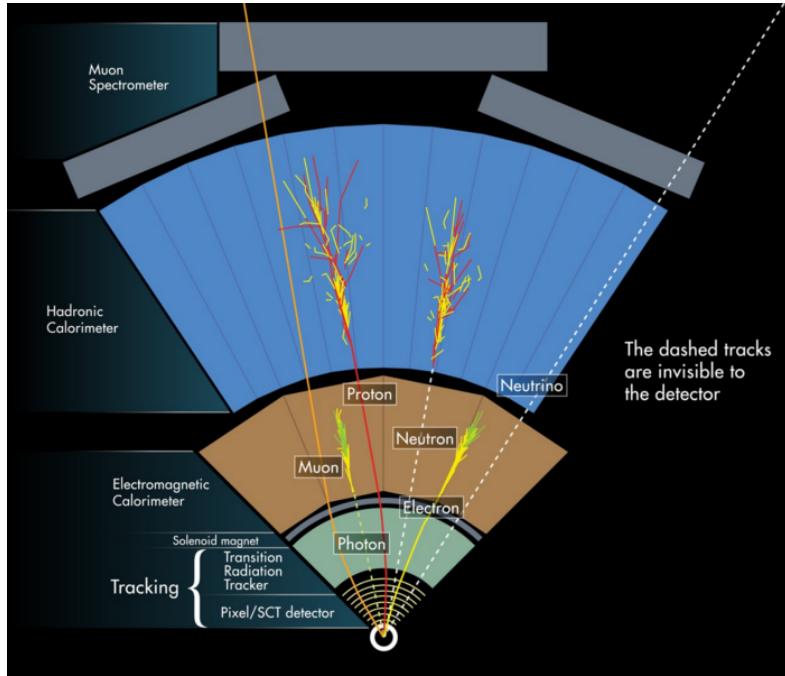


Figure 5.1: Fraction of the transversal plane of ATLAS. Each particle leaves a different signature in each layer [206]. By signature is meant the particular distribution of energy deposition. This scheme is fundamental to understand the object reconstruction in this chapter.

energy deposits as multiple objects, overlap removal criteria (see Section 5.5) are used.

5.1 Tracks and vertices

The detection and measurement of the momentum of the charged particles is an essential aspect of any large particle physics experiment. Regardless of the medium through which a charged particle travels, it always leaves a trail of ionised atoms and liberated electrons. By detecting this it is possible to reconstruct the trajectory of a charged particle.

The trajectories followed by particles are referred to as “tracks”. For charged particles, the tracks are reconstructed using, mainly, the information of the ID (see Section 3.3.2) and, in the case of muons, the MS is also used (see Section 3.3.4). A charged particle passing through the ID will interact with its active sensors, the pixel detector and SCT providing three-dimensional measurements as space-points. While each two-dimensional hit in the pixel detector is directly translated into a space-point, for the SCT two one-dimensional hits are needed to reconstruct one

space-point. These space-points can be given by a single pixel activation or by several neighbouring pixels activated simultaneously (named cluster). Since the ID is immersed in a solenoidal magnetic field, the charged particles have their trajectories curved by the Lorentz force, this allows to accurately calculate its p_T using the sagitta method [148].

The algorithms described in Section 3.4 are of fundamental importance to reconstruct high-quality tracks. This reconstruction is performed in two stages, the inside-out and the outside-in procedures [207]. The first is initiated from the centre of the ID and works outwards. This method is also used for the reconstruction of the primary vertex. The inside-out algorithm starts by grouping the hits in the Pixel and SCT and merging them into clusters that are used to define the space-points. Secondly, the space-points are combined in groups of three to form the track seeds. Then, a pattern-recognition algorithm named Kalman filter [208] is applied to build track candidates from the seeds. This is accomplished by adding extra clusters from the remaining layers of the ID that are compatible with the estimated trajectory of the particles. The Kalman filter provides several track candidates, so an ambiguity-solver algorithm is applied to perform a stringent selection of the candidates. This compares the individual track candidates by measurements of the track quality. Finally, the track candidates are then put through a high-resolution global χ^2 fit, which allows to further rejection of track candidates with a poor fit.

The inside-out method accounts for the majority of tracks reconstructed in the ATLAS detector but it is complemented by the outside-in, which starts in the TRT and works inwards. This method is used to find small-track segments in the ID that were missed.

The identification of the primary vertex is also of crucial importance for the object reconstruction. This vertex identifies the IP in which the hard-scattering process takes place. Therefore, the vertices are defined by relating the origin of the track with individual points. The reconstruction of the vertex is done in two complementary steps. First, the tracks are associated with vertex candidates (vertex finding). Second, an iterative χ^2 fit is used to determine the best final three-dimensional location of the vertex.

Sagitta method

The linear momentum (or just momentum) of a particle (\vec{p}) is one of the most important magnitudes in high-energy physics experiments because it provides information about the energy of that particle. It is possible to determine the p_T of charged particles by measuring the curvature caused by the magnetic field. In principle, particles should have a straight trajectory but the magnetic field (B)

curves its trajectory. The p_T relates to the bending radius (r) by the Lorentz force:

$$m \frac{v^2}{r} = vqB,$$

from which one derives

$$p_T = rqB,$$

where q is the electrical charge of the particle and v its speed. The r is determined using the arc length (l) and the sagitta (s), which is the distance from the centre of the trajectory arc to the midpoint of its chord. Figure 5.2 shows in red the definition of sagitta. The radius is deduced by:

$$r^2 = (l/2)^2 + (r - s)^2 \rightarrow r = \frac{(l/2)^2 + s^2}{2s}.$$

For high p_T particles $s \ll r$ and, hence, it is possible to approximate $r \sim \frac{l^2}{8s}$. The main uncertainty on p_T is the uncertainty on the sagitta and it can be modelled with a Gaussian distribution.

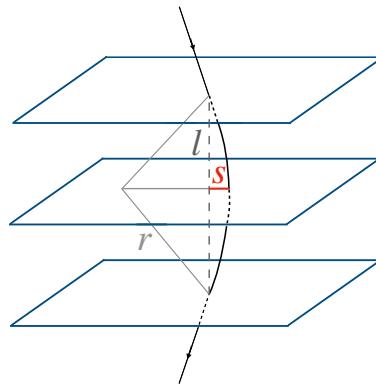


Figure 5.2: The arc represents the path of the particle. The layers of the tracker are drawn in blue. With the sagitta (red) and the arc length, the radius of curvature can be determined. The more energetic a particle is, the larger is its bending radius.

5.2 Charged leptons

The reconstruction of the charged leptons is a fundamental piece of many analyses, including the one developed in this thesis where the final state consists of two light-flavoured-charged leptons (i.e. e/μ) and one τ -lepton decaying into hadrons. Therefore, the identification and reconstruction of the three leptons flavours is a key part of this thesis and is described through the following subsections.

5.2.1 Electrons

The reconstruction of electrons¹ and photons is accomplished through the identification of energy deposits in the ECAL. For the electrons, particle tracks recorded in the ID are required [209, 210].

Electrons

In the analysis presented in this work, there are two final-state light-leptons that can be electrons. Therefore, accurate and efficient electron identification is crucial to measure our process of interest. Figure 5.3 presents a schematic representation of the ingredients composing the process of electron reconstruction and identification. When an electron travels through the detector, it leaves traces in the ID and energy deposits in the ECAL. The calorimeter signal activates the LVL1 trigger and electron candidates are selected from an initial match between the ECAL energy clusters and the ID tracks. The track associated with the electron must also pass on requirements on the longitudinal impact parameter $z_0 \cdot \sin \theta < 0.5$ mm and the transverse impact parameter $\frac{d_0}{\sigma(d_0)} < 5$ where $\sigma(d_0)$ is the uncertainty on d_0 .

A typical electron candidate is expected to generate on average 12 hits in the inner tracker system, which includes one hit in the IBL layer, three hits in the silicon pixel layers, and eight hits in the SCT (4 double-sided silicon-strips layers). Furthermore, approximately 35 straw hits are produced in the TRT system for an electron of p_T larger than 500 MeV. Finally, the electron moves to the ECAL, where the majority of its energy is collected by the second layer.

The first step in the electron reconstruction is to build the clusters in the calorimeters. To do so, the space in the ECAL is divided into small elements of dimension $\Delta\eta \times \Delta\phi = 0.025 \times 0.025$ that combine the subdetector layers. These elements are called towers. A presampler in the $|\eta| < 1.8$ region also gathers the energy and, along the first three layers of the ECAL, is used to determine the total energy per tower. Clusters are seeded by individual towers with energy above 2.5 GeV and are searched for within the ECAL middle layer. Once the candidate clusters have been established, the next step is to associate them with the tracks reconstructed in the ID using the tracking algorithms.

When multiple tracks can be linked to a specific EM calorimeter cluster, it is necessary to designate a primary electron track. This selection is performed through an algorithm that evaluates the $\eta\phi$ distance between the extrapolated

¹Note that the term electrons is used to collectively refer to electrons and positrons.

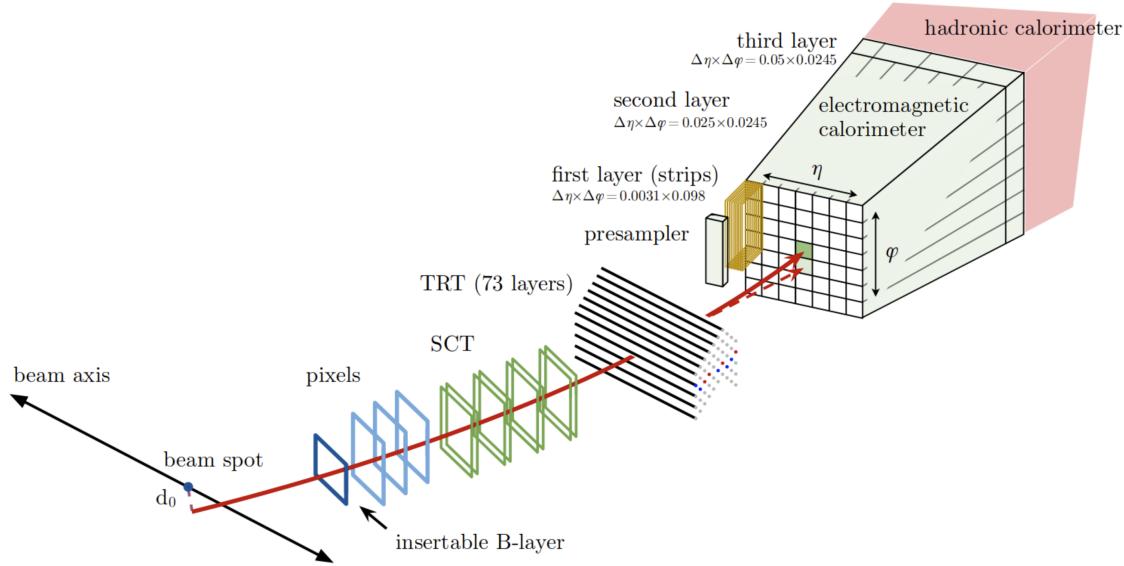


Figure 5.3: Trajectory of an electron through the detector. The hypothetical path of the electron is represented by a solid red line, while the trajectory of a bremsstrahlung photon generated in the tracking system material is represented by a dashed red line.

tracks and the cluster barycentre and considers the quantity of hits in the silicon detectors and the number of hits in the innermost silicon layers.

Electrons may arise from either the primary hard-scattering event, such as the decay products of W or Z bosons (referred to as prompt electrons), or as the decay products of secondary particles with relatively long lifetimes, such as b -hadrons (these are the so called non-prompt electrons). An example of non-prompt electron is presented in Figure 5.4. The identification of prompt electrons is achieved through the use of a likelihood discriminant constructed from measurements taken in the ID and ECAL. The measured quantities are selected based on their effectiveness in distinguishing prompt-isolated electrons from energy deposits resulting from hadronic jets, from converted photons and non-prompt electrons. The discriminant considers the properties of the primary electron track, the lateral and longitudinal growth of the EM shower in the ECAL, and the spatial compatibility of the primary electron track with the cluster. Different operating points (working points) can be defined by setting fixed values for the likelihood discriminant. These are **tight**, **medium** and **loose**. The **tight** category is the most stringent, while the **loose** category is much more permissive in terms of accepting something as an electron. In Section 6.3, these categories are used to define the trigger selection and the electron definition used in this analysis.

Regarding its charge, it is identified by the curvature of its track under the magnetic field. If the curve of the trajectory is not very pronounced, it can lead

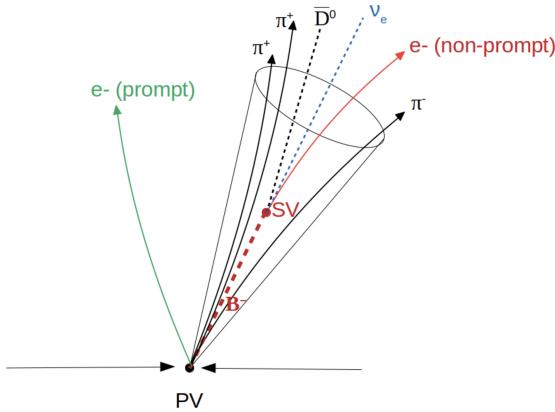


Figure 5.4: A prompt electron depicted in green. The cone symbolises a jet containing several hadrons. The dashed red line corresponds to a b -hadron (B^-), which decays into a c -hadron (\bar{D}^0), a neutrino (ν_e), and a non-prompt electron (red). The non-prompt electron is originated from the secondary vertex while the prompt from the primary vertex.

to misidentification of the charge. This phenomenon, known as charge flip, plays a role in the analysis described in this thesis by being a relatively relevant source of background (as it is later presented in Table 6.18). The charge flip can also emerge from bremsstrahlung radiation.

Additionally, to differentiate the prompt electrons in signal processes from the background misidentifications, isolation requirements are applied to the tracks and the calorimeter clusters. These are described in Reference [211].

The estimation of electron energy is derived from energy deposits in the ECAL and is subsequently refined through a series of calibrations to mitigate residual discrepancies between data and simulation [212]. This calibration process includes the inter-calibration of the multiple calorimeter layers, corrections for energy shifts caused by pile-up, corrections that improve the uniformity of the energy response and changes to the overall energy scale. Each correction is applied to the data to enhance accuracy. Additionally, a correction to account for the difference in energy resolution between data and simulation is also applied to the simulated samples. The different uncertainties associated with this calibration stage are considered in the ATLAS analyses.

Finally, the electron-identification efficiency is determined for both real and simulated data using a data-driven technique known as the tag-and-probe method [213]. The tag-and-probe is employed with the $Z \rightarrow e^- e^+$ and $J/\psi \rightarrow e^- e^+$ decay modes. The tag-and-probe method relies on the preparation of an unbiased sample of physics objects, the *probe* objects, which are used to calculate efficiencies and resolutions. The data sample is selected by means of an independent *tag* object [214].

Photons

The process of photon reconstruction closely mirrors that of electron reconstruction, with the primary distinction being the absence of tracks in the tracker, unless a photon undergoes conversion into an electron-positron pair, in which case the corresponding tracks must be retrieved.

The identification working points are established with the ECAL information. The distinction between prompt photons and background photons is achieved by applying selections based on quantities that characterise the shape and properties of the corresponding EM shower, as well as by implementing isolation criteria for the photon candidate.

5.2.2 Muons

Muons can pass through every component of the ATLAS detector without being fully absorbed in the ECAL. Therefore, the reconstruction of muon candidates within the ATLAS experiment involves a combination of information from the ID, the MS, and the calorimeters [215]. Muon candidates with $|\eta| < 2.5$ are considered for reconstruction [216]. In the MS, track reconstruction is accomplished by grouping hits into local track segments using a Hough transform [217]. These segments are then merged to form track candidates, and a fitting procedure is employed to determine the trajectory of the muon within the magnetic field. Depending on the subdetectors involved in the muon reconstruction process, different types of muons can be identified:

- Combined muons: This type of muon is identified by matching MS tracks to ID tracks and performing a combined track fit using the hits from both systems. The energy loss in the calorimeters is taken into account during the fitting process. This definition works only for muons within $|\eta| < 2.5$, i.e. the ID acceptance.
- Inside–out muons: An *inside–out* algorithm is utilised to reconstruct this category of muons. It extrapolates ID tracks to the MS and searches for at least three aligned MS hits, which are then used for a combined track fit.
- Muon-spectrometer extrapolated: These muons arise when a MS track cannot be matched to an ID track. These muons are reconstructed in a range $2.5 < |\eta| < 2.7$ where the ID does not cover. In such cases, the parameters of the MS track are extrapolated to the beam pipe to define the reconstructed muon. This type of muon object is also referred to as stand-alone muon.

- Segment-tagged muons: This group of muons is identified by extrapolating ID tracks to the MS and searching for matching segments. A muon is considered segment tagged if an ID track is successfully matched to at least one MS segment, and the muon parameters are directly obtained from the ID track fit. This allows an increase in the acceptance of muons which have crossed only one layer of the MS chamber.
- Calorimeter-tagged muons: In this scenario, muons are identified by extrapolating ID tracks through the calorimeters to search for energy deposits consistent with those of a minimum-ionising particle, i.e. a particle whose mean energy loss rate through matter is close to the minimum value. If a match is found, the muon is identified as calorimeter-tagged, and its parameters are again obtained from the ID track fit. This type of reconstructed muons recovers acceptance in the region where the MS is only partially instrumented.

Prompt muons are identified by applying specific requirements on the number of hits in the ID and the MS, track-fit properties, and variables that test the compatibility between measurements in the two systems. Similarly to the electrons, the identification of muons requires that $z_0 \cdot \sin \theta < 0.5$ mm and $\frac{d_0}{\sigma(d_0)} < 3$. The stringency of these requirements leads to three primary working points: **tight**, **medium** and **loose**. As for electrons, the **tight** muons must meet more stringent requirements than the **loose** muons. The segment-tagged and calorimeter-tagged muons are considered in the loosest definition. Additionally, two working points are designed for extreme phase space regions: the high- p_T working point, which ensures optimal momentum measurement for muons with $p_T > 100$ GeV, and the low- p_T working point, which addresses muons less likely to be fully reconstructed as tracks in the MS due to their low momentum.

To distinguish prompt muons objects from the misidentified muons, isolation criteria are applied. By demanding that $\Delta R = 0.2$ cone around the tracks in the ID and around the energy clusters in the calorimeters, the non-prompt muons can be rejected [216].

Corrections have been made to the simulated muon momentum scale and resolution to address the observed discrepancies between the experimental data and simulation. These discrepancies are studied using $Z \rightarrow \mu^-\mu^+$ and $J/\psi \rightarrow \mu^-\mu^+$ decays. After implementing the corrections, data and simulation agree to the per mille level for the muon momentum scale and to the percent level for the muon momentum resolution. The uncertainties arising from these corrections are quantified and propagated to the different ATLAS analyses.

5.2.3 Hadronically decaying taus

On the one hand, the leptonically-decaying τ -leptons (τ_{lep}) cannot be differentiated from prompt electrons or muons because τ -decay takes place within several millimetres of the IP, i.e., before reaching the detector. Therefore, the τ_{lep} s are not identified. On the other hand, identifying hadronically-decaying τ -leptons (τ_{had}), while possible, is a challenging task. In this section, the identification and reconstruction of τ_{had} is described.

The τ -lepton, being the most massive known lepton, exhibits a lifetime of approximately 2.9×10^{-13} s [218], which corresponds to an averaged travelled distance of 87 μm in vacuum (assuming a momentum similar to the τ mass). It predominantly decays into final states consisting of hadrons, accounting for approximately 65% of its total decay modes. When doing so, it produces a τ -jet containing a small number of charged and neutral hadrons as it is shown in Figure 5.5. The formation of jets is described in Section 5.3.

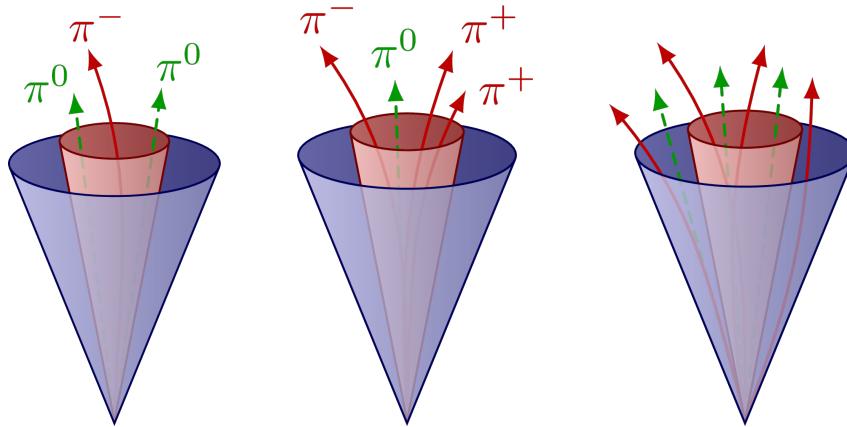


Figure 5.5: Isolation (blue) and signal cones (red) of hadronically-decayed taus in the left and centre images. For comparison, a quark/gluon-initiated jet is presented on the right. Note how the τ -jets are more collimated.

Hadronic decays of τ -leptons exhibit distinct properties in terms of vertex displacement, track multiplicity, and kinematics. When the momentum of the τ -lepton is large compared to its mass, a very collimated jet (the reconstruction of jets is described in Section 5.3) is produced. For example, if the τ -lepton carries a transverse momentum $p_T(\tau) > 40 \text{ GeV}$, around 90% of its energy is constrained within a cone of radius $\Delta R = 0.2$. Another relevant property is that the τ -lepton decays exhibit a low charged-track multiplicity of just one or three prongs². Also, a rel-

²The term “prong” refers to the number of charged particles in the final state of the tau decay. As most detectors use a tracking chamber to identify charged particles, tau decays can be classified into those that provide a single track (1-prong) and those that provide three tracks (3-prong).

event fraction of the EM energy deposition in the calorimeters is due to photons coming from the decay of neutral pions. In Table 5.1, the main τ -decay BRs are shown. All these properties are exploited for the τ_{had} identification.

Decay Mode	Fit Result (%)
$\mu^- \bar{\nu}_\mu \nu_\tau$	17.3937 ± 0.0384
$e^- \bar{\nu}_e \nu_\tau$	17.8175 ± 0.0399
$\pi^- \nu_\tau$	10.8164 ± 0.0512
$\pi^- \pi^0 \nu_\tau$	25.4941 ± 0.0893
$\pi^- 2\pi^0 \nu_\tau$	9.2595 ± 0.0964
$\pi^- 3\pi^0 \nu_\tau$	1.0429 ± 0.0707
$\pi^- \pi^- \pi^+ \nu_\tau$	8.9868 ± 0.0513
$\pi^- \pi^- \pi^+ \pi^0 \nu_\tau$	2.7404 ± 0.0710
$\pi^- \omega \nu_\tau$	1.9494 ± 0.0645

Table 5.1: Most relevant τ decay modes as listed by Reference [10].

The τ_{had} reconstruction involves the use of the anti- k_t algorithm [219–221] with radius $R = 0.4$. See Section 5.3 for more details about the anti- k_t algorithm. Local hadronic calibration is used to calibrate the topological clusters that serve as input to the algorithm. The τ_{had} candidates are required to have $p_{\text{T}} > 5$ GeV and $|\eta| < 2.5$ [222].

It is important to properly identify the primary vertex, as τ leptons travel away from the primary vertex before decaying. To ensure optimal performance in high-pile-up scenarios, the Tau Jet Vertex Association [223] algorithm is employed to determine the primary vertex of the τ_{had} . This minimises the influence of additional interactions, which could potentially lead to tau tracks failing to meet the z_0 impact parameter requirement [224].

Tracks are associated with the τ_{had} if they are in the core region $\Delta R < 0.25$ around the τ_{had} direction and satisfy the following criteria: $p_{\text{T}} > 1$ GeV, at least two associated hits in the pixel layers of the ID, and at least seven hits in total in the pixel and the SCT layers. The τ_{had} candidates are categorised as 1-prong or multi-prong (primarily composed of three tracks). To correctly establish the charge and the number of charged decay products of a τ_{had} , the tracks that are associated with the τ_{had} decay need to be correctly distinguished from the rest. The classification of the tracks associated with a τ candidate is done using recurrent neural network (RNN) based on Keras and TensorFlow with three connected dense layers [222, 225]. The RNN categorises tracks associated with τ -jet candidates into four categories: **Tau Tracks**, **Conversion Tracks**, **Isolation Tracks**, and **Fake Tracks**. **Tau Tracks**, are tracks from charged decay products of the tau lepton,

essential for determining the charge and number of decay products. **Conversion Tracks** are from electrons or positrons resulting from photon conversion, used in identifying pi-zero mesons and rejecting truth electrons reconstructed as tau candidates. **Isolation Tracks** likely originate from quark or gluon jets, characterised by higher multiplicities and a softer transverse momentum spectrum. **Fake Tracks** include tracks that do not fit into the other categories, mainly misreconstructed or pile-up tracks.

Afterwards, a RNN-based τ -identification algorithm is trained using simulated samples of τ_{had} candidates [222, 225]. The architecture of this RNN is the same as the one for track classification. The training varies according to the prongness of the τ_{had} candidate. The RNN uses a combination of low-level input variables for individual tracks and clusters that are associated with the τ_{had} candidate as well as several high-level observables calculated from track and calorimeter quantities. The working points of the RNN for τ_{had} identification are presented in Table 5.2.

Truth τ_{had} efficiency (%)		
Working point	1-prong	3-prong
Tight	60	45
Medium	75	60
Loose	85	75
Very Loose	95	95

Table 5.2: List of defined working points with fixed truth τ_{had} selection efficiencies for the RNN classifier [222].

5.3 Jets

When a quark or gluon is produced in high-energy processes, it cannot exist in isolation due to the colour confinement as it is stated in Section 1.4. An exception to this rule are the top quarks, whose lifetime is smaller than the hadronisation time by two orders of magnitude and, hence, they are detected by their decay products. For the gluons and the rest of quarks, hadronisation showers take place and jet-clustering algorithms merge the clusters and tracks produced by these jets to reconstruct them. There are two algorithms to reconstruct jets: the cone algorithms [226] and the sequential clustering algorithms [219].

In the majority of ATLAS analyses, the sequential clustering “anti- k_t ” algorithm is used [219–221] to analyse the data from hadronic collisions. To model a jet as a cone, this algorithm uses a specific choice of radius parameter (R) defining the

radial size of the jet. The distance between all pairs of objects i and j (d_{ij}) and the distance between the objects and beam pipe (d_{iB}) are used in:

$$d_{ij} = \min(p_{Ti}^{2f}, p_{Tj}^{2f}) \frac{\Delta R_{ij}^2}{R^2},$$

$$d_{iB} = p_{Ti}^{2f},$$

where

$$\Delta R_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$$

and p_{Ti} , y_i and ϕ_i are respectively the transverse momentum, the rapidity and the azimuthal angle of object i . The parameter f accounts for the relative power of the energy versus geometrical (ΔR_{ij}) scales. For the anti- k_t , f is set to -1 . Other clustering algorithms use different choices of f such as $f = 0$ (Cambridge/Aachen algorithm [20]) or $f = 1$ (inclusive p_T algorithm [219]).

The algorithm iterates over the topological-cluster objects of the calorimeter as it follows: first it proceeds to identify the smallest distances with among all the combinations of d_{ij} and d_{iB} . If the distance is a d_{iB} , the entity i is labelled as “jet” and removed from the list of entities. If, on the contrary, it is a d_{ij} , the objects i and j are merged together. This way, before clustering among themselves, soft components (low- p_T) tend to be merged to the hard ones (high- p_T). Then the distances are recalculated and the process repeated. This is done iteratively until all entities are assigned to a particular jet.

If a particle from the hard-scattering process does not have other hard-scattering particles as neighbours within a $2R$ distance, all soft particles will be assigned to it, resulting in a perfectly conical jet. But if another hard particle is present in that $2R$ distance, then there will be two hard jets and it will be impossible for both to be perfectly conical.

Typically, the cone size R is selected to be 0.4 or 0.6, though the most standard used in ATLAS is 0.4. If $R = 1$, the jet is labelled as Large- R and if $R = 0.4$ then as Small- R jet. Large- R jets are used in boosted analyses and for tagging the top-quark-induced jets or the jets produced from the W , Z and Higgs bosons [227, 228]. In the studies presented in this thesis, only Small- R jets are considered.

5.3.1 Bottom-quark-induced jets

The identification of jets originating from the hadronisation of b -quarks (b -jets) is referred to as b -tagging. The goal of b -tagging is to discriminate b -jets from the jets produced by c -quarks (c -jets) or by gluons or quarks of other flavours (referred to as “light” jets). This identification plays an important role in the analysis

carried in this thesis since at LO there may be two b quarks final state of the tHq production³.

In general, it is a challenging task to determine which quark flavour produced the jet. It is also difficult to distinguish between quark-initiated jets or gluon-initiated jets. However, if a b quark is created, the hadronisation will produce a jet of hadrons, one of which will be a b -type hadron (B hadron). The B weakly hadrons turn out to be relatively-long-lived particles (1.5×10^{-12} s [10]). If this larger longevity is combined with the Lorentz time-dilation that particles experience when produced in high-energy collisions, it results in the B hadron traveling on average a few millimetres before decaying.

As a result, the experimental signature of a b quark is a jet of particles emerging from the point of collision (primary vertex) and a secondary vertex resulting from b -quark decay that is several mm away from the primary vertex as Figure 5.6 shows. Therefore, the capacity to resolve secondary vertices from the parent vertex is crucial for identifying b -quark jets. Other features that are used to identify the b -jets are its high mass, the properties of the b -quark fragmentation and the fact that the decay of a B hadron will on average have a higher charged track multiplicity in the decay than other hadron decays.

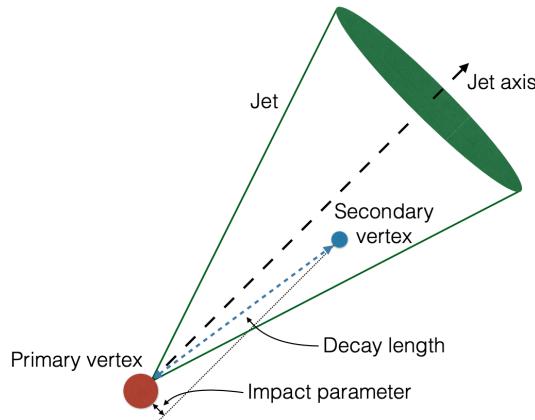


Figure 5.6: Illustration of the production of a b -jet with its characteristic second vertex [229].

The identification of b -jets involves a two-step process. Initially, low-level algorithms are employed to reconstruct the primary characteristics of the b -jets. Subsequently, the outcomes of these algorithms are combined in high-level algorithms

³In the Feynman diagrams in Figures 2.12a and 2.12b the initial-state b quark can be produced from gluon splitting into a $b\bar{b}$ pair (4FS), resulting in one \bar{b} quark (which is generally too soft and too forward to be detected) in the final state. Additionally, the top quark decays into a W -boson and a b quark, producing another final-state b quark.

that consist of multivariate classifiers. The various low-level algorithms can be categorised into three groups:

- Impact-parameter-based algorithms: These algorithms employ the properties of individual tracks associated with a jet. Tracks originating from b -type hadron decay have distinct characteristics, such as large impact parameters (d_0, z_0). Algorithms like IP2D and IP3D [230] use the impact parameter significances of tracks within a jet to distinguish between b -jets and light-jets. Multivariate Analysis⁴ (MVA) methods are used by other algorithms such as the RNN1P algorithm [231], which exploits spatial and kinematic correlations among tracks originating from the same B hadron using a recurrent neural network⁵ (RNN).
- Secondary-vertex-based algorithms: These type of algorithms use information from secondary vertices to create discriminative variables for b -tagging. For instance, SV1 [232] is a likelihood-based tagger that considers the invariant mass of particles in the secondary vertex, the ratio of track energies, the number of two-track vertices, and the ΔR separation between the primary-secondary vertex and the jet direction.
- Decay-chain reconstruction: These algorithms aim to reconstruct the complete decay chain of the B hadron. The JetFitter algorithm [233] is an example of this type. It uses the topology of weak b - and c -hadron decays within the jet. It employs a Kalman filter [208] to find a common line connecting the primary, bottom, and charm vertices, enabling the reconstruction of the flight path of the B hadron and vertex positions.

High-level taggers, such as MV2 and DL1 [230, 231], use the outcomes of low-level algorithms to determine the probability of a jet being classified as a b -, c -, or light-jet. MV2 employs a Boosted Decision Tree (BDT) architecture, while DL1 utilises a Deep Neural Network (NN). These high-level algorithms incorporate input from the IP3D, SV1, and JetFitter algorithms, along with the kinematic characteristics of the jets, including p_T and η . While MV2 is specifically trained to separate b -jets from c - and light-jet, the DL1 algorithm provides a multidimensional output which not only tags the b -jets but also the c - and light-jets.

⁴MVA is a statistical approach that analyses multiple variables together to identify patterns and relationships in data. In this thesis, the MVA methods mentioned/used are the BDTs and NNs.

⁵A neural network (NN) is a mathematical machine learning model composed of interconnected artificial neurones that use activation functions and weights to process input data, perform nonlinear transformations, and learn from training examples through iterative adjustments of the weights to solve various tasks, such as pattern recognition and regression. It performs similar tasks as those of the BDT (see Appendix B).

One of the algorithms that composes the DL1 series is the **DL1r** [231]. Its implementation as a multi-class NN architecture allows for a more compact memory usage compared to the previous BDT-based MV2c10 algorithm [229]. The NN topology is comprised of fully connected hidden layers, and the hyperparameters are optimised to enhance the performance of b -tagging. The ultimate **DL1r** b -tagging discriminant is formulated as follows [234]:

$$D_{\text{DL1r}} = \ln \left(\frac{p_b}{f_c \cdot p_c + (1 - f_c) \cdot p_{\text{light}}} \right) \quad (5.1)$$

where p_b , p_c , p_{light} and f_c represent the b -jet, c -jet and light-flavour jet probabilities, and the effective c -jet fraction in the background training sample, respectively [235]. The b -tagged jets are jets whose values of the **DL1r** are above a certain threshold, hereafter referred to as working points. The efficiency of identifying b -jets [236] and the mis-tag rate of c -jets [237] are measured using $t\bar{t}$ event data. While the calibration of light-jets relies on events that involve a Z boson, following a procedure similar to the one described in Reference [238]. The correction factors derived from these calibration analyses are subsequently applied to correct the simulated events. In Section 6.3.5.1 is discussed how the **DL1r** algorithm is applied in this thesis.

5.4 Missing transverse energy

According to the principle of momentum conservation, the total sum of transverse momenta of all detected particles should be zero in the transverse plane. However, the presence of undetected particles can lead to an imbalance in this calculation, resulting in missing transverse momentum (\vec{E}_T^{miss}). This \vec{E}_T^{miss} is typically associated with SM neutrinos, but it can also occur due to particles escaping the acceptance of the detector or being poorly reconstructed due to limitations in detector acceptance, finite detector resolution, detector inefficiencies presence of (temporary or permanent) dead regions, or any sources of noise.

Therefore, the measurement of the magnitude of the missing transverse momentum (E_T^{miss}) plays a crucial role in various analyses, such as the study of top-quark polarisation (see Section 2.1.4.1), which involves a final state with one neutrino (as in the search for the tHq process presented in this work) or more neutrinos.

The E_T^{miss} is reconstructed by calculating the magnitude of the negative vector sum of the transverse momenta of all detected particles [239, 240]. This includes

contributions from various particles such as leptons, photons, jets, and soft-event signals⁶.

5.5 Overlap removal

In ATLAS, the reconstruction of physics objects within the detector is primarily performed independently for each object. Ambiguity in the observed detector signatures can lead to the double-counting of signals, causing multiple physics objects to be defined simultaneously. This is what is known as overlap. For instance, charged particle tracks in the inner tracking volume accompanied by energy deposits in the calorimeters could be interpreted as both an electron and a hadronic jet. This ambiguity can arise from misidentification and duplication of objects or from the production of particles in close proximity (non-isolation), which may bias the reconstruction of one or both objects.

To address these reconstruction ambiguities, overlap removal is an essential step in all ATLAS analyses. In this analysis, the overlap removal is implemented based on the geometric proximity (ΔR) between reconstructed objects. In Section 6.3.6, the details of how this is done are provided. However, a recent advancement in the ATLAS core software introduced Global Particle Flow (GPF) links between jet constituents and physics objects that share a common detector element, such as a track or calorimeter cluster. These GPF links offer a cleaner approach for removing overlaps between jets and other physics objects. By explicitly examining shared detector signals among physics objects, it becomes possible to identify instances of double-counting of energy. If such overlaps are found, a set of criteria can be applied to determine which objects should be vetoed, ensuring accurate event reconstruction.

⁶Soft-event signals refer to reconstructed charged-particle tracks that are associated with the hard scattering vertex but not with any specific hard object.

Chapter 6

Search for the tHq production with a τ_{had} in the final state

Cinquanta quilos pesa el xino.

—RAFAEL AGULLÓ-IRLES

This analysis aims to set the best estimation of the tHq SM cross-section in the $2\ell + 1\tau_{\text{had}}$ channel¹ using a dedicated search². In the ongoing analysis within the ATLAS collaboration, the $y_t = -1$ hypothesis is also studied. To explore the inverted coupling hypothesis, the complete tH production (i.e. the $tHq + tWH$ processes) should be studied as both processes are sensitive to the sign of y_t (see Section 2.3.3.2). However, a complete study of $y_t = -1$ is outside of the scope of this thesis. Here, just the $y_t = -1$ hypothesis for the tHq is considered.

This chapter is organised as follows: Section 6.1 introduces the different channels in which the ongoing effort within the ATLAS collaboration is divided. Section 6.2 provides an overview of the detector-collected datasets and MC-event-simulation samples used in this analysis. In this section is also discussed how each

¹The tHq production with final states involving τ_{had} was also studied by one of my collaborators, being publicly available in her PhD thesis dissertation [241].

²Note that in the already published searches shown in Table 2.1 there is just one dedicated tHq study, it uses partial Run 2 data, and it does not consider final states with hadronically decaying taus.

background process affects this analysis. In Section 6.3, the specific object definition employed in the analysis is discussed. Section 6.4 explores various aspects of the signal process, including the generation of parton-level information, the assignment of light-lepton origin, and the event reconstruction. In Section 6.5, a description of the estimation of the background yields is presented. Section 6.6 covers the signal separation, along with the definition of control and validation regions, and outlines the MVA methods used. Later on, a comprehensive treatment of systematic uncertainties is provided in Section 6.7. Finally, the likelihood fit procedure and the resulting outcomes are presented and discussed in Section 6.8.

6.1 Channels of the search of the tHq process

The study of the tHq production can be classified according to the number of light-flavour leptons (ℓ), i.e. electrons or muons, and the number of τ_{had} in the final state. The study of the 1ℓ channel uses only the $tHq(H \rightarrow b\bar{b})$, which is the most dominant decay mode for the Higgs boson with a 58% BR as reported in Figure 2.8. There are other Higgs-boson-decay channels that can provide the 1ℓ signature but these either have a small cross-section or their backgrounds are much larger. For the multileptonic (ML) final-state channels, the $H \rightarrow WW^*$, $H \rightarrow \tau^-\tau^+$ and $H \rightarrow ZZ^*$ decay modes are considered. These three Higgs-boson-decay modes combined account for a total of 31% BR (see Figure 2.8). Even though $H \rightarrow \mu^-\mu^+$ could also provide a ML final state, it is not considered due to its low BR. The same happens with the other decay modes that, due to their small cross-section, are considered irrelevant.

Therefore, the considered ML channels are 2ℓ SS, 3ℓ , $2\ell + 1\tau_{\text{had}}$ and $1\ell + 2\tau_{\text{had}}$. The former two are explored in References [242, 243]. Table 6.1 presents the different channels being currently studied by the ATLAS collaboration. Some gaps, represented by “—”, can be seen in Table 6.1. These correspond to the channels with more than three leptons in the final state as well as the 2ℓ OS and $1\ell + 1\tau_{\text{had}}$. The reason to not include the 2ℓ OS channel in the ongoing search by the ATLAS collaboration is because the $t\bar{t}$ background is so large that the sensitivity of that channel is negligible. Similarly, a huge background makes the $1\ell + 1\tau_{\text{had}}$ to have very poor sensitivity. Finally, events with more than three leptons are discarded from these analyses because they represent a small fraction of the total tHq production.

For each ML final-state channel, there is a different probability distribution of being produced based on a particular Higgs-boson-decay mode as Table 6.2 shows. For instance, the $H \rightarrow \tau\tau$ is the most likely decay mode in the $1\ell + 2\tau_{\text{had}}$ channel but not in the 3ℓ channel.

	0 τ_{had}	1 τ_{had}	2 τ_{had}
1 ℓ (e/μ)	$tHq (H \rightarrow b\bar{b})$ 1 ℓ	—	$tHq (WW^*/ZZ^*/\tau\tau)$ 1 $\ell + 2\tau_{\text{had}}$
2 ℓ (e/μ)	$tHq (H \rightarrow WW^*/ZZ^*/\tau\tau)$ 2 ℓ SS	$tHq (H \rightarrow WW^*/ZZ^*/\tau\tau)$ 2 $\ell + 1\tau_{\text{had}}$	—
3 ℓ (e/μ)	$tHq (H \rightarrow WW^*/ZZ^*/\tau\tau)$ 3 ℓ	—	—

Table 6.1: Different channels that are being studied in the ongoing analysis by the ATLAS collaboration for the tHq production. The classification is done according to the presence of light-flavoured leptons and hadronically-decaying taus in the final state. The $2\ell + 1\tau_{\text{had}}$ channel is partitioned into two subcategories depending on the relative sign of the electric charge exhibited by the two charged light leptons: 2ℓ OS + $1\tau_{\text{had}}$ and 2ℓ SS + $1\tau_{\text{had}}$. The multiplicity requirements on each final-state objects ensure the orthogonality between the various channels.

Channel	Probability (%)		
	$H \rightarrow \tau\tau$	$H \rightarrow WW^*$	$H \rightarrow ZZ^*$
2 $\ell + 1\tau_{\text{had}}$	63	32	5
1 $\ell + 2\tau_{\text{had}}$	96	3	1
2 ℓ SS	17	80	3
3 ℓ	14	69	17

Table 6.2: Percentage probability of the Higgs-boson-decay modes within different final-state ML channels. The numbers are obtained with MC-simulated samples at parton level. The used events are those passing the requirement defined by the topology of each channel. The statistical uncertainty is smaller than 0.1% for all the values in this table. Note that these percentages are normalised to consider only these three decay modes.

The $2\ell + 1\tau_{\text{had}}$ channel is further subdivided in two sub-channels depending on the charge of the light-charged leptons. The so-called 2ℓ SS + $1\tau_{\text{had}}$ channel is defined when the two light leptons have the same electric charge. In contrast, when there are two light leptons with opposite electric charge, the 2ℓ OS + $1\tau_{\text{had}}$ channel is defined.

The work presented in this thesis is focused on the two $2\ell + 1\tau_{\text{had}}$ sub-channels, which are treated separately since these two sub-channels have different background compositions, being the 2ℓ SS + $1\tau_{\text{had}}$ the one with the lower background contribution and, therefore, higher sensitivity. The Feynman diagrams illustrating these two processes are depicted in Figure 6.1. Although the diagrams exhibit a resemblance for both final-state channels, the challenges encountered during both analyses differ significantly.

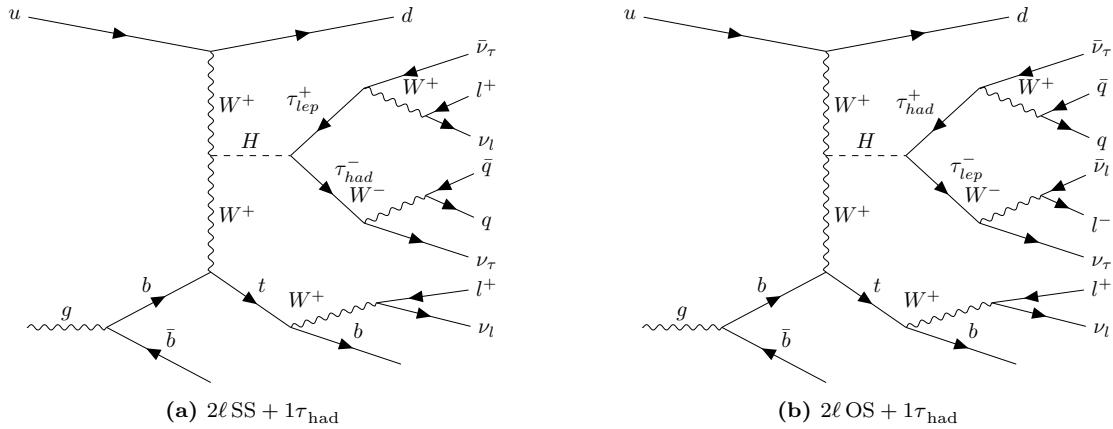


Figure 6.1: Representative LO Feynman diagrams (4FS) for the tHq ($2\ell + 1\tau_{\text{had}}$) in the $H \rightarrow \tau_{\text{had}}\tau_{\text{lep}}$ decay channel (dominant decay mode). Note that the two charged light-flavoured leptons in (a) have the same positive electrical charge and in (b) these leptons have opposite charges.

6.2 Data and simulated events

In this section the particularities of the data collected by the detector and the MC-simulated event samples are presented. The generalities of the data gathering and the production of the MC event samples are described in Chapter 4.

All the search channels share the same object selection and use a common set of MC NTuples, which are produced with the SingleTopAnalysis³ package using the TOPQ1 derivations as input. These TOPQ1 derivations contain a single-lepton filter that requires at least one electron or muon satisfying $|\eta| \leq 2.5$. Additionally, this lepton should have $p_T > 20$ GeV for 2015 data and above 25 GeV for both the 2016–2018 data and the MC event simulations. The produced NTuples have at least two e/μ in their final state, one satisfying the condition above and the other must present a $p_T > 10$ GeV.

After the NTuples are generated with the SingleTopAnalysis package, a post-processing framework named `tHqLoop` which further manipulates the real and simulated data samples to skim and slim⁴ them.

³The SingleTopAnalysis package is a ROOT-based software based on AnalysisTop. AnalysisTop is the standard analysis software within the Athena framework for Run 2 analyses in the Top-quark Working Group.

⁴The slimming is done by removing unnecessary branches. In ROOT, a branch represents a variable associated with each event or entry in a TTree, which is a hierarchical data structure used for storing and analysing data in the ROOT framework.

6.2.1 ATLAS-collected data samples

The real-data samples used in this analysis correspond to the events recorded by the ATLAS detector from pp collisions with 25 ns bunch spacing delivered by the LHC at $\sqrt{s} = 13$ TeV during Run 2. This corresponds to a total integrated luminosity of $L^{\text{Run 2}} = 140 \text{ fb}^{-1}$ (see Section 4.2.1). The uncertainty corresponding to this integrated luminosity is measured by the LUCID-2 detector to be 0.83% [179, 244]. This data-taking period ranges from 2015 to 2018 and, for each year, a different luminosity and uncertainty are measured, as is shown in Table 6.3. This data-taking period also presents different pile-up through the years. Figure 6.2 presents the average number of interactions in each crossing between proton bunches. As can be seen, the average pile-up increased every year.

Year	2015	2016	2017	2018
Peak $\mathcal{L}(t)$ ($\times 10^{33} \text{ cm}^2 \text{ s}^{-1}$)	5	13	16	19
Total delivered L (fb^{-1})	4.0	38.5	50.2	63.4
L registered by ATLAS (pb^{-1})	$3244.54 \pm 1.13\%$	$33402.2 \pm 0.89\%$	$44630.6 \pm 1.13\%$	$58791.6 \pm 1.10\%$
Periods	D–H,J	A–G,I,K,L	B–F,H,I,K	B–D,F,I,K,L,M,O,Q
Run numbers	276262–284484	297730–311481	325713–340453	348885–364292
Number of events	220.58M	1057.84M	1340.80M	1716.77M

Table 6.3: Peak luminosity, integrated luminosity delivered by the LHC and cumulative luminosity collected by the ATLAS detector at $\sqrt{s} = 13$ TeV during Run 2 per year [245]

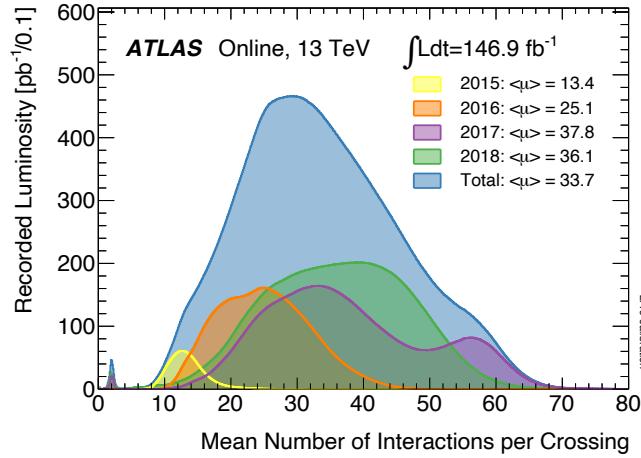


Figure 6.2: Luminosity-weighted distribution of the mean number of interactions per crossing $\langle\mu\rangle$ for full Run 2 and per year with pp collisions data during stable beams at $\sqrt{s} = 13$ TeV [179].

As introduced in Section 4.2.1, the GRLs is an xml file that collects the luminosity blocks that are considered good to be used for physics analyses. This is

done by demanding that the LHC had stable beams and that all the detectors and subdetectors were operating correctly. The GRL is used to filter the registered data at the luminosity blocks⁵ level. Events were selected from a shared data stream using the unprescaled single-lepton triggers, which are detailed in Section 6.3.1.

6.2.2 Simulated event samples

The simulated event samples used in this analysis were produced using different MC-event generators and simulation frameworks. The generalities of the ATLAS simulation chain are described in Section 4.3.

The MC production is divided into campaigns, where the centre-of-mass energy, geometry and conditions used in production correspond to a running period of the LHC. Three campaigns are used: mc16a, mc16d, and mc16e, corresponding to periods 2015-2016, 2017, and 2018, respectively. Various event generators interfaced with shower/hadronisation generators are used in this analysis.

- For the pile-up modelling (see Section 4.3.1.5), PYTHIA 8.186 is used with the ATLAS third set of tuned parameters for minimum-bias events (A3 tune [246]) and the NNPDF2.3LO set of PDFs [247].
- For the hard-scattering process, unless it is specified differently, the NNPDF3.0NLO PDF set [248] is used.
- For the PS, unless explicitly stated otherwise, the PYTHIA generator is used with A14 [249] tune and the NNPDF2.3LO PDF set.
- The top-quark decay is simulated at LO either by PYTHIA 8 or HERWIG 7 in the PS or by MADSPIN [186, 204] to preserve spin correlations, while the decay of the Higgs boson is generated either by PYTHIA 8 or HERWIG 7 in the PS. The decays of bottom and charm hadrons are simulated using the EVTGEN program (either version 1.6.0 or 1.7.0) [205].
- The MC-simulated events are weighted to match the distribution of the average number of interactions per bunch crossing, $\langle \mu \rangle$, observed in the real detector data. A rescaling factor of 1.03 ± 0.04 is applied to the $\langle \mu \rangle$ value of the samples to improve agreement between data and simulation [250].
- The FS event samples are used as the baseline whenever available. For certain cases such as the tHq signal process and the tWH and four top-quark

⁵A luminosity block corresponds to about 1 or 2 minutes of data taking and has around $\sim 10^5$ events. It is a unit of known luminosity.

background processes, AFII event samples are used as baseline. Most of the systematic effects are evaluated using alternative AFII samples, except for specific systematic uncertainties such as $t\bar{t}/tW$ interference, $t\bar{t}Z$, $t\bar{t}W$, and tWZ modelling, where FS event samples are available and, therefore, used. The details about AFII and FS simulation strategies are provided in Section 4.3.1.6

These MC event samples are used to assess efficiency and resolution models and to estimate systematic uncertainties. The details of each simulation event sample for each process are provided in the subsequent subsections. A summary of all tHq signal and background processes is presented in Table 6.4. The relevance of the processes listed in Table 6.4 is not uniform. In this section, it is also discussed the significance of each background process, highlighting their respective importance and how they mimic the final state of the signal process.

Process	Generator	Order (scheme)	PDF set	Parton shower	PDF set (tune)
Signal					
tHq	MADGRAPH5_AMC@NLO 2.6.2	NLO (4FS)	NNPDF3.0NLO nf4	PYTHIA 8.230	NNPDF2.3LO (A14 tune)
Backgrounds					
$t\bar{t}$	POWHEG BOX v2	NLO (5FS)	NNPDF3.0NLO	PYTHIA 8.230	NNPDF2.3LO (A14 tune)
$V+jets$	SHERPA 2.2.1	NLO+LO	NNPDF3.0NNLO	-	-
Diboson	SHERPA 2.2.1-2	NLO+LO	NNPDF3.0NNLO	-	-
Triboson	SHERPA 2.2.2	NLO+LO	NNPDF3.0NNLO	-	-
$t\bar{t}Z$	MADGRAPH5_AMC@NLO 2.3.3	NLO	NNPDF3.0NLO	PYTHIA 8.210	NNPDF2.3LO (A14 tune)
$t\bar{t}W$	SHERPA 2.2.10	NLO	NNPDF3.0NNLO	-	-
$t\bar{t}H$	POWHEG BOX v2	NLO (5FS)	NNPDF3.0NLO	PYTHIA 8.230	NNPDF2.3LO (A14 tune)
t -channel	POWHEG BOX v2	NLO (4FS)	NNPDF3.0NLO nf4	PYTHIA 8.230	NNPDF2.3LO (A14 tune)
tW	POWHEG BOX v2	NLO (5FS, DR)	NNPDF3.0NLO	PYTHIA 8.230	NNPDF2.3LO (A14 tune)
s -channel	POWHEG BOX v2	NLO	NNPDF3.0NLO	PYTHIA 8.230	NNPDF2.3LO (A14 tune)
tZq	MADGRAPH5_AMC@NLO 2.3.3	NLO	NNPDF3.0NLO	PYTHIA 8.230	NNPDF2.3LO (A14 tune)
tWH	MADGRAPH5_AMC@NLO 2.8.1	NLO (5FS, DR)	NNPDF3.0NLO	PYTHIA 8.245p3	NNPDF2.3LO (A14 tune)
tWZ	MADGRAPH5_AMC@NLO 2.3.3	NLO	NNPDF3.0NLO	PYTHIA 8.212	NNPDF2.3LO (A14 tune)
$t\bar{t}t$	MADGRAPH5_AMC@NLO 2.2.2	NLO	NNPDF3.1NLO	PYTHIA 8.186	NNPDF2.3LO (A14 tune)
$t\bar{t}\bar{t}\bar{t}$	MADGRAPH5_AMC@NLO 2.3.3	NLO	NNPDF3.1NLO	PYTHIA 8.230	NNPDF2.3LO (A14 tune)
ggH	POWHEG BOX v2	NLO	CT10	PYTHIA 8.210	CTEQ6L1 (AZNLO tune)
$q\bar{q}H$	POWHEG BOX v1	NLO	CT10	PYTHIA 8.186	CTEQ6L1 (AZNLO tune)
WH	PYTHIA 8.186	LO	NNPDF2.3LO	-	-
ZH	PYTHIA 8.186	LO	NNPDF2.3LO	-	-

Table 6.4: Summary of the baseline-nominal-simulated signal and background event samples used in the tHq analyses. The alternative simulations are described in Section 6.7.2.

6.2.2.1 Simulated tHq signal samples

The tHq simulated event sample is generated by using the MADGRAPH5_AMC@NLO 2.6.2 [251] generator at NLO, employing the NNPDF3.0NLO nf4 [248] PDF set. The μ_R and μ_F scales are set to a default scale based on $0.5 \times \sum_i \sqrt{m_i^2 + p_{T,i}^2}$, where i runs over all the particles

generated by the \mathcal{M} calculation. The simulation event samples are generated in the 4FS scheme (see Section 2.1.2.2 for the scheme discussion). This is referred to as the nominal signal sample.

The normalisation of the tHq signal samples is performed with respect to the cross-section predictions obtained from NLO generator. For pp collisions at $\sqrt{s} = \sqrt{13}\text{ TeV}$, the cross-section of the ML channels corresponds to $\sigma_{\text{NLO}}(tHq_{\text{ML}}) = 16.7\text{ pb}$ using a top-quark mass of $m_t = 172.5\text{ GeV}$. This cross-section is a factor 3.6 times smaller than the one for the $tHq(b\bar{b})$ production, with $\sigma_{\text{NLO}}(tHq_{b\bar{b}}) = 60.1\text{ fb}$.

6.2.2.2 Simulated background samples

The background can be defined as everything in a subset of the data that imitates the signal processes without truly being a signal event. Therefore, in this study, whatever mimics the signature of an associated tHq production with $2\ell + 1\tau_{\text{had}}$ final state is referred to as background.

To perform the physics analysis, it is fundamental to subtract the background events from the dataset as much as possible to achieve higher signal purity. By doing this, the analysed dataset resembles more to the process that is desired to study. This procedure is the so called “event selection” and it is described in Section 6.6.

In the subsequent section, the MC simulation of the background processes is presented, highlighting the main ones and their specific characteristics. Furthermore, an explanation is provided for each of these processes to clarify how they contribute to the background. Later on, in Section 6.5, a comprehensive description of the background estimation is discussed.

Top-quark pairs

The production of $t\bar{t}$ events constitutes the main background source for both $2\ell + 1\tau_{\text{had}}$ sub-channels. When the leptons from the top-quark decay are electrons or muons, the $t\bar{t}$ process can mimic the $2\ell + 1\tau_{\text{had}}$ signature if one of the quarks produces a jet that is wrongly reconstructed as a τ_{had} . This background particularly relevant in the $2\ell\text{ OS} + 1\tau_{\text{had}}$ channel, where it surpasses the signal yields by three orders of magnitude. As it is discussed in Section 6.5, the misidentification of quarks as if they were τ_{had} constitutes the main source of background. If one of the leptons in the $t\bar{t}$ diagram was a hadronically-decaying tau and the other an e/μ , the $2\ell + 1\tau_{\text{had}}$ signature could be obtained if one of the b -jets is wrongly identified

as an e/μ . This second scenario of misidentified e/μ is less common but can still happen.

Regarding its generation, the $t\bar{t}$ events are simulated using Powheg Box v2 [186, 252–254] and the NNPDF3.0NLO PDF set [248]. The μ_R and μ_F scales are set to the default scale of $\sqrt{m_t^2 + p_T^2}$. The parameter h_{damp} , controlling the matching between the Powheg generator and high- p_T radiation, is set to 1.5 the mass of the top quark [255]. All these samples are generated in the 5FS.

The $t\bar{t}$ sample is normalised to the cross-section prediction at NNLO in QCD including the resummation of NNLL soft-gluon terms calculated using TOP++ 2.0 [256–261]. For pp collisions at $\sqrt{s} = 13$ TeV, this cross-section corresponds to $\sigma_{\text{NNLO+NNLL}}(t\bar{t}) = 832 \pm 51$ fb using a top-quark mass of $m_t = 172.5$ GeV.

Single vector boson

This background corresponds to the $Z + \text{jets}$ and $W + \text{jets}$ productions, which are also referred to as $V + \text{jets}$. For $Z + \text{jets}$, it can mimic the $2\ell \text{OS} + 1\tau_{\text{had}}$ signature when one of the b -quark from the gluon splitting is misidentified as a τ_{had} . Conforming, along with $t\bar{t}$ the main background in that particular channel.

The single-vector-boson processes are simulated using the SHERPA 2.2.1 generator [262]. The \mathcal{M} accuracy is NLO for up to two partons and LO for up to four partons. The PS is also simulated with SHERPA, which is based on the Catani–Seymour dipole factorisation and the cluster hadronisation model [263]. The \mathcal{M} for a given jet multiplicity are matched to the PS using a colour-exact variant of the MC@NLO algorithm [264]. The $V + \text{jets}$ samples are normalised to a NNLO prediction [265].

Top-quark pair in association with a single vector boson

Right after the single-boson and $t\bar{t}$ processes, the $t\bar{t}V$ processes (i.e. $t\bar{t}Z$ and $t\bar{t}W$) are the most relevant background, particularly in the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel.

There are various ways in which these processes can mimic the final state of the signal. Since it can have three leptons in the final state, the object multiplicity can be replicated and $t\bar{t}W$ can be an irreducible background. It can also mimic the $2\ell + 1\tau_{\text{had}}$ signature if misidentification takes place.

The $t\bar{t}W$ events are simulated using the SHERPA 2.2.10 generator at NLO accuracy in QCD with the NNPDF3.0NNLO PDF set [248]. This framework incorporates a strategy that merges multiple legs, allowing for the integration of one extra parton at NLO and two at LO. An event-by-event correction factor was applied to

this sample that provide virtual NLO EWK corrections. Additionally, a simulation sample at LO accuracy in QCD was also generated with SHERPA 2.2.10 but for the $t\bar{t}W + j$ final state to emulate EW corrections to $t\bar{t}W$ production. The generation of $t\bar{t}Z$ events event is conducted with the MADGRAPH5_AMC@NLO 2.3.3 [251] generator, providing \mathcal{M} at NLO in α_s and using the NNPDF3.0NLO PDF set. The μ_F and μ_R scales are set to the default of $0.5 \times \sum_i \sqrt{m_i^2 + p_{T,i}^2}$, where the sum runs over all the particles generated from the \mathcal{M} calculation.

For $t\bar{t}W$, the cross-sections are calculated at NLO QCD and NLO EW accuracy using the MADGRAPH5_AMC@NLO FxFx framework. The predicted value is $\sigma_{\text{NLO QCD} + \text{NLO EW}}^{\text{pred}}(t\bar{t}Z) = 722.48^{+70.2}_{-77.7}$ fb. For $t\bar{t}Z$, the cross-sections are calculated at NLO QCD and NLO EW accuracy using MADGRAPH5_AMC@NLO as reported in Reference [83]. The obtained cross-section at $\sqrt{s} = 13$ TeV is $\sigma_{\text{NLO QCD} + \text{NLO EW}}^{\text{pred}}(t\bar{t}Z) = 0.88^{+0.09}_{-0.11}$ pb.

Top-quark pair in association with a Higgs-boson

The $t\bar{t}H$ process is discussed in depth in Section 2.3.2. If the Higgs boson and one of the top quarks decay to leptons and the other top quark to $q\bar{q}$, the multiplicity in leptons of the $2\ell + 1\tau_{\text{had}}$ channel is reproduced and the $t\bar{t}H$ is reconstructed as a tHq process. The mimicking of the signal process can also take place by other decay modes of the $t\bar{t}H$ if a quark-initiated jet is misidentified as a τ_{had} . This background is particularly important in the $2\ell \text{ SS} + 1\tau_{\text{had}}$ channel.

To simulate the $t\bar{t}H$ processes, the POWHEG BOX v2 generator with the NNPDF3.0NLO PDF set is used. The functional form of the μ_R and μ_F scales was set to $\sqrt[3]{m_T(t) \cdot m_T(\bar{t}) \cdot m_T(H)}$, where m_T is the transverse mass.

At $\sqrt{s} = 13$ TeV, the cross-section is computed to NLO QCD and NLO EW precision with MADGRAPH5_AMC@NLO [83] and it is $\sigma_{\text{NLO QCD} + \text{NLO EW}}^{\text{pred}}(t\bar{t}H) = 507^{+35}_{-50}$ fb.

Diboson

The diboson background is another one with significant contribution. In the case of the WZ diboson process, there can be three leptons in the final state and, hence, the object multiplicity of the $2\ell + 1\tau_{\text{had}}$ channel can be replicated without any type of misidentification. For the ZZ diboson process there can be either two or four leptons. In the former situation, misidentification is necessary to mimic the $2\ell + 1\tau_{\text{had}}$ signature. In the latter, one of the leptons should not be reconstructed to reproduce the $2\ell + 1\tau_{\text{had}}$ final-state-object multiplicity.

Samples of diboson final states (VV) are simulated using the SHERPA 2.2.1 or 2.2.2 generator depending on the process. The samples use \mathcal{M} s at NLO accuracy in QCD. The NNPDF3.0NNLO set of PDFs is used, along with the dedicated set of tuned PS parameters developed by the SHERPA authors.

Single-top-quark processes

The three single-top-quark productions described in Section 2.1.2.2 are simulated but only the tW production plays a relevant role. In the 2ℓ OS + $1\tau_{\text{had}}$ channel, tW stands out as a significant background, but its impact is not as significant as $t\bar{t}$ and $Z + \text{jets}$. All three single-top-quark processes are modelled using POWHEG BOX v2 generator in conjunction with PYTHIA 8. The DR scheme is employed to handle the interference between tW and $t\bar{t}$.

Single-top quark in association with a boson

Another family of background processes is the tX , which is composed by tZq , tWZ and tWH . The tWH process exhibits heightened sensitivity to deviations of y_t from the SM expectation and a more inclusive tH search has to include this process as part of the signal (this is not the case for this thesis). The tZq production presents a challenge due to its close resemblance to the tHq signal. When stringent selection requirements are applied to discriminate the signal from all the background processes, the reduction of tZq is considerably more modest compared to other processes.

The tWH events are generated at NLO in the 5FS using MADGRAPH5_AMC@NLO 2.8.1 with the NNPDF3.0NLO. The events are interfaced with PYTHIA 8.245p3 using the A14 tune and the NNPDF2.3LO PDF set. The functional form of the μ_F and μ_F scales is set to the default scale $0.5 \times \sum_i \sqrt{m_i^2 + p_{T,i}^2}$, where the sum runs over all the particles generated from the \mathcal{M} calculation. The DR scheme is employed to handle the interference between tWH and $t\bar{t}H$.

The tZq sample is simulated at NLO in the 4FS using MADGRAPH5_AMC@NLO 2.3.3 with the NNPDF3.0NLO PDF set and then interfaced with PYTHIA 8.230 using the A14 tune and the NNPDF2.3LO PDF set. The functional form of the μ_F and μ_F scales is set to $4\sqrt{m_b^2 + p_{T,b}^2}$, where m_b is the mass of the m_b -type quark [266].

The tWZ process is simulated at NLO in the 5FS using MADGRAPH5_AMC@NLO 2.3.3. The top quark and the Z boson are decayed at LO using MADSPIN. Here, the decay of the Z boson is restricted to a pair of

charged leptons. The μ_F and μ_R scales are set to the top-quark mass. To address the interference between tWZ and $t\bar{t}Z$, the DR scheme is employed.

Other non-dominant processes

Other minor backgrounds considered in this analysis are the triboson (VVV), the Higgs-boson processes (ggF , VBF , VH), three top quark (ttt) and four top quarks ($t\bar{t}t\bar{t}$). The simulation procedure of the baseline samples for all these processes is summarised in Table 6.4.

6.3 Object definition

In Chapter 5, a general overview of the object reconstruction process is provided. Different possible definitions for the physical objects are discussed in that chapter. In contrast, this section focuses on presenting the specific definitions employed for this analysis at hand.

6.3.1 Triggers

To select events from collisions, a two-level-trigger-system is employed [267] as described in Section 3.3.6. During the data-taking period spanning 2015–2018, varying pile-up conditions necessitated the use of different single-lepton (electron or muon) triggers. The trigger names employed in this analysis are listed in Table 6.5. These triggers are combined using the logical OR operation.

Year	Single-electron trigger	Single-muon trigger
2015	HLT_e24_lhmedium_L1EM20VH	HLT_mu20_iloose_L1MU15
	HLT_e60_lhmedium	HLT_mu50
	HLT_e120_lhloose	
2016–2018	HLT_e26_lhtight_nod0_ivarloose	HLT_mu26_ivarmedium
	HLT_e60_lhmedium_nod0	HLT_mu50
	HLT_e140_lhloose_nod0	

Table 6.5: Employed single-lepton trigger depending on the light-lepton flavour and year.

- The single-electron triggers rely on the likelihood-based algorithm for prompt-electron identification and non-prompt electron rejection described in Sec-

tion 6.3.2. Isolation and identification requirements are applied as well. Besides, various lower transverse energy (E_T) thresholds ranging from 20 GeV to 26 GeV are applied for low- p_T electrons. For high- p_T electrons, two additional complementary triggers with lower- E_T thresholds from 60 GeV to 140 GeV are employed.

- Muons in this analysis are triggered by matching tracks reconstructed in both the MS and ID, along with isolation criteria. Similar to the electron triggers, to mitigate efficiency losses due to isolation at high p_T , several lower- p_T thresholds from 20 GeV to 50 GeV are set.

6.3.2 Electrons

In this analysis, the electron candidates must meet several criteria: they should have a p_T greater than 10 GeV, be within the pseudorapidity range of $|\eta^{\text{clust}}| < 2.47$, and pass the **tight** identification level. Electron candidates are excluded if their calorimeter clusters lie within the transition region between the barrel and the endcap sections of the EM calorimeter, defined as $1.37 < |\eta^{\text{clust}}| < 1.52$. Additionally, these electrons must pass the tight isolation test based on the “Prompt Lepton Improved Veto” (PLImprovedTight) isolation working point [268]. This isolation working point uses a multivariate analysis technique, combining shower shape and track information, to effectively differentiate between prompt electrons and those that arise from backgrounds such as hadronic jets, photon conversions, or the decay of heavy-flavour hadrons. Additionally, the associated track must satisfy $|z_0 \sin(\theta)| < 0.5$ mm and $|d_0|/\sigma(d_0) < 5$.

In certain analysis regions, additional requirements are imposed on electrons. These include the application of the **Electron Charge ID Selector Tool** (ECIDS), a BDT-based tool which enhances the rejection of electrons with misidentified-electrical charges. Another tool referred to as **Electron Ambiguity Tool** is used to suppress the misidentification of photons reconstructed as electrons.

6.3.3 Muons

The preselected-muon candidates used for the overlap-removal process must satisfy $p_T > 7$ GeV, $|\eta| < 2.5$, and pass the **medium** identification criteria. This working points imposes conditions on the number of hits in the ID and MS subsystems, as well as on the significance of the charge-to-momentum ratio [215, 216].

Isolation criteria are established using the level of isolation determined by a multivariate likelihood algorithm. These requirements (**tight**) are applied to suppress contributions from misidentified or non-prompt muons [268].

Moreover, the track associated with the muon candidates must satisfy $|z_0 \sin(\theta)| < 0.5$ mm and $|d_0|/\sigma(d_0) < 3$, as mentioned in Section 5.3.

A summary of the specific-object-definition criteria for electrons and muons is presented in Table 6.6.

	Electron	Muon
Identification	tight	medium
Isolation	tight	tight
Acceptance	$p_T > 10 \text{ GeV}$, $ \eta^{\text{clust}} < 2.47$ except $1.37 < \eta^{\text{clust}} < 1.52$	$p_T > 10 \text{ GeV}$, $ \eta < 2.5$
Impact parameter	$ d_0/\sigma(d_0) < 5.0$ $ z_0 \sin(\theta) < 0.5$ mm	$ d_0/\sigma(d_0) < 3.0$ $ z_0 \sin(\theta) < 0.5$ mm
Extra selections	ECIDS, e/γ ambiguity-cuts	
Overlap removal		See Section 6.3.6

Table 6.6: Summary of the electron- and muon-object definitions used in this analysis.

6.3.4 Hadronically-decaying taus

The selection criteria for τ_{had} candidates are outlined in Table 6.7 and adhere to the guidelines established by the Tau CP group [225, 269]. The τ_{had} objects must satisfy the criteria of $p_T > 20 \text{ GeV}$, $|\eta| < 2.5$, excluding the range $1.37 < |\eta^{\text{clust}}| < 1.52$, and have either 1 or 3 associated tracks. This requirement is applied because, as mentioned in Section 5.2.3, the τ -leptons decay to one or three prongs. To distinguish τ_{had} candidates from other objects, they must pass the **medium** (**loose**) JetID requirement, which is defined using a RNN, as well as fail the cut on the electron BDT. No specific veto is applied for muons. The energy calibration is performed using the MVA Tau Energy Scale (MVA TES) method. Scale factors for leptons are applied to account for efficiency and energy scale corrections.

6.3.5 Jets

Jets are reconstructed using the anti- k_t jet algorithm [221] on particle-flow objects [270], with a distance parameter of $R = 0.4$, implemented in the FASTJET

	τ_{had}
Acceptance	$p_T > 20 \text{ GeV}$, $ \eta^{\text{clust}} < 2.5$ except $1.37 < \eta^{\text{clust}} < 1.52$
Number of tracks	1 or 3
Identification	RNN Medium
Electron veto	electron BDT Loose
Overlap removal	See Section 6.3.6

Table 6.7: Summary of the τ_{had} object definitions used in this analysis. The selection requirements for actual τ_{had} are applied in addition to the pre-selected objects used for the overlap removal.

package [271] (referred to as `AntiKt4EMPflowJets` jet collection). The Jet Vertex Tagger (JVT) [272, 273], recommended by the Jet/ E_T^{miss} CP group, is used to select jets. Jets are retained if they have $p_T > 20 \text{ GeV}$ and fall within the pseudorapidity range of $|\eta| < 4.5$. Additionally, jets with $p_T < 60 \text{ GeV}$ and $|\eta| < 2.4$ must satisfy $\text{JVT} > 0.5$ to meet the criteria of the `Tight` JVT working point. For forward jets with $2.5 < |\eta| < 4.5$ and $p_T < 120 \text{ GeV}$, an alternative JVT working point (`fJVT`) is applied, requiring $\text{fJVT} < 0.4$ along with a timing requirement on the jet [272]. The jet definition and b -tagging requirements are summarised in Table 6.8.

The jet calibration procedure follows the standard method recommended by the Jet/ E_T^{miss} CP group, which corrects the jet energy to match, on average, the true jet energy at the particle level and applies in-situ corrections for data [274].

6.3.5.1 Identification of b -tagged jets

In this study, the `DL1r` algorithm, a MVA method for b -tagging, is employed [234, 275, 276] (see Section 5.3.1). The jets are b -tagged if the values of the `DL1r` discriminant exceed certain thresholds or working points. Four working points are defined for the `DL1r` tagger, corresponding to selecting 85%, 77%, 70%, and 60% of b -jets in $t\bar{t}$ simulated events. To assess the b -tagging performance comprehensively, the efficiency of the `DL1r` tagger is measured using collision data. The `DL1r` tagger discriminant is defined in Equation 5.1.

6.3.6 Overlap removal

Once all the objects are identified, to avoid that a single signature is identified as different physical objects, the overlap removal is applied, as Section 5.5 introduced.

Jets	
Collection	<code>AntiKt4EMPflowJets</code>
Acceptance	$p_T > 20 \text{ GeV}$, $ \eta < 4.5$
Jet Vertex Tagger	JVT > 0.5 if $ \eta < 2.4$ and $p_T < 60 \text{ GeV}$ fJVT < 0.4 if $2.5 < \eta < 4.5$ and $p_T < 120 \text{ GeV}$
Overlap removal	See Section 6.3.6
<i>b</i> -tagging jet	
Acceptance	$p_T > 20 \text{ GeV}$, $ \eta < 2.5$
<i>b</i> -tagging	<code>DL1r</code> algorithm

Table 6.8: Summary of the jet selection criteria and *b*-tagging.

To avoid the double-counting, in this analysis, the pre-selected **Loose** leptons and jets are used. Then, the following steps are applied to resolve the ambiguities:

1. Any electron found with a track overlapping with any other electron is removed. The electron is also removed if it shares a track with a muon (except if it is a calorimeter muon) since the electron object is very likely to be identified from the tracks that the muon produced in the ID.
2. Any calorimeter muon (see Section 5.2.2) found to share a track with an electron is removed. This measure is taken because calorimeter muons objects are reconstructed in the region where the MS is not fully instrumented.
3. Any jet found within a $\Delta R \leq 0.2$ of an electron is removed, as it is very possible that the jet corresponds to the electron.
4. Any electron subsequently found within $\Delta R \leq 0.4$ of a jet is removed in order to reduce the impact of non-prompt electrons.
5. Any jet with fewer than 3 tracks associated to it found within $\Delta R \leq 0.2$ of a muon is removed. This is done to reduce the number of fake jets from muons depositing energy in the calorimeters.
6. Any muon subsequently found within $\Delta R \leq 0.4$ of a jet is removed to reduce the contribution from muons from heavy-flavour decays within a jet.
7. Any τ_{had} found within a $\Delta R \leq 0.2$ of an electron is removed.
8. Any τ_{had} found within a $\Delta R \leq 0.2$ of any type of muon is removed. If the tau has $p_T > 50 \text{ GeV}$, it will only be removed if it is found to overlap with a combined-type muon.

9. Any jet found within a $\Delta R \leq 0.2$ of a τ_{had} is removed.

The overlap removal procedure is implemented by applying the criteria in the order specified.

6.3.7 Missing transverse momentum

As discussed in Section 5.4, the \vec{E}_T^{miss} is reconstructed by summing the negative vector of p_T of reconstructed and calibrated particles and jets after performing the overlap removal. Additionally, a soft term is included, which consists of charged-particle tracks associated with the hard scatter vertex [277, 278]. The purpose of the soft term is to account for low-momentum particles that may not be identified among the final state objects [279–281]. The E_T^{miss} serves as a measurement of the undetectable particles in an event and is subject to energy losses caused by detector inefficiencies, acceptance limitations, and energy resolution. In this analysis, the main source of E_T^{miss} are the neutrinos in the final state.

6.4 Study of the tHq signal process

In this section, a study on some of the signal properties is presented. First of all the generation of the truth-level information of the tHq processes is briefly presented in Section 6.4.1. Later, the method to determine the parent particle for the light leptons in the $2\ell\text{SS} + 1\tau_{\text{had}}$ channel is presented in Section 6.4.2. Finally, the reconstruction of the kinematic properties of the Higgs boson and the top quark is discussed in Section 6.4.3. It is interesting to note how each of these studies needs the previous one, i.e. the reconstruction of the signal process needs the leptons assignment, and the lepton association uses the truth-level information.

6.4.1 Validation of parton-level simulations

As already presented in Section 4.3.1, the parton-level information refers to the MC-generated events before taking into account the effects of the interaction of the particles with the matter of the detector. It may also include the PS and hadronisation information for a given process. If the parton-level-simulation information is kept in the reconstructed MC event sample and can be matched to reconstructed objects.

In the ATLAS top-quark-physics group, a dedicated software package is used analyse, administrate and store the parton-level information. This package is referred to as `TopPartons` (see Appendix A.1). The kinematic information and the true identity⁶ of each of the particles in the event is saved in the NTuples through this library. To confirm the correct performance of the whole software, theoretical calculations have to be carried and compared to the output of the program. In Appendix A the details of this task are discussed in detail. Note that this development is not done only for the $2\ell + 1\tau_{\text{had}}$ channel but for all ML channels.

From the calculations in Appendix A.2, it is deducted that from all tHq events considered in the ML searches, only a 3.72% decay into a $2\ell + 1\tau_{\text{had}}$ final state. From these, as can be seen in Table 6.9, more than in 80% of cases the τ_{had} is produced in the Higgs-boson-decay chain.

Higgs-boson decay channel	Origin of τ_{had}		Total
	Top quark	Higgs boson	
$H \rightarrow \tau\tau$	5.06	64.06	69.13
$H \rightarrow WW^*$	9.01	18.01	27.02
$H \rightarrow ZZ^*$	2.22	1.64	3.85
Total	16.29	83.71	100.00

Table 6.9: Contribution as a percentage of each Higgs-boson decay channel to the $2\ell + 1\tau_{\text{had}}$ final state obtained from calculations combining the BRs of the considered decays. Here is also presented whether the τ_{had} is generated from the top quark or the Higgs-boson decay chain. The discrepancies with the first row of Table 6.2 are due to the reconstruction efficiency and selection requirements.

Having a proper parton-level information is fundamental for this analysis because it is used in several different tasks: from the determination of the light-lepton origin in the $2\ell \text{ SS} + 1\tau_{\text{had}}$ channel to the estimation of the τ_{had} -misidentification contribution.

6.4.1.1 Comparison between software and calculation results

With all decay scenarios incorporated into the `TopPartons` code, and having calculated from the expected production fractions, the last step to validate the parton-level simulations is to compare these two elements and ensure that they are consistent. The metric used to perform the comparison is the ratio between the

⁶Identity refers to which particle it is. The identity is commonly referred to as PDG-ID, which is the particle numbering scheme defined by the Particle Data Group. It assigns a unique code to each type of particle. For instance, PDG-ID(H boson)= 25 and PDG-ID(t quark)= 6.

tHq ($2\ell + 1\tau_{\text{had}}$) event yields in a particular Higgs-boson-decay channel and all the events in that particular decay mode:

$$\frac{\text{Number of events}(H \rightarrow \text{Decay channel} \rightarrow 2\ell + 1\tau_{\text{had}})}{\text{Number of events}(H \rightarrow \text{Decay channel})}.$$

In Table 6.10, the BR-based calculations performed in Appendix A.2 are put alongside the results of the parton-level truth informations described in Appendix A.1. As can be seen, the agreement between the calculations and the `TopPartons` output is resonable for the two main Higgs-boson-decay channels. For these, the disagreement is of 1.2% for the $\tau\tau$ and 5.9% for the WW . In contrast, a discrepancy of 60% is found in the ZZ decay mode. The origin of the variance in the ZZ decay is not due to bungs in the code that generates the parton-level information. Instead, it is originated from the omission of certain configurations in the ZZ decay mode within the evaluation tool. Even though the conflict in ZZ is much larger than for the other channels, this is not so problematic since the $H \rightarrow ZZ^*$ accounts for a very small part of the total tHq events in the $2\ell + 1\tau_{\text{had}}$ final-state channel (3.86%).

Yields ratio	Calculation	<code>TopPartons</code> result
$\frac{H \rightarrow \tau\tau \rightarrow 2\ell + 1\tau_{\text{had}}}{H \rightarrow \tau\tau}$	0.1246	0.1232 ± 0.0057
$\frac{H \rightarrow WW^* \rightarrow 2\ell + 1\tau_{\text{had}}}{H \rightarrow WW^*}$	0.0141	0.0151 ± 0.0009
$\frac{H \rightarrow ZZ^* \rightarrow 2\ell + 1\tau_{\text{had}}}{H \rightarrow ZZ^*}$	0.0164	0.0100 ± 0.002

Table 6.10: Theoretical predictions compared to the `TopPartons` output. The uncertainty on the second column corresponds to the statistical uncertainty.

6.4.2 Light-lepton-origin assignment

The two light leptons in the final state of the $2\ell + 1\tau_{\text{had}}$ channel can originate either from the Higgs boson or the top quark. The ambiguities regarding the origin of these light-flavoured leptons make the reconstruction of the top-quark and Higgs-boson systems extremely difficult. Nevertheless, the electric charge of these leptons could provide useful information to probe their origins.

To know whether the light-flavoured leptons in the final state are originated from the Higgs boson or the top quark is very beneficial to both reconstruct the event and design variables at reconstruction level with high discrimination power. The variables using the lepton assignment information play a relevant role not only

in the definition of the signal-enriched section but also the in the determination of the control regions (see Section 6.6.4) to constrain the most important background processes.

According to the calculations performed by combining the BR of the Higgs boson, the top quark and all its decay products (see Section 6.4.1), in the $2\ell + 1\tau_{\text{had}}$ channel of tHq production, the τ_{had} is produced 83.7% of times as a product of the Higgs-boson decay in opposition to the 16% in which it comes from the top-quark disintegration.

Origin association for $2\ell \text{ OS} + 1\tau_{\text{had}}$

In the dominant scenario (τ_{had} originated from the Higgs-boson system) the association of which light-flavoured lepton comes from the top-quark decay and which one comes from the Higgs-boson decay can be done directly if these two leptons have opposite electric charges, i.e. in the $2\ell \text{ OS} + 1\tau_{\text{had}}$ channel. Since the Higgs boson is neutrally charged, the sum of the charge of its decay products should be zero. Therefore, in the $2\ell \text{ OS} + 1\tau_{\text{had}}$ channel, while the light lepton with opposite charge to that of the τ_{had} is the one coming from the Higgs boson, the other lepton, i.e. the one with the same charge as τ_{had} , is the one originated from the top-quark decay.

Origin association for $2\ell \text{ SS} + 1\tau_{\text{had}}$

In contrast to the $2\ell \text{ OS} + 1\tau_{\text{had}}$ channel, in the case of τ_{had} coming from the Higgs boson, when the two light leptons have the same electric charge, $2\ell \text{ SS} + 1\tau_{\text{had}}$, it is not possible to know, a priori, which of the leptons comes from the top-quark system and which from the Higgs-boson decay.

To perform this association for the $2\ell \text{ SS} + 1\tau_{\text{had}}$ three methods are tested. These are:

- **Initial approach:** A NN based on the Keras framework was trained to perform the assignment task [282]. In labelling the data, it was assumed that the leading lepton (ℓ_1) always originates from the top quark. Truth-level studies have shown that in most cases, the lepton coming from the decay chain of the top quark is the leading lepton. This is because the top quark typically carries more momentum than the Higgs boson. However, it should be noted that this assumption is only correct 61.1% of the time. Therefore, this unreliable labelling of data produces an unreliable NN and, hence, this approach is discarded. The performance of this method is evaluated in Section 6.4.2.7, where it is compared to the baseline technique for leptons association.

- **Cut-based classification:** The most simple method to carry out the lepton assignment involves employing variables capable of distinguishing the origin of the lepton in the 2ℓ OS + $1\tau_{\text{had}}$ scenario, where the origin is known. Then, some criteria are applied to these variables to define an algorithm to assign the lepton origin. The visible Higgs-boson mass (m_H^{vis}) and the reconstructed top-quark mass (m_t^{reco}) are used for this purpose and the logic is the following:

If $\Delta(m_H^{\text{vis}}) > 57$ GeV: Assign lepton to top quark for which $m_H^{\text{vis}}(\ell^{\text{top}}) > m_H^{\text{vis}}(\ell^H)$.

If $\Delta(m_H^{\text{vis}}) < 57$ GeV: Assign lepton to top quark for which $m_t^{\text{reco}}(\ell^{\text{top}}) > m_t^{\text{reco}}(\ell^H)$,

where $\Delta(m_H^{\text{vis}}) = m_H^{\text{vis}}(\ell_1) - m_H^{\text{vis}}(\ell_2)$. This algorithm provides an accuracy of about 80% when evaluated on the 2ℓ OS + $1\tau_{\text{had}}$ sample exclusively.

- **BDT-based method:** To accurately assign the origin of the light leptons in the 2ℓ SS + $1\tau_{\text{had}}$ scenario, a gradient BDT method was developed. The BDT is implemented using the Toolkit for Multivariate Data Analysis (TMVA) library of ROOT [283, 284], whose technicalities are discussed in Appendix B.

This BDT-based method uses labels derived from the truth-level information and is trained using reconstruction-level variables. Subsequently, it can later predict the lepton origin for unlabelled data. The methodology employed in this approach is thoroughly described in this section, covering the creation of the labels, the training process, and the application of the BDT model. The results of this technique are presented in Section 6.4.2.7.

Among the three developed methods, the BDT-based approach provides the best results as it is presented in Section 6.4.2.7. The implementation of the BDT-based approach can be outlined through the following procedural steps:

1. **Labelling:** Creation of a label for supervised training through the use of truth-level information and the establishment of categories for classification. The categories “Type 1” and “Type 2” are defined.
2. **Feature selection:** Selection of reconstruction-level input features with a discriminatory capacity between Type 1 and Type 2.
3. **Hyperparameter optimisation:** Optimisation of the training hyperparameters.
4. **Negative-weight usage:** The choice of the negative-weights-treatment strategy. This matter is discussed in Appendix D.
5. **Training:** Supervised training of the model to classify events according to the origin of the light lepton.

6. **Evaluation:** Application of scores and search of the optimal classification threshold ($\text{BDT}_{\text{Threshold}}^{\text{Lepton Assignment}}$).

All these steps are described in the following sections.

6.4.2.1 Labelling the $2\ell \text{SS} + 1\tau_{\text{had}}$ with the reconstruction-level and truth-level matching

Even though at reconstruction level it is not known which are the parents of the particles in the final state, at parton level this information is accessible, in other words, the origin⁷ of the light leptons is known. For a given simulated event, it is possible to access to both the reconstruction-level and parton-level information simultaneously. Having the parton-level leptons, whose parents are known, and the reconstruction-level leptons, whose parents need to be identified, it is possible to compare them to create an association and, therefore, identify which parton-level lepton corresponds to which reconstructed lepton. The aim of this relation is to assign the leading (ℓ_1) and sub-leading (ℓ_2) light leptons at reconstruction level to the “lepton from top-quark-decay chain” (ℓ_{top}) and “lepton from Higgs-boson decay chain” (ℓ_{Higgs}) at truth level.

In order to link the reconstruction-level light leptons to the parton-level light leptons, a $\Delta R < 0.01$ cone around each of the reconstructed leptons is built. Observe in Figure 6.3 that, with a cone of that size, almost all reconstruction-level leptons find a parton-level match. When inside that cone there is exactly one truth-level light lepton, there is what is called “a match”. To identify properly the lepton origin in an event, it is required that both leptons at reconstruction level have a match. There are two different cases for this. The first situation is that in which the leading-light lepton is ℓ_{top} and the sub-leading is ℓ_{Higgs} . For the sake of simplicity, this configuration is named “Type 1” and it is represented in Figure 6.4a. The second double-matching combination is the other way around, the leading-light lepton is ℓ_{Higgs} and the sub-leading is ℓ_{top} . Pictured in Figure 6.4b, this type of events are called “Type 2”. On the contrary, if only one of the two reconstructed light leptons is matched (Figure 6.4c), none of the leptons are classified. Finally, in the scenario in which none of the parton-level leptons fall into the cones (Figure 6.4e), no assignation takes place.

To perform this labelling, it is required that the τ_{had} is originated in from the Higgs-boson system. This is imposed in order to guarantee that there are both a ℓ_{top} and a ℓ_{Higgs} . This condition is satisfied to more than 80% of events. The

⁷By origin of a light lepton is meant whether this particle comes from the Higgs-boson-decay chain or the top-quark-decay chain.

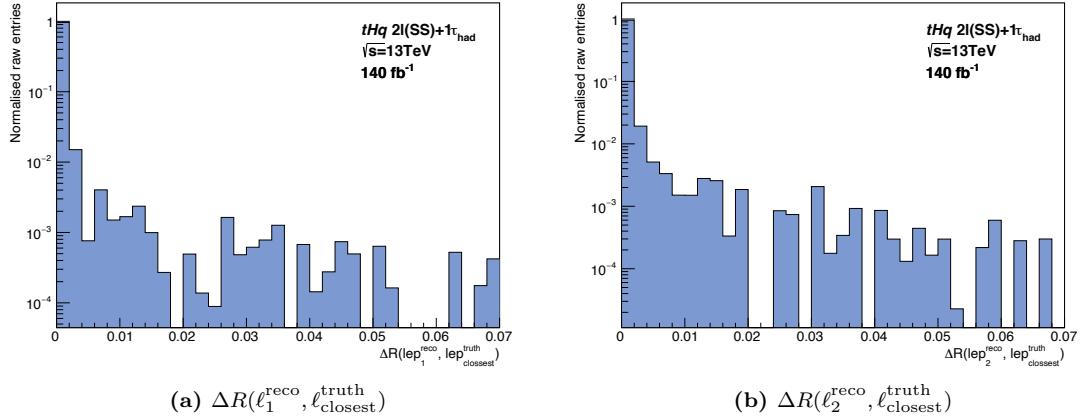


Figure 6.3: Normalised distribution of the ΔR distance between the reconstructed light leptons and the closest parton-level lepton in the 2ℓ SS + $1\tau_{\text{had}}$ channel. The events are weighted using the misidentification scale factors. Note how, by demanding $\Delta R < 0.01$, the leptons find a match in the majority of situations.

Higgs-boson decay channels used for these studies are the $H \rightarrow \tau\tau$ (one τ decaying leptonically and the other hadronically) and the $H \rightarrow WW^*$. The $H \rightarrow ZZ^*$ channel has not been included since its impact in the on the $2\ell + 1\tau_{\text{had}}$ production when the τ_{had} comes from the Higgs boson is negligible. If the τ_{had} is originated in the Higgs-boson system, only 2.0% of the events correspond to the $H \rightarrow ZZ^*$ decay channel, in contrast to the 76.5% of the $H \rightarrow \tau\tau$ and the 21.5% of the $H \rightarrow WW^*$. These numbers are presented in Table 6.11. Once all these conditions are applied, a minimum distance between each of the reconstructed leptons to its correspondent parton-level lepton, $\Delta R_{\min}^{\ell_1, \ell_2}$ is demanded. The corresponding entry counts at each step are summarised in Table 6.12. Note that the numbers are not the event yields but the entries, i.e. counting the raw number of MC events with no weights into consideration. More information about event weights is given in Appendix D.

Channel	Fraction (%)
$H \rightarrow \tau\tau$	76.52
$H \rightarrow WW^*$	21.52
$H \rightarrow ZZ^*$	1.956

Table 6.11: Contribution of each Higgs-boson decay channel to the $2\ell + 1\tau_{\text{had}}$ final state when demanding that the τ_{had} is originated from the Higgs-boson decay chain. The numbers in this table are calculated from the rightmost column of Table 6.9.

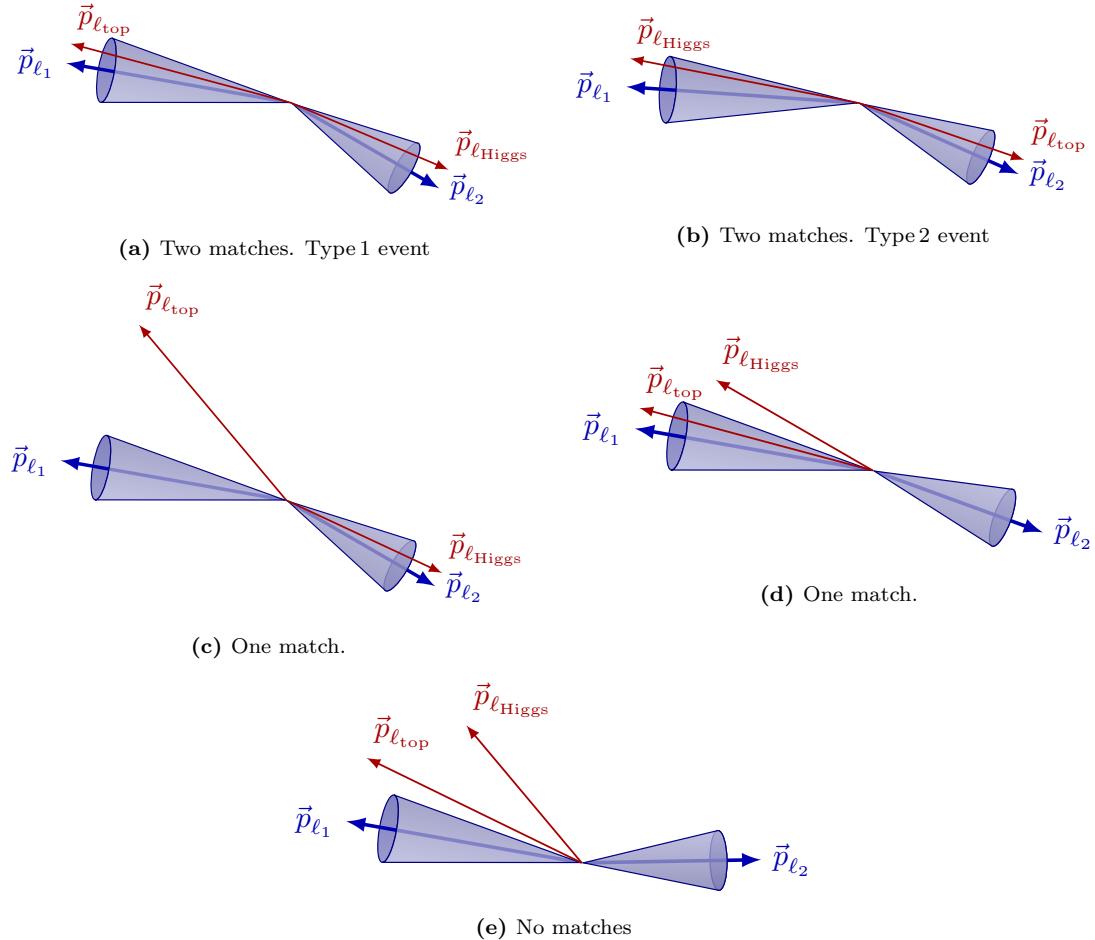


Figure 6.4: Different scenarios for the association between reconstruction-level (blue arrow) and parton-level (red arrow) light leptons. Note that the labels ℓ_{top} and ℓ_{Higgs} are only available for the parton-level particles. The labelling of the events is performed only for the cases in (a) and (b).

6.4.2.2 BDT input features

The choice of input variables for training a BDT is a crucial factor for achieving good classification accuracy. The chosen variables must exhibit the ability to effectively differentiate between Type 1 and Type 2. Nine variables are used. Figure 6.5 presents the distributions of the highest ranked. Appendix C.1 shows the distributions of the other five. In these plots can be seen the Type 1 and Type 2 present different profiles for the same variable. This divergence in shapes indicates the efficiency of these variables for classifying.

Stage	Entries
Total tHq ($2\ell + 1\tau_{\text{had}}$) event sample	43922
$2\ell \text{SS} + 1\tau_{\text{had}}$ event sample	18140
Selection: $H \rightarrow \tau_{\text{lep}}\tau_{\text{had}}/W^+W^-$, $W \rightarrow e/\mu/\tau_{\text{lep}}$	15446
$\Delta R(\ell_1^{\text{reco}}, \ell_{\text{closest}}^{\text{truth}}) < 0.01$ and $\Delta R(\ell_2^{\text{reco}}, \ell_{\text{closest}}^{\text{truth}}) < 0.01$	14680

Table 6.12: Unweighted (raw) events at each step of the labelling process. The first row corresponds to the entire simulated event sample of just tHq events in the $2\ell + 1\tau_{\text{had}}$ channel after the preselection requirements (see Section 6.6.1). The second row applies the condition on the charge of the light leptons. The third refers to the subset of $2\ell \text{SS} + 1\tau_{\text{had}}$ events for which it is computed the ΔR distance between the truth- and reconstruction-level leptons. The $W \rightarrow e/\mu/\tau_{\text{lep}}$ condition enforces that the τ_{had} is produced in the Higgs-boson system. Finally, the last row demands that both leptons at reconstruction level are matched to different truth-level leptons.

The TMVA package is capable of ranking variables within its model based on how often the variables are used to split decision tree nodes, and by weighting each split occurrence by the separation gain-squared⁸ ($< S^2 >$) it has achieved and by the number of events in the node [285]. The variables employed in the model, ordered according to their respective levels of importance, are as follows:

- $m_{\text{vis}}^{\text{opt1}}(H)$: Mass of the τ_{had} and the leading⁹ light-lepton.
 $< S^2 > = 3.734 \times 10^{-1}$.
- $\Delta\eta(\tau_{\text{had}}, \ell_1)$: $\Delta\eta$ between the τ_{had} and the leading light-lepton.
 $< S^2 > = 2.457 \times 10^{-1}$.
- $m_{\text{vis}}^{\text{opt2}}(H)$: Mass of the combined τ_{had} and the sub-leading light-lepton.
 $< S^2 > = 2.025 \times 10^{-1}$.
- $\Delta\eta(\tau_{\text{had}}, \ell_2)$: $\Delta\eta$ between the τ_{had} and the sub-leading light-lepton.
 $< S^2 > = 1.864 \times 10^{-1}$.
- $m_{\text{pred}}^{\text{opt1}}(t)$: Mass of the leading b -tagged jet, the leading light-lepton, and the predicted energy for the corresponding neutrino. The prediction for the neutrino from the top-quark decay is described in Section 6.4.3.1.
 $< S^2 > = 1.596 \times 10^{-1}$.
- $\Delta R(b\text{-tagged jet}, \ell_1)$: ΔR between the leading b -tagged jet and the leading light-lepton.
 $< S^2 > = 1.142 \times 10^{-1}$.

⁸See Equation B.5 in Appendix B for more details.

⁹Note that the term “leading” refers to the p_{T} -leading. This applies for all its uses in this thesis.

- $m_{\text{pred}}^{\text{opt2}}(t)$: Mass of the leading b -tagged jet, the sub-leading lepton and the predicted energy for the corresponding neutrino of the ℓ_2 . $\langle S^2 \rangle = 1.104 \times 10^{-1}$.
- $\Delta R(b\text{-tagged jet}, \ell_2)$: ΔR between the leading b -tagged jet and the sub-leading light-lepton. $\langle S^2 \rangle = 1.009 \times 10^{-1}$.
- $\Delta\eta(\text{closest } b\text{-tagged jet}, \ell_1)$: $\Delta\eta$ between the leading lepton and the closest b -tagged jet to that lepton. $\langle S^2 \rangle = 7.401 \times 10^{-2}$.

The separation-power-based ranking of the BDT^{Lepton Assignment} input variables is derived by counting how often the variables are used to split decision tree nodes. Then, each split occurrence is weighted by the separation-gain squared it has achieved and by the number of events in the node [285].

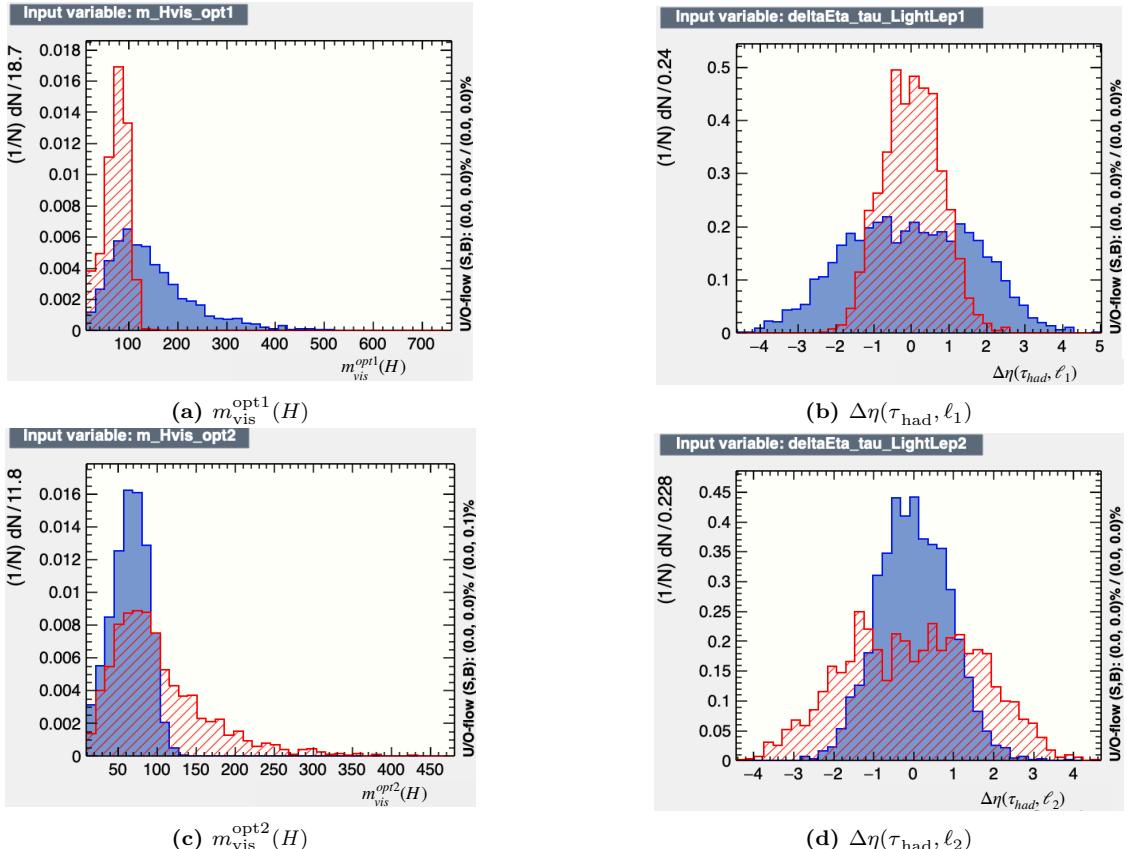
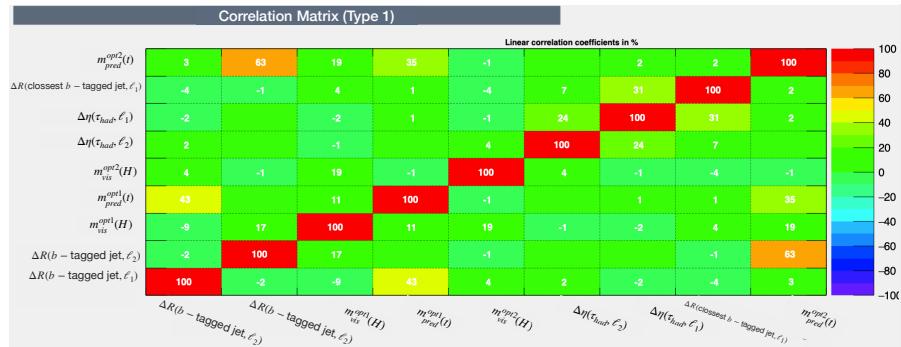


Figure 6.5: Normalised distributions of the highest ranked variables for the light-lepton assignment.

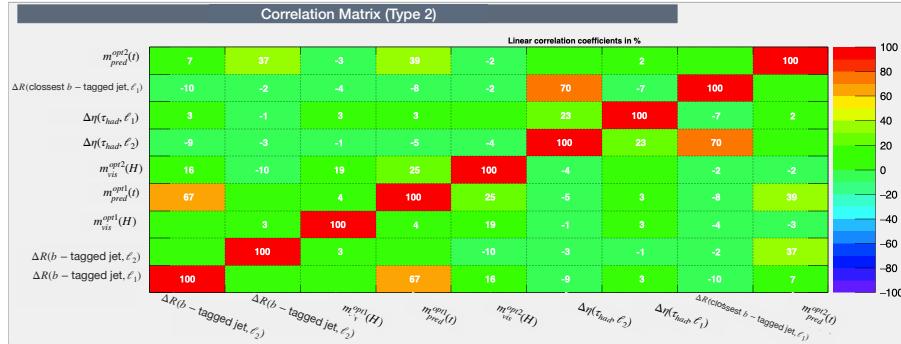
Furthermore, it is important that the selected variables are not highly correlated, as correlations can exacerbate model complexity and result in redundant in-

formation that provides no improvement to model performance. Figure 6.6 present the correlation matrices for the final input variables. A correlation matrix is a square matrix that provides a comprehensive view of the correlation coefficients between multiple variables. Note that these matrices show the bidimensional linear correlation coefficients but the BDT uses N-dimensional relations so relevant higher-order relations cannot be uncovered by the matrices.

When using larger sets of input features, as expected, an increase in the number of highly correlated pairs of variables is observed. To address this concern, for each correlated variable pair, the variable with lower separation power is removed from the model. The separation power ranking is automatically built during the training.



(a) Events with ℓ_1 from the top quark and ℓ_2 from the Higgs boson.



(b) Events with ℓ_1 from the Higgs boson and ℓ_2 from the top quark.

Figure 6.6: Matrices showing the linear correlation coefficients of the BDT input variables for the Type 1 and Type 2 samples. The correlation coefficients range from -100 to 100 , being 0 the value for totally independent variables. Higher order functional or non-functional relationships may not, or only marginally, be reflected in the value of linear correlation coefficient.

The list of variables used in the assignment BDT is chosen such that it does not overlap with that of the BDT trained to separate the signal from the background (Section 6.6). This is done in order to avoid potential biases, since the input variables of the latter are built with the former.

6.4.2.3 Optimisation of the lepton-assignment BDT hyperparameters

The hyperparameters are the parameters whose values control the learning process. They are not part of the final model but determine the values of the model parameters that the learning algorithm acquires. The values of the hyperparameters are optimised to maximise the performance of the assignment. While certain ML libraries such as PyTorch, scikit-learn, XGBoost, or TensorFlow, offer functionality for discovering the most effective hyperparameter values, the corresponding capabilities within the TMVA package of ROOT remain in a developmental stage, and presently, they are not even documented. Tests are carried to identify optimal hyperparameters using a grid-search-based algorithm, with the receiver operating characteristic (ROC) area serving as the primary figure of merit. However, better performances are achieved through the manual fine-tuning of hyperparameters.

A discussion about the different hyperparameters and their meaning is provided in the Appendix B.5. The set of hyperparameters used to train the model is presented in Table 6.13.

Hyperparam.	Value	Meaning
Type	Gradient	Boosting type for the trees.
MaxDepth	3	Maximum depth of cell tree.
Shrinkage	0.2	Learning rate for GradBoost algorithm.
NTrees	10^3	Number of trees in method.
nCuts	40	Number of grid points in variable range used in finding optimal cut in node splitting.

Table 6.13: Hyperparameters tuned for the BDT^{Lepton Assignment} training. The rest of hyperparameters are set to their default values (defined in Appendix B.5).

6.4.2.4 Treatment of the negatively-weighted events

Negatively-weighted events can pose challenges when using a BDT or other ML techniques such as NN. These issues are discussed in more detail in Section B.3. The origin of the negatively-weighted events is explained in Appendix D.1.

When training a BDT, it is necessary to address the issue of negatively-weighted events since the algorithm cannot directly handle negative weights. Two common approaches to handle this situation are: ignoring negative weights during training and using the absolute values of the weights.

In the case of ignoring negative weights during training, the BDT algorithm treats all events with negative weights as if they have zero weights (i.e. their

information is lost). This approach avoids any potential complications arising from negative weights but still preserves the positively weighted events in the training process. On the other hand, using the absolute values of the weights involves taking the magnitude of the negative weights, effectively treating them as positive weights. This second strategy allows the BDT to incorporate the information from these events while disregarding the sign of the weights.

Other options are provided by the TMVA library of ROOT, as described in Section B.3.1. However, after conducting several tests, it is observed that these alternative techniques do not exhibit comparable performance in terms of both separation power and stability when compared to the approach of ignoring negative weights.

After careful evaluation and experimentation, it is determined that excluding events with negative weights yielded better results than using the absolute weights of the events. This approach demonstrated improved performance in terms of the desired outcomes of the BDT^{Lepton Assignment} training.

However, to ensure the validity of this approach, it should be verified that the subset of positively weighted events accurately represents the behaviour of the data. Approximately 40% of the signal entries have negative weights and, hence, ignoring this subset could affect the distribution of the samples. Therefore, before discarding these events during training, it is necessary to examine whether the shape of the data distributions is significantly affected.

The shapes of the distributions using all events versus using only positively-weighted events are compared and shown in Appendix C.1. The distributions exhibit perfect compatibility within the statistical uncertainty bands. Additionally, the concentration of negative weights is practically the same for both categories, with 36.1% for Type 1 and 36.6% for Type 2. This indicates a balanced representation, as removing negative weights does not introduce bias towards either category.

In conclusion, using only the positively weighted events in the training process does not present any significant issue.

6.4.2.5 Training of the lepton-assignment BDT

The primary objective of the BDT^{Lepton Assignment} is to differentiate between the Type 1 and Type 2 categories, similar to the separation of signal and background with the region-definition BDTs. One important difference between the region-definition BDTs and the BDT^{Lepton Assignment} is that while the former is using all the simulated signal and background samples, the latter is exclusively trained on $2\ell SS + 1\tau_{had}$ signal events. This approach is justified by the objective of determining

the origin of each lepton in the signal events, a classification that is meaningful only within the context of signal processes

The use of TMVA offers the advantage of directly working with the ROOT-formatted NTuples, eliminating the need to convert them to numpy arrays or pandas dataframes.

In order to mitigate the effects of low statistics, the k -folding method (carefully described in the Appendix B.4) from cross validation is implemented using five folds. Thus meaning that the data are split in five sub-sets named folds and five BDTs are trained. Each model uses four folds for training and one for test, implying that for each BDT the train/test split is 80%/20%. After removing the negatively-weighted events, 9362 raw events are used for building the model. Of those 5518 are Type 1 and 3844 Type 2. Therefore, each of the five BDTs uses 7490 events in the training and 1872 in the classification. This cross-validation technique allows to use all the events in the dataset for the train. No validation dataset is used but the different models are compared and if all of them behave similarly, there is no overtraining.

Due to the use of k -folding, the training of the BDT^{Lepton Assignment} results in five distinct ML models. If overtraining had occurred, these models would not generalise effectively, leading to noticeable differences in the BDT score¹⁰ distributions between models. Therefore, it is crucial to ensure that the models are consistent with one another. Figure 6.7 displays the BDT score distributions for all five folds simultaneously. The purpose of this visualisation is to assess the compatibility of the five models and verify the absence of overtraining. As observed in the figure, the BDT score distributions are consistent across the folds, indicating that there is no evidence of overtraining. The AUC of each fold is presented in Table 6.14. Unfortunately, TMVA documentation does not provide method for retrieving logloss directly .

Additionally, it is essential to compare the distributions of the train and test samples for each model. The greater the similarity between these distributions, the better the performance of the model. However, it is crucial to note that an exact agreement between the train and test samples could indicate a potential bug or issue with the model. Figure 6.8 presents, for a single model, the BDT^{Lepton Assignment} simultaneously displaying the train and test samples while distinguishing between the two event types. The train versus test plots of the models corresponding to the other folds, are in Figure B.9 of Appendix B. By examining the figure, it can be observed that the train and test samples exhibit compatibility, indicating that the models are performing well and are not overfitting to the training data.

¹⁰The score of the BDT is the result of the model prediction for a given event.

Fold	AUC
1	0.933
2	0.932
3	0.944
4	0.947
5	0.928
Mean	0.9372 ± 0.0073

Table 6.14: AUC for the five folds of the BDT^{Lepton Assignment}. In the last row, the mean is presented with the corresponding standard deviation.

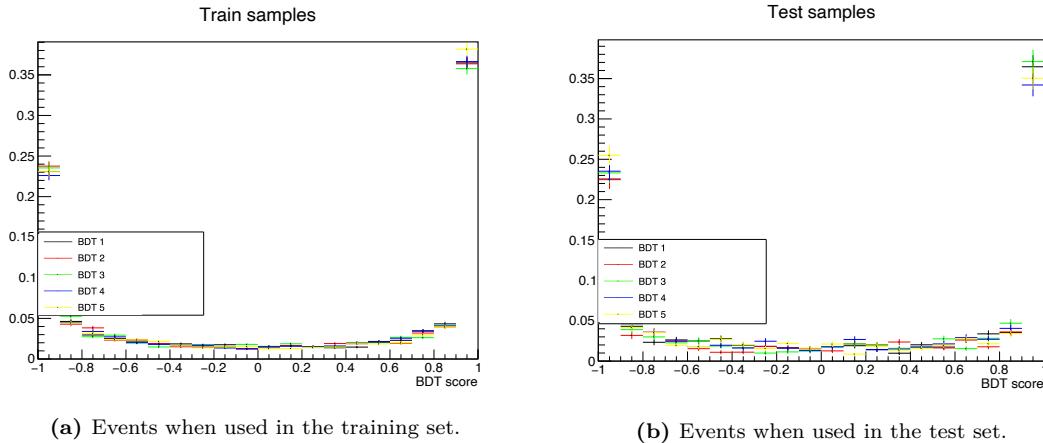


Figure 6.7: Comparison of the profiles of the BDT^{Lepton Assignment} scores for the different models (folds). Observe that the scores exhibit similar distribution shape, compatible within the statistical uncertainty, indicating that none of the the models are affected by any particular subset of data used for training.

Figures 6.9a and 6.9b present the ROC curves and BDT score, respectively. The former is separated by folds and the latter by categories and combining the data from all five folds. The ROC curves illustrates the balance between correctly identifying Type 1 instances and misclassifying Type 2 instances at various classification thresholds. A higher ROC curve and a larger area under the ROC curve (AUC) indicate improved classification performance. A more comprehensive explanation of these concepts can be found in Appendix B.6. Notably, the Type 1 and Type 2 categories replace the conventional positive and negative instances. The substantial AUC observed in the ROC curves signify strong performance. For a detailed view of the ROC curves for each fold individually, refer to Figure B.10 in the Appendix B.

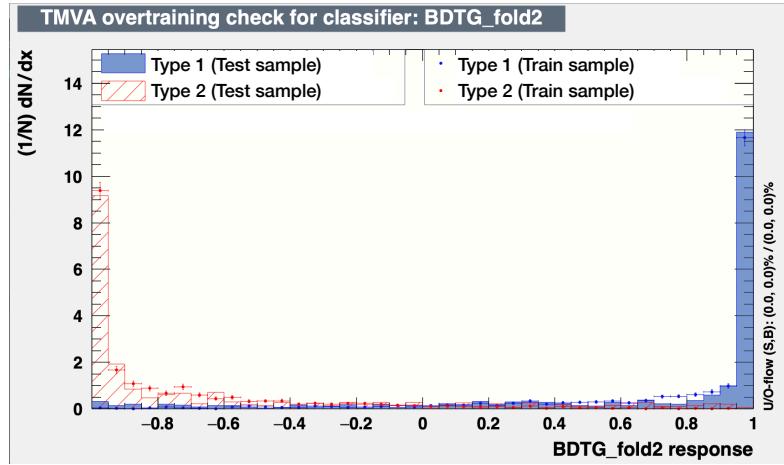


Figure 6.8: $\text{BDT}^{\text{Lepton Assignment}}$ distributions for test and train samples superimposed in one of the folds.

Regarding the $\text{BDT}^{\text{Lepton Assignment}}$ score in Figure 6.9b, it is evident that the Type 1 and Type 2 categories exhibit distinct peaks at opposite extremes of the distribution. This significant separation confirms the effectiveness of the $\text{BDT}^{\text{Lepton Assignment}}$ when differentiating between the two categories. It is worth noting that in Figure 6.9b, the bullets representing the train sample and the shadowed area representing the test samples appear identical due to the way TMVA combines the folds in a single histogram. However, the accurate assessment of the test versus train comparison is shown in Figure B.9.

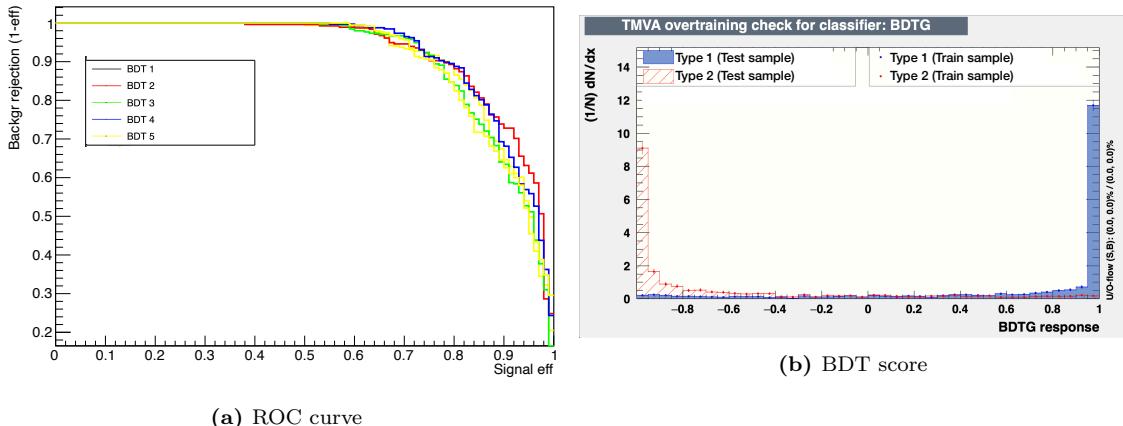


Figure 6.9: (a) ROC and (b) score combining all folds for the $\text{BDT}^{\text{Lepton Assignment}}$. Each curve in (a) represents a different fold.

6.4.2.6 Model application and classification-threshold selection

Although the model-building tool is developed as an independent and self-contained application, the $\text{BDT}^{\text{Lepton Assignment}}$ model is integrated into the tHqLoop framework. As a result, during the production of post-processed NTuples (i.e. the tHqLoop output), the lepton origin in the $2\ell \text{SS} + 1\tau_{\text{had}}$ is already used for variable construction. This represents a great advantage since it simplifies the post-processing of NTuples.

Since k -folding is used, there are several models and, hence, it is necessary to define a criteria to decide which model apply to each event. On the one hand, for the events used in the training, the BDT used is the one in which that event was part of the test sample. On the other hand, for the events that do not take part in the training, the choice of BDT is randomised using the variable `event number`. Since the event number has no physical meaning, it is equivalent to randomly assigning a BDT to each event.

Note that the $\text{BDT}^{\text{Lepton Assignment}}$ output is not a category but a real number ($\text{BDT}^{\text{Lepton Assignment}}$). To define to which category assign the event, a threshold value is selected. All the events with $\text{BDT}^{\text{Lepton Assignment}}$ scores above the selected threshold are classified as Type 1 and, contrary, if the BDT score is below as Type 2. A search of the optimal threshold is presented in Table 6.15 and $\text{BDT}_{\text{Threshold}}^{\text{Lepton Assignment}} = -0.315$ is the best value for separating the two categories.

$\text{BDT}_{\text{Threshold}}^{\text{Lepton Assignment}}$	-0.45	-0.4	-0.35	-0.33	-0.32	-0.315	-0.31	-0.395	-0.3
Accuracy (%)	88.12	88.03	87.97	88.23	88.33	88.39	88.36	88.32	88.15

Table 6.15: Different studied $\text{BDT}^{\text{Lepton Assignment}}$ thresholds compared to its correspondent accuracy. The first row is the value on the lepton-assignment-BDT-score threshold ($\text{BDT}_{\text{Threshold}}^{\text{Lepton Assignment}}$) that is used to separate between Type 1 and Type 2 events. If for an event the score is larger than threshold value, the leading lepton is considered to come from the top quark. The accuracy on the second row is calculated using the truth-level information.

6.4.2.7 Results of the BDT-based method for lepton assignment

The result obtained by this method can be compared to the one given by the cut-based alternative method described at the beginning of this section. A summary of the performance of the two methods is given in Table 6.16. The performance of the BDT-based method method surpasses that of the alternative approaches.

Accuracy of the light-lepton-origin assignment			
	Total (100%)	$H \rightarrow \tau\tau$ (83.08%)	$H \rightarrow WW$ (16.92%)
BDT-based method	88.39%	88.44%	88.18%
Cut-based method	83.86%	84.24%	81.80%

Table 6.16: Accuracy calculated by comparing the prediction of the method to the true value. The true value is obtained using the labelling described in Section 6.4.2.1. This labelling is only available for $H \rightarrow \tau\tau$ and $H \rightarrow WW^*$. Only the events with successful reconstruction-level and truth-level matching are used. From these events, 83% are $H \rightarrow \tau\tau$ and 17% $H \rightarrow WW^*$. The NN-based initial approach is not presented since it does not have any proper labelling.

6.4.3 Reconstruction of the top quark and the Higgs boson

In order to suppress background events, it is desirable to reconstruct the kinematics of both the top quark and the Higgs boson. The accurate reconstruction of these particles is a challenging task due to the presence of four neutrinos in the final state, of which at least three are from the Higgs boson and one is from the top quark¹¹. In this work, the strategy used to fully reconstruct the top and the Higgs boson consists on first reconstructing the top-quark system and then the Higgs-boson chain. Since it is known the total E_T^{miss} , if the missing energy is reconstructed for the top quark, it is possible to access the \vec{E}_T^{miss} (Higgs) based on:

$$\vec{E}_T^{\text{miss}} (\text{total}) = \vec{E}_T^{\text{miss}} (\text{Higgs}) + \vec{E}_T^{\text{miss}} (\text{top}).$$

The possible final-state configurations when the Higgs boson decays to $\tau^-\tau^+$ are described in Table 6.17. The first two final-state columns in this table represent 92.7% of $H \rightarrow \tau\tau$ events while the last one can be neglected. This means that it is a relatively safe option to assume that the τ_{had} is produced from the decay of the Higgs boson. This is calculated from the numbers in Table 6.9.

Before reconstructing the top quark or Higgs boson, the two charged-light leptons in the final state have to be assigned to their corresponding parents. In most of cases, one ℓ comes from the Higgs boson and the other from the top quark.

6.4.3.1 Reconstruction of the top quark

To reconstruct the top quark, it is essential to have complete knowledge of the 4-momenta of both the W boson and the b quark. Therefore, if the neutrino is reconstructed, the top-quark system can be reconstructed as well.

¹¹This is assuming the most common decay channel, $H \rightarrow \tau\tau$.

Parent particle	Decaying particle	Decay sub-channels		
H	τ	$\tau_{\text{had}} \nu_\tau$	$\tau_{\text{had}} \nu_\tau$	$\ell_1 \nu_{\ell_1} \nu_\tau$
	τ	$\ell_1 \nu_{\ell_1} \nu_\tau$	$\ell_2 \nu_{\ell_2} \nu_\tau$	$\ell_2 \nu_{\ell_2} \nu_\tau$
t	W	$\ell_2 \nu_{\ell_2}$	$\ell_1 \nu_{\ell_1}$	$\tau_{\text{had}} \nu_\tau$
	b	$b\text{-jet}$	$b\text{-jet}$	$b\text{-jet}$

Table 6.17: Possible configurations of the final state objects in the $2\ell + 1\tau_{\text{had}}$ channel when the Higgs boson decays to τ_{had} and τ_{lep} .

To properly reconstruct the top quark it is necessary to determine the z-component of the neutrino momentum, $p_z(\nu_{\text{top}})$. In order to achieve this, various hypotheses are examined using parton-level information to derive $p_z(\nu_{\text{top}})$ from the z-component momentum of ℓ_{top} . One of the initial hypotheses tested is whether the z-component of the neutrino momentum in the top-quark system exhibits a linear dependence with a coefficient α on the z-component of the momentum of the light lepton, $p_z(\ell_{\text{top}})$. Being the parameter α a real number ranging from -1 to 1 , the hypothesis $p_z(\nu_{\text{top}}) = \alpha \cdot p_z(\ell_{\text{top}})$ is tested in Figure 6.10. This hypothesis is tested by studying the parton-level information of the top quark and its neutrino. The results indicate that the dependence between these two parts is not linear and, hence, this hypothesis for $p_z(\nu_{\text{top}})$ is not supported.

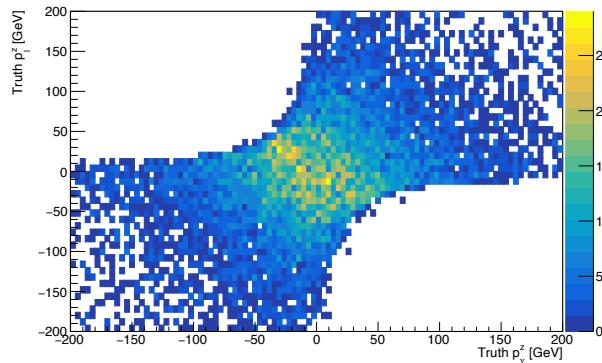


Figure 6.10: Distributions comparing $p_z(\ell_{\text{top}})$ and $p_z(\nu_{\text{top}})$ at parton level to check if there is a dependence.

Regarding the $\vec{E}_{\text{T}}^{\text{miss}}$ of the top-quark system, constraints have to be imposed on the ν_{top} . Through the correlation analysis between the neutrino and the top-quark

lepton, the following constraints are found in a linear fit:

$$p_T(\nu_{\text{top}}) = \frac{1615.98 \text{ GeV}^2}{p_T(\ell_{\text{top}})},$$

$$\phi(\nu_{\text{top}}) = \phi(\ell_{\text{top}}) \pm \frac{\pi}{2},$$

where $\phi(\nu_{\text{top}})$ and $\phi(\ell_{\text{top}})$ are, the azimuthal angles of the neutrino and the reconstructed ℓ_{top} , respectively, and $p_T(\nu_{\text{top}})$ and $p_T(\ell_{\text{top}})$ their transverse momentums. The behaviour represented by these constraints can be observed in Figure 6.11.

To resolve the sign ambiguity in the \pm of the second constraint, the azimuthal angle of the b quark originated on the top-quark decay ($\phi(b_{\text{top}})$) considered. The following condition is imposed:

$$\phi(b_{\text{top}}) - \phi(\ell_{\text{top}}) \geq 0 \rightarrow \phi(\nu_{\text{top}}) = \phi(\ell_{\text{top}}) - \frac{\pi}{2},$$

$$\phi(b_{\text{top}}) - \phi(\ell_{\text{top}}) < 0 \rightarrow \phi(\nu_{\text{top}}) = \phi(\ell_{\text{top}}) + \frac{\pi}{2}.$$

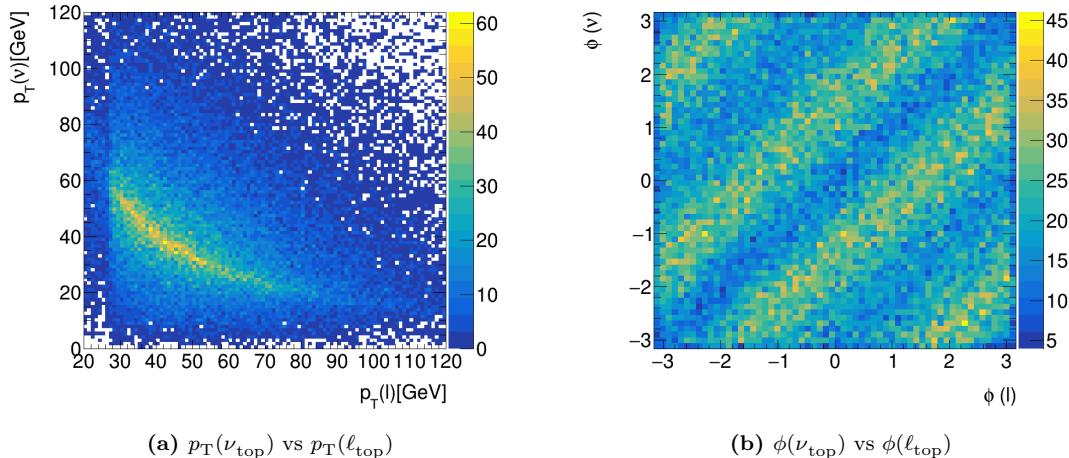


Figure 6.11: Truth-level distributions comparing the p_T (a) and ϕ (b) of the ν_{top} and ℓ_{top} . With these two plots are obtained the imposed conditions at reconstruction level on $p_T(\nu_{\text{top}})$ and $\phi(\nu_{\text{top}})$.

The calculation of the reconstructed top-quark mass, m_t^{reco} , incorporates several variables, including the lepton associated with the top quark, the leading b -jet, and the neutrino $p_T(\nu_{\text{top}})$ and $\phi(\nu_{\text{top}})$. However, in this calculation, the $p_z(\nu_{\text{top}})$ component is set to zero due to its unknown value. With this it is possible to reconstruct the mass of the top quark and to calculate $p_T^{\text{miss}}(H)$. The latter is used in the reconstruction of the Higgs boson.

6.4.3.2 Reconstruction of the Higgs boson

A straightforward method for reconstructing the Higgs-boson mass involves summing its visible decay products. Assuming the $H \rightarrow \tau\tau$ decay mode:

$$m_H^{\text{vis}} = \text{mass}(\tau_{\text{vis},1} + \tau_{\text{vis},2}),$$

where τ_{vis} is either the visible part of the τ_{had} or the light lepton from the τ_{lep} .

However, this approach overlooks the contribution of neutrinos and is therefore insufficient. By neglecting smaller contributions to the $E_{\text{T}}^{\text{miss}}$, such as energy loss in detector material, it becomes possible to relate $p_{\text{T}}^{\text{miss}}(H)$ to the measured $E_{\text{T}}^{\text{miss}}$ and the $p_{\text{T}}^{\text{miss}}(t)$ by imposing the relation:

$$p_{x,y}^{\text{miss}}(H) = p_{x,y}^{\text{miss}}(\text{measured}) - p_{x,y}^{\text{miss}}(\text{top, reconstructed}).$$

This characteristic can be effectively employed in established methods for reconstructing the Higgs-boson mass. The quantities $p_{\text{T}}(\nu_{\text{top}})$ and $\phi(\nu_{\text{top}})$, calculated in the previous section, are used in this relation.

There are several techniques to reconstruct the $\tau^-\tau^+$ system. Among these, the most popular methods are:

- Partial invariant mass: Also known as *Transverse Mass Method*, it uses the m_H^{vis} and the transverse mass of the $\tau^-\tau^+$ system:

$$m_{\text{T}}^2 = m^2(\tau_{\text{vis}1}, \tau_{\text{vis}2}, E_{\text{T}}^{\text{miss}}(\tau\text{-syst})).$$

This technique was first used in the UA1 experiment [286] and has the advantage that it can be applied to all signal event topologies for the $H \rightarrow \tau^-\tau^+$ channel (i.e. $\tau_{\text{had}}\tau_{\text{had}}$, $\tau_{\text{had}}\tau_{\text{lep}}$ and $\tau_{\text{lep}}\tau_{\text{lep}}$). The dedicated studies have shown that the distributions of the reconstructed m_H using this method is lower than the one obtained with the parton-level m_H .

- Collinear approximation [287]: This is one of the most commonly used techniques for the reconstruction of the invariant mass of a $\tau\tau^+$ system with the presence of invisible decays products. It is mainly based on the assumption that τ -lepton and all its decay products are collinear. In our search, these can be translated into having a boosted Higgs boson, which is not the case for the tHq production. Another inconvenience of this method is that it is very sensitive to the resolution of $E_{\text{T}}^{\text{miss}}$, being likely to overestimate the mass.
- In the bibliography there are other mechanisms to solve this issue such as the Recursive Jigsaw Reconstruction [288] or the Kinematic Likelihood Fitter [289]. These other methods are not explored in this thesis.

- A more sophisticated technique that outperforms the mentioned above is the `MissingMassCalculator` (MMC), which was originally developed for the $H \rightarrow \tau\tau$ analysis [290, 291]. The MMC does not suffer from any of the limitations of the previous methods. It allows for a more complete reconstruction of event kinematics with improved resolution for the mass of Higgs-boson-decay. This method is based on assuming that the source of $E_{\text{T}}^{\text{miss}}(H\text{-system})$ is due to the neutrinos from τ decays exclusively. Then, for each event, it scans over all possible configurations of the visible and invisible τ -decay products. It is based on the requirement that mutual orientations of the neutrinos and other decay products are consistent with the mass and decay kinematics of a τ lepton. This is achieved by minimising a likelihood function defined in the kinematically allowed phase space region. Due to its better functioning, the MMC is the technique used in this analysis.

6.5 Background estimation

In this analysis, background events are categorised into two distinct types: “reducible” backgrounds and “irreducible” backgrounds.

- **Irreducible backgrounds:** Irreducible backgrounds involve final states with the same physical objects as the signal. The primary irreducible backgrounds in this analysis include Diboson, tW , $t\bar{t}Z$, $t\bar{t}H$, $t\bar{t}W$ and tZq .
- **Reducible backgrounds:** The reducible backgrounds are originated from inaccuracies in the reconstruction process. In other words, some objects are incorrectly reconstructed as one of the objects present in the signal process. The misidentification of one or more objects can make non-signal process have the same final-state objects as the signal process. If the experimental tools and techniques are improved, the effects of this type of background can be mitigated.

In the $2\ell + 1\tau_{\text{had}}$ channel, the dominant backgrounds consist of reducible events where jets misidentified as τ_{had} are present. This is particularly observed in the $t\bar{t}$ and $Z + \text{jets}$ backgrounds. Additionally, other backgrounds include diboson (VV) and $t\bar{t}X$ productions, such as $t\bar{t}H$, $t\bar{t}Z$, and $t\bar{t}W$. Note that some background processes can belong to both categories simultaneously.

In Section 5.2.3 it is already noted that the task of rejecting quark- and gluon-originated jets for τ_{had} reconstruction is challenging. Particularly, the difficulty in

distinguishing hadronic taus from jets leads to a high rate of misidentification. But this is not the only type of reducible background that plays a role in the tHq search in the $2\ell + 1\tau_{\text{had}}$ final state. The sources of reducible background are:

- Gluon-initiated jet misidentified as τ_{had}
- Quark-initiated misidentified as τ_{had}
- Electrons and muons misidentified as τ_{had}
- Jets or non-prompt leptons¹² misidentified as prompt light leptons
- Electrons with wrongly reconstructed charge.

Table 6.18 provides a comprehensive overview of the estimated contribution of the different sources of reducible backgrounds in the $2\ell + 1\tau_{\text{had}}$ channel. As can be seen, the misidentification of τ_{had} is the most important source of background in the $2\ell \text{ OS} + 1\tau_{\text{had}}$ channel while for the $2\ell \text{ SS} + 1\tau_{\text{had}}$ the fraction of misidentified particles is much smaller.

Backgrounds	$2\ell \text{ OS} + 1\tau_{\text{had}}$	$2\ell \text{ SS} + 1\tau_{\text{had}}$
Irreducible	472 ± 6	81.3 ± 1.6
Gluon identified as τ_{had}	1567 ± 27	9.2 ± 0.6
Quark identified as τ_{had}	5640 ± 70	23.0 ± 1.1
Other misidentified τ_{had}	2240 ± 50	2.8 ± 0.4
Misidentified e/μ	28.3 ± 1.9	19.6 ± 1.6
e charge flip	340 ± 13	33.0 ± 3.2

Table 6.18: Origin of the backgrounds determined by the matching of events at truth and reconstruction levels. The yields have been obtained at preselection level. Note that one event can fall in several categories simultaneously [241].

Accurately determining the expected fraction of each background process is crucial for obtaining high-quality results. While the yields for irreducible backgrounds are reliably estimated in the MC-based simulations, determining the abundance of reducible backgrounds is not a straightforward task. This difficulty arises because reducible backgrounds can heavily rely on the specifics of the physics simulation, particularly in non-perturbative regions where simulation reliability may be questionable [268]. Additionally, these backgrounds are influenced by the modelling of the material composition and response of the detector.

¹²Electrons and muons from meson decays and electrons from photon conversions. The heavy-flavour hadron decays produce a weak boson that decays to leptons.

Among the approaches for the estimation of misidentification rates, the *template fit method* stands out, especially for taking advantage of the high number of available MC samples. This data-driven strategy to estimate the abundance of the different reducible backgrounds involves matching reconstructed objects in simulated datasets with the corresponding simulated-parton-level objects using $\Delta R = 0.2$ cones. By doing this for hadronic taus, the templates for physical objects (e , μ , τ_{had} , quark-initiated jet, gluon-initiated jet, and unknown¹³) are created. These templates are constructed in two iterations. The first iteration finds the truth label of the reconstructed leptons and the second the does the same for the jets. The composition at parton-level of the reconstructed objects is presented in Figure 6.12. There it can be seen that while the light-lepton objects in the tHq process are typically well reconstructed, the reconstructed τ_{had} from the backgrounds is not a true τ in most cases.

After preparing the templates, it becomes essential to determine their normalisation factors (also referred as scale factors). This is done by fitting the templates to match the data in background-enriched control regions specially dedicated to estimating the templates.

The estimation of the misidentification-based background events and the assessment of the corresponding uncertainty are done separately for the $2\ell \text{OS} + 1\tau_{\text{had}}$ and $2\ell \text{SS} + 1\tau_{\text{had}}$ channels. For the $2\ell \text{OS} + 1\tau_{\text{had}}$ channel, the estimation of the scale factors is conducted as follows:

1. To compensate for the potential mismodelling of quark- and gluon-initiated jets mimicking τ_{had} , the template fit of the quark and gluon components is performed in a region enriched with misidentified- τ_{had} . This region of the phase space is defined in a way that the τ_{had} candidates do not pass the `medium` identification criteria (see Section 6.3.4 for identification working points). The template fit is used here is the nominal method. It is assumed that the scale factors derived this way can be extrapolated to the rest of the analysis regions and, hence, these are applied to all MC samples. The scale factors are derived separately for:
 - 1-prong and 3-prong τ_{had} decays.
 - Various p_{T} (τ_{had}) bins.
 - 1 and 2 b -jets multiplicity.

Therefore, different scale factors are derived for different regions of the phase space. No significant dependence on $|\eta(\tau_{\text{had}})|$ is observed.

¹³Refers to objects for which the correspondent truth object is not found.

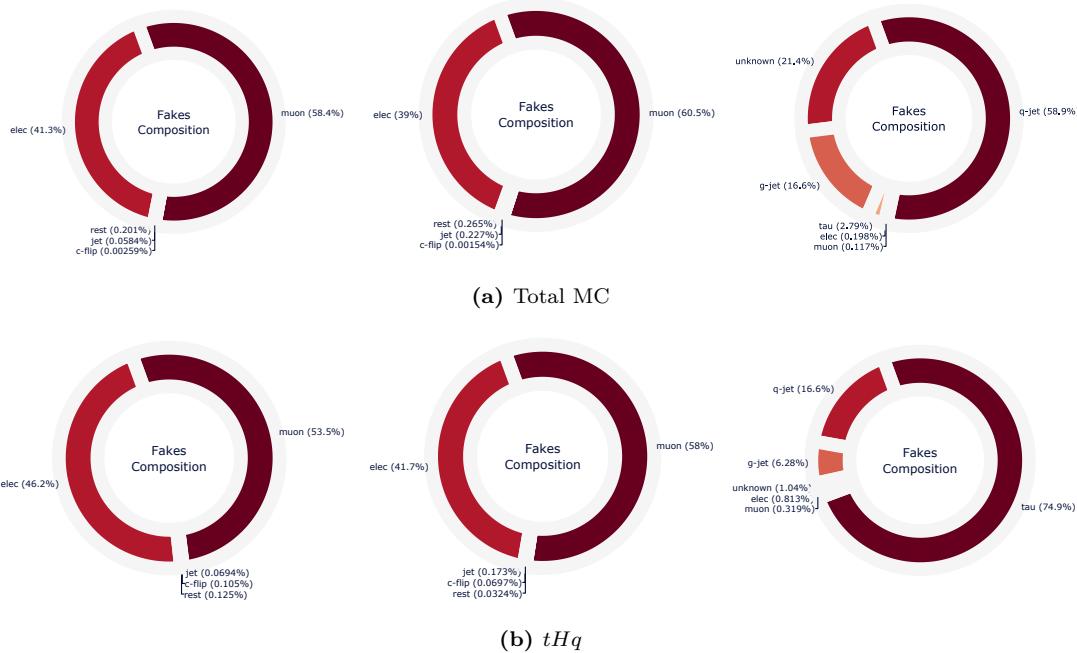


Figure 6.12: Composition at parton level of the reconstructed leading (left) and sub-leading (middle) leptons, as well as τ_{had} (right) passing the selection requirements in the $2\ell + 1\tau_{\text{had}}$ analysis with oppositely charged light leptons. These fractions are calculated using (a) all MC event samples and (b) MC samples of the tHq production.

2. Secondly, the number of real and misidentified τ_{had} candidates is counted in the same region. With these counts, the correction factors that align the MC to the data are derived. This is known as the *counting method* and it is the alternative approach used to assess the uncertainty of the nominal method. The counting method uses same the same binning, prongness and b -jet-multiplicity divisions as the template fit.
3. Finally, using the samples with adjusted misidentified- τ_{had} rates, the estimation of the misinterpretation rates for light leptons is calculated. Similarly to the τ_{had} case, the correction factors to be applied to MC are determined to match the templates to the data in a misidentified- ℓ -enriched region. This region is defined by requiring that the light leptons fail their identification and isolation requirements described in Table 6.6. A simplified classification scheme is used to construct the templates for the leptons.

Meanwhile, for the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel, the condition on the electric charge of the light leptons produces a drastic decrease in the influence of the τ_{had} -

misidentification rates. This is shown in Figure 6.13, which is the same as Figure 6.12 but only with $2\ell \text{SS} + 1\tau_{\text{had}}$ events.

Due to the limitation of the statistical sample, the techniques applied for the estimation of the backgrounds arising from misidentified objects in the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel are a simplified version of those used in the $2\ell \text{OS} + 1\tau_{\text{had}}$. In the $2\ell \text{SS} + 1\tau_{\text{had}}$ only the counting method is used and, regarding the binning, the sample is not split by the number of b -tagged jets

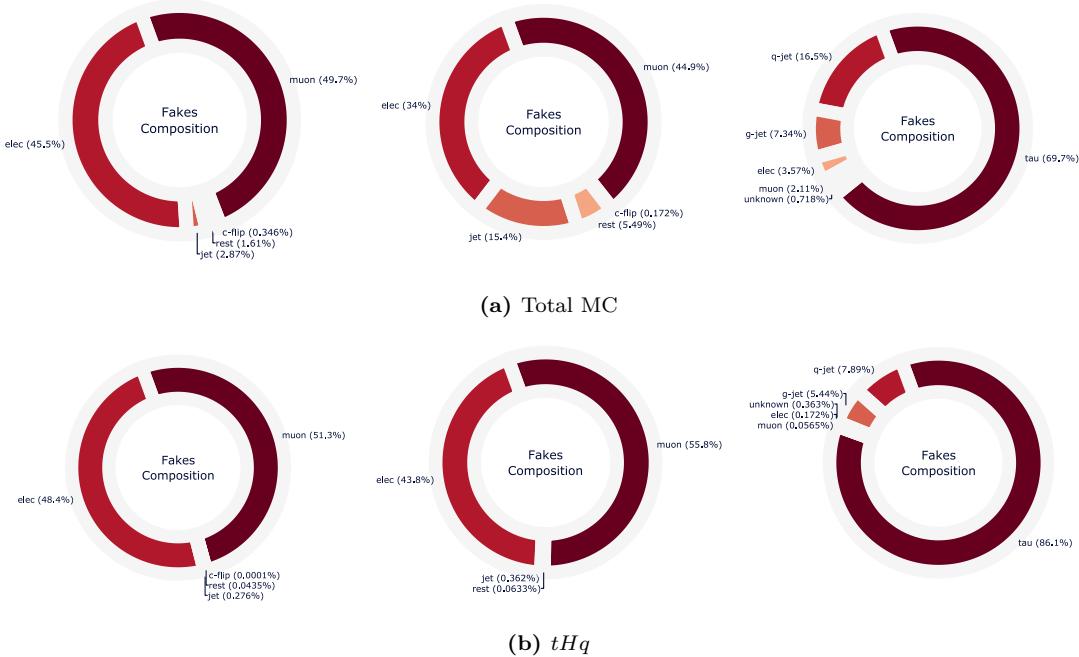


Figure 6.13: Composition at parton level of the reconstructed leading (*left*) and sub-leading (*middle*) leptons, as well as τ_{had} (*right*) in the $2\ell \text{SS} + 1\tau_{\text{had}}$ analysis channel. The numbers are computed for the combined MC samples (a) and tHq production only (b). An electron charge ID selector is applied to this data. Note that while the pie charts distinguish between quark- and gluon-initiated jets, these two processes use a single template in the template fit method for the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel.

As Figure 6.13 shows, while the fraction of correctly identified τ_{had} improves in the $2\ell \text{SS} + 1\tau_{\text{had}}$, the influence of jets being misinterpreted as leptons is considerable for the sub-leading leptons (see the green section in the second pie chart of Figure 6.13a). To minimise the background arising from incorrect identification of the electron charge, an ECIDS is implemented in the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel [209].

6.6 Event selection

The event selection consists of enriching the relative contribution of the signal over background. By doing so, a region of the phase space is defined that is enriched with signal events. This region is called signal region (SR).

The signal selection is performed in several steps and using different methods. First of all, a preselection region (PR) is defined, where the physical objects are selected according to the detector acceptance and some other physics criteria. The PR is described in Section 6.6.1. Then, discriminant variables are defined and used as input features for a BDT that can distinguish between signal-like and background-like events by creating a discriminant known as “BDT score”. The training of the BDTs for the region definition is discussed in Section 6.6.2. Finally, by applying requirements on the BDT outputs, the SR is defined (see Section 6.6.3). Additional regions dedicated to control the modelling of the background processes can also be defined (see Section 6.6.4). The regions of the phase space enriched with particular background processes are referred to as control (CR) or validation regions (VR). The difference between CR and VR is that while the former are considered into the fit calculations (discussed in Section 6.8), the latter are not. The SR, CRs and VRs are all orthogonal and are subspaces of the PR.

Two figures of merit are used to optimise both the fraction of signal events in the data and the absolute number of signal events. These metrics are the S/B or purity and the signal significance, respectively:

- **Purity:** The purity of a process is defined as the ratio between the event yields of the target process and the total yields. For the signal process, instead of purity, the the signal to background ratio (S/B) is used.
- **Significance:** This metric does not only account for the relative fraction of the process of interest but also the total amount of events. Using the significance as metric enhances the importance of keeping enough statistics.

The definition of the significance estimator used in this work is the one given in Reference [292]:

$$\text{Significance} = \sqrt{2[(t+r)\ln(1+t/r) - r]} \approx \frac{t}{\sqrt{t+r}},$$

where t is the number of events of the target process and r is the number of yields for the rest of processes combined. This can be used not only to evaluate the signal significance but also the significance of the given background processes in the dedicated background-enriched regions.

6.6.1 Preselection

The starting point on the region definition is a collection of events that are reconstructed as tHq events and with objects that pass the reconstruction requirements defined in Section 6.3. Firstly, preselection requirements are applied to guarantee the orthogonality between the different tHq channels. These requirements are the obvious ones in terms of number of final-state light leptons and hadronically decaying taus and are exclusive of the $2\ell + 1\tau_{\text{had}}$ channel.

- One hadronically-decaying tau: $n(\tau_{\text{had}}) = 1$.
- Two light-flavoured-charged leptons: $n(e/\mu) = 2$.

The multiplicity of jets is constrained as well. From the Feynman diagrams in Figure 6.1 it can be seen that one non- b -tagged jet and one b -tagged jet can be expected. Although there are two b -type quarks in the Feynman diagram, the second b -quark frequently goes undetected because its p_{T} distribution is peaking around 2 or 3 GeV (see Figure 6.14) and, hence, it does not pass the p_{T} threshold of the detector. For this reason, when only one jet is b -tagged, it is assumed to either be the b from the top-quark decay or a jet from secondary radiations. These choices are motivated by maximising the signal acceptance while minimising the background contamination.

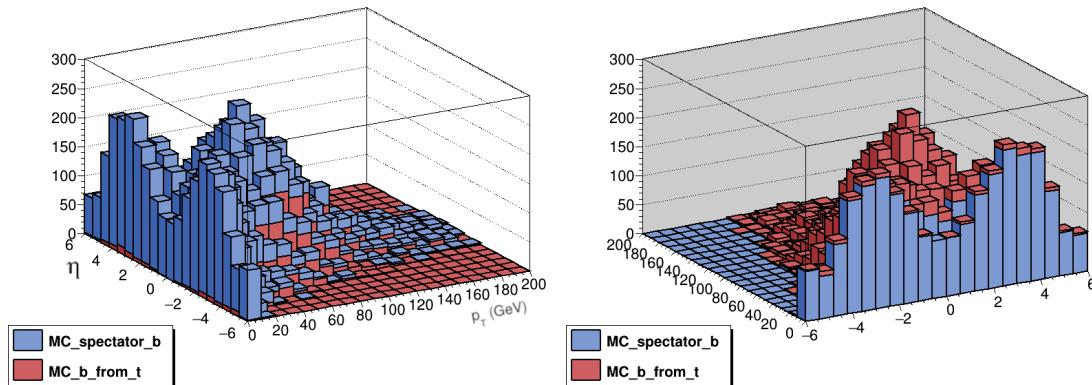


Figure 6.14: Truth-level p_{T} vs η distributions for the spectator (or second) b quark and the b quark produced in the top quark decay for the tHq events.

- In total from two to six jets are requested. In the distribution of jets within the tHq sample can be observed that, as expected, the number of tHq events drop at higher jet multiplicity: $n(\text{jet}) = [2, 6]$. This broad range in jet multiplicity is chosen strategically, as it will later be incorporated in the training of the BDTs.

- From which exactly one or two are b -tagged jets: $n(b\text{-jet}) = [1, 2]$.
- For its detection and identification, the jets are required to be in $|\eta(\text{jet})| < 4.5$ with $p_T(\text{jet}) > 20 \text{ GeV}$.
- For b -tagged jets the pseudorapidity requirement is tighter $|\eta(b\text{-jet})| < 2.5$ because that is the coverage of the ID, which is necessary for b -tagging.
- The preselection conditions also account for the geometrical acceptance of the detector and the trigger thresholds presented in Section 3.3. Reconstructed electrons, muons and taus must satisfy the acceptance requirements discussed in Sections 6.3.2, 6.3.3 and 6.3.4. The p_T requested for electrons and muons is based on single-lepton triggers (see Section 6.3.1). This requires the leading lepton to have $p_T > 27 \text{ GeV}$, the sub-leading lepton to have $p_T > 20 \text{ GeV}$ and the sub-sub-leading one to have $p_T > 14(20) \text{ GeV}$ for e/μ (τ_{had}).

The higher is the p_T of an object, the more efficient is its reconstruction. Therefore, setting the p_T requirements at the PR aims to balance between achieving good reconstruction quality and ensuring a sufficient amount of data (statistics) for analysis.

Other constraints are also applied

- To remove the low- E_T^{miss} backgrounds, a minimum E_T^{miss} of 5 GeV is imposed. By demanding a maximum E_T^{miss} of 800 GeV only one event is lost. By doing this, the BDTs can later use the entire E_T^{miss} distribution to perform cuts.
- The sum of the charge of the leptons must be ± 1 .

A summary of the preselection requisites is presented in Table 6.19 and the effect of such conditions can be observed in Figure 6.15. The events that pass the preselection requirements conform the PR and the yields at this level are presented in Table 6.20. Using the events in the PR, the BDTs presented in Section 6.6.2 are trained to define the SR and CRs, as it is described in Sections 6.6.3 and 6.6.4, respectively.

6.6.2 BDTs for region definition

In order to distinguish the tHq signal process from the backgrounds and to create background-dedicated regions, several gradient BDTs for binary classification are trained, optimised and injected. These are:

Object	Multiplicity	Momentum (GeV)	Pseudorapidity
Light leptons	$n(e/\mu) = 2$	$p_T(e) > 14$ $p_T(\mu) > 14$	$ \eta(e) < 2.47$, $ \eta(e) \notin [1.37, 1.52]$ $ \eta(\mu) < 2.50$
Hadronic tau	$n(\tau_{\text{had}}) = 1$	$p_T(\tau_{\text{had}}) > 20$	$ \eta(\tau_{\text{had}}) < 2.50$, $ \eta(\tau_{\text{had}}) \notin [1.37, 1.52]$
Jets	$n(\text{jet}) = [2, 6]$	$p_T(\text{jet}) > 20$	$ \eta(\text{jet}) < 4.5$
b -tagged jets	$n(b\text{-jet}) = [1, 2]$	$p_T(b\text{-jet}) > 20$	$ \eta(b\text{-jet}) < 2.5$
E_T^{miss}		$p_T(E_T^{\text{miss}}) \in [5, 800]$	

Table 6.19: Preselection requirements. Additionally, all leptons are required to fulfil the tight-lepton definition. The relative sign on the light-leptons electrical-charge is also used in the preselection to separate the $2\ell \text{ SS} + 1\tau_{\text{had}}$ and $2\ell \text{ OS} + 1\tau_{\text{had}}$ channels.

- BDT($tHq|_{\text{OS}}$): Trained to discriminate the tHq process from all backgrounds simultaneously in the $2\ell \text{ OS} + 1\tau_{\text{had}}$ channel. This BDT is later used to define the SR for tHq in this channel.
- BDT($t\bar{t}|_{\text{OS}}$): Trained to separate $t\bar{t}$ from the rest of processes but, in practice, what it does is discerning between $t\bar{t}$ and $Z + \text{jets}$ in the $2\ell \text{ OS} + 1\tau_{\text{had}}$ channel. It is used to define the regions for these two background processes.
- BDT($tHq|_{\text{SS}}$): Trained for the $2\ell \text{ SS} + 1\tau_{\text{had}}$ channel to discriminate the tHq signal from the rest of processes.

The steps involved in building these models closely resemble those described for the model in Section 6.4.2. The primary distinction lies in the training datasets: while the model for lepton assignment is trained using the tHq ($2\ell \text{ SS} + 1\tau_{\text{had}}$) samples, the BDTs for region definition are trained utilising the simulated samples from all processes (both signal and background) within a channel. Another key difference is that these BDTs are developed using the XGBoost software library, which specialises in gradient boosting [293].

Like in the case of the BDT^{Lepton Assignment}, the k -folding cross-validation technique (see Appendix B.4) with $k = 5$ is used to ensure that the performance assessment is not biased by a particular random split of training and test data. This split is based on the `event number` variable.

6.6.2.1 Negative-weights strategy

Typically two strategies are employed to address negatively-weighted events: either disregarding these events or using their absolute weights. Section 6.4.2.4 delves into the advantages and disadvantages of each approach, and Appendix D offers further discussion on this topic. Here, roughly 30% of the simulated events have negative MC weights.

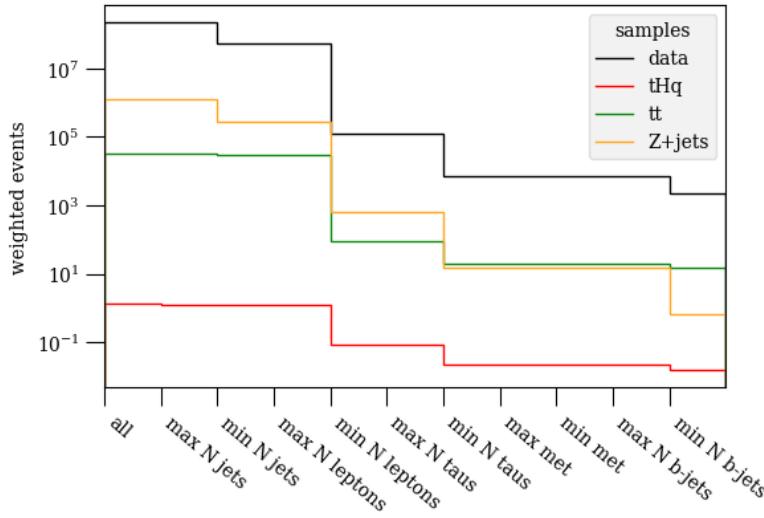


Figure 6.15: Cutflow on the PR conditions for the data tHq , $t\bar{t}$ and $Z + \text{jets}$ samples in the $2\ell + 1\tau_{\text{had}}$ channel. The bin with *All* refers to the number of events before applying any PR requirement.

For the $\text{BDT}(tHq|_{\text{OS}})$ and $\text{BDT}(t\bar{t}|_{\text{OS}})$, only the positive weights are employed. This approach is validated by testing that the shape of the distributions using only positively-weighted events have the same shape as when using all the events. For the $\text{BDT}(tHq|_{\text{SS}})$, the absolute weight of the events is used in the training. This choice is favoured by the scarcity of events in the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel. Discarding negatively-weighted events would significantly reduce the size of the training sample, preventing the $\text{BDT}(tHq|_{\text{SS}})$ from learning to effectively separate signal and background.

Nevertheless, alternative weight-handling strategies are tested and it is observed that the performance of the models is worse when different approaches are different to the ones used in the BDTs presented in this section.

6.6.2.2 Discriminant variables

To enhance the capability of the BDT to discriminate between processes, several variables are built. Some of these kinematic variables are useful to improve the signal to background separation. As it is introduced in Section 6.4.2, it is observed that the variables that refer to the light leptons according to its origin provide better discrimination than the ones that use the p_T -based ordering. This is illustrated on Figure 6.16, where the distance of the light leptons to the leading b -jet is a feature that provides larger separation power when the e/μ is referred by its origin rather

Process	$2\ell + 1\tau_{\text{had}}$	$2\ell \text{ OS} + 1\tau_{\text{had}}$	$2\ell \text{ SS} + 1\tau_{\text{had}}$
tHq	3.2 ± 0.1	1.93 ± 0.11	1.22 ± 0.09
tWH	3.4 ± 1.1	2.4 ± 1.1	0.95 ± 0.22
$t\bar{t}$	5700 ± 160	5690 ± 160	24.8 ± 1.7
$Z + \text{jets}$	3800 ± 400	3800 ± 400	0.7 ± 0.4
$t\bar{t}H$	86 ± 13	61 ± 13	25.2 ± 1.3
$t\bar{t}W$	134 ± 25	92 ± 23	42 ± 9
$t\bar{t}Z$	160 ± 26	139 ± 26	21.2 ± 1.0
tWZ	22 ± 10	19 ± 10	2.6 ± 1.7
tZq	45 ± 6	40 ± 6	5.4 ± 0.9
tW	260 ± 40	260 ± 40	1.2 ± 1.0
Diboson	190 ± 130	180 ± 130	10 ± 6
minor bkgs	16 ± 8	14 ± 8	1.8 ± 1.0
Total background	10400 ± 500	10300 ± 500	137 ± 12
Data	10323	10221	102
S/B (%)	0.031	0.019	0.89
Significance	0.031	0.019	0.104

Table 6.20: Event yields at PR level for $2\ell + 1\tau_{\text{had}}$ channel and its two sub-channels. The uncertainty corresponds to both the statistical and systematic uncertainties.

than by its p_T . Therefore, the classification of the light leptons is substituted from leading (ℓ_1) and sub-leading (ℓ_2) leptons to ℓ_{top} and ℓ_{Higgs} .

An initial set of variables is proposed for each BDT based on their separation power. In some particular cases, variables that provide good classification accuracy to the BDT are not included in the initial set because they bias the separation power towards a particular process. This is the case of the E_T^{miss} in the $\text{BDT}(tHq|_{\text{OS}})$. This variable (see Figure 6.19b) offered such a good separation between $Z + \text{jets}$ and tHq that the model would become so specialised in discriminating these two processes that it would negatively affect the separation between the tHq and $t\bar{t}$ processes.

After selecting the preliminary set of features, it is optimised using an iterative method focusing on the impact of the variables over the BDT performance as well as their correlations with other variables. The iterative algorithm used to rank the variables is detailed in Appendix B.2. Through this algorithm, the information of the effect of removing each initial variable from the model training is obtained and, hence, a decision can be made on which variables are kept. The rule of thumb is to keep only the features that would improve the performance of the BDT. The

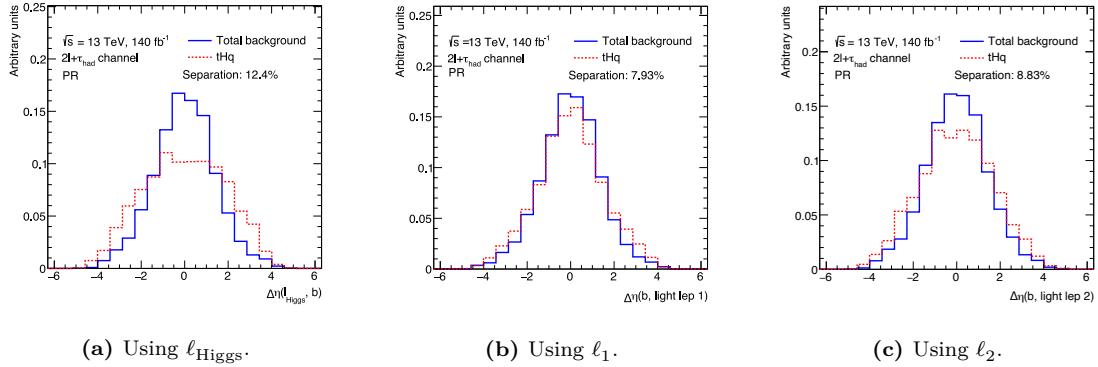


Figure 6.16: Separation plots for the distance in azimuthal angle between the a light lepton and the leading b -tagged jet. Note that higher discrimination is achieved when the light lepton is identified by its origin (a) rather than when they are classified based on its p_T ordering (b, c).

study of the correlations among the input features is presented in Appendix B.7.3. When a pair of variables presents a high correlation, one of them is removed from the model. The list of all variables used is presented in Table 6.21. The ranking of these variables is presented in Figure 6.17. In this figure can be checked which variables are used most times to generate splits in each BDT.

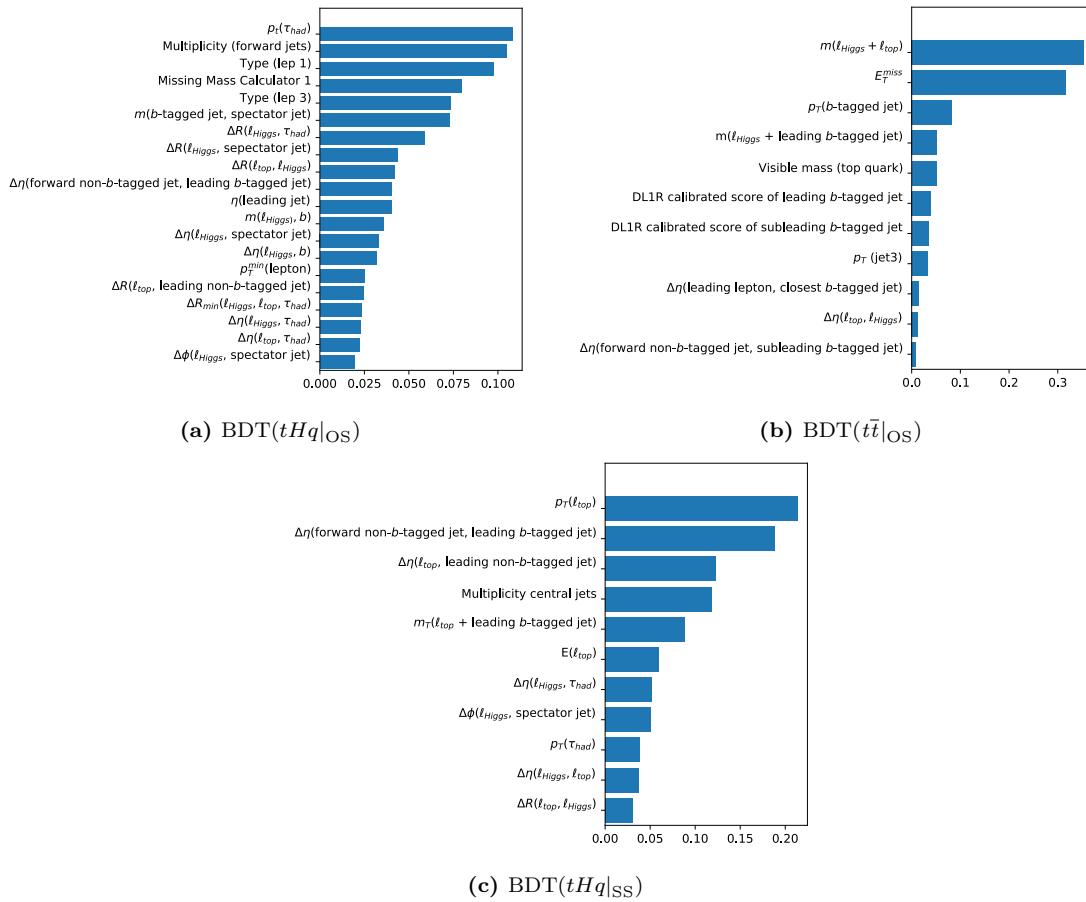


Figure 6.17: Rankings of input variables for each region-definition BDT. The rankings are obtained with the feature importance tool of the XGBoost. The x-axes correspond to the values given by XGBoost to evaluate the accuracy brought by the variable to the BDT.

Variable name	BDT($tHq _{\text{OS}}$)	BDT($t\bar{t} _{\text{OS}}$)	BDT($tHq _{\text{SS}}$)
$\Delta\eta(\ell_{\text{Higgs}}, \tau_{\text{had}})$	✓	-	✓
$\Delta R(\ell_{\text{top}}, \text{leading non-}b\text{-tagged jet})$	✓	-	-
$\Delta\eta(\text{forward non-}b\text{-tagged jet, leading }b\text{-tagged jet})$	✓	-	✓
Visible mass (top quark)	-	✓	-
p_{T} (b -tagged jet)	-	✓	-
$m(\ell_{\text{Higgs}}, b\text{-tagged jet})$	✓	✓	-
Missing Mass Calculator 1	✓	-	-
$p_{\text{T}}(\ell_{\text{top}})$	-	-	✓
$E(\ell_{\text{top}})$	-	-	✓
$\Delta R_{\min}(\ell_{\text{Higgs}}, \ell_{\text{top}}, \tau_{\text{had}})$	✓	-	-
$m(b\text{-tagged jet, spectator jet})$	✓	-	-
$\Delta\eta(\text{forward non-}b\text{-tagged jet, subleading }b\text{-tagged jet})$	-	✓	-
DL1R calibrated score of subleading b -tagged jet	-	✓	-
DL1R calibrated score of leading b -tagged jet	-	✓	-
$\Delta\phi(\ell_{\text{Higgs}}, \text{spectator jet})$	✓	-	✓
$\Delta R(\ell_{\text{top}}, \ell_{\text{Higgs}})$	✓	-	✓
Multiplicity (forward jets)	✓	-	-
$\Delta\eta(\text{leading lepton, closest }b\text{-tagged jet})$	-	✓	-
$\Delta\eta(\ell_{\text{top}}, \ell_{\text{Higgs}})$	-	✓	✓
$\Delta R(\ell_{\text{Higgs}}, \tau_{\text{had}})$	✓	-	-
$p_{\text{T}}^{\min}(\text{lepton}) \parallel p_{\text{T}}$ of the softest lepton	✓	-	-
$\Delta R(\ell_{\text{Higgs}}, \text{spectator jet})$	✓	-	-
Type (ℓ_1)	✓	-	-
Type (ℓ_3)	✓	-	-
$m_{\text{T}}(\ell_{\text{top}} + \text{leading }b\text{-tagged jet})$	-	-	✓
$\Delta\eta(\ell_{\text{top}}, \text{leading non-}b\text{-tagged jet})$	-	-	✓
$\Delta\eta(\ell_{\text{Higgs}}, \text{leading }b\text{-tagged jet})$	✓	-	-
$m(\ell_{\text{Higgs}} + \ell_{\text{top}})$	-	✓	-
$p_{\text{T}}(\tau_{\text{had}})$	✓	-	✓
$p_{\text{T}}(\text{jet 3})$	-	✓	-
$\eta(\text{leading jet})$	✓	-	-
$\Delta\eta(\ell_{\text{Higgs}}, \text{spectator jet})$	✓	-	-
$E_{\text{T}}^{\text{miss}}$	-	✓	-
$\Delta\eta(\ell_{\text{top}}, \tau_{\text{had}})$	✓	-	-
Multiplicity central jets	-	-	✓

Table 6.21: List of variables included in the training of the BDTs for region definition. The ✓ symbol indicates that the variable is used for that BDT model. Whenever a b -tagged jet is mentioned, unless stated otherwise, it refers to the leading- b -tagged jet.

As it is indicated in Figure 6.17a, the two most important variables for the $\text{BDT}(tHq|_{\text{OS}})$ are the p_T of the τ_{had} and the multiplicity of jets with $2.5 < |\eta| < 4.5$. Since the reconstructed τ_{had} is more boosted when it is generated from the Higgs-boson decay than in the main backgrounds (see Figure 6.18a), it produces good separation. In the tHq process, the jet initiated by the spectator quark is typically directed towards the beam pipe. This happens because a light quark from one proton interacts with a bottom quark from the other proton through the exchange of a virtual W boson. The light quark, now a spectator quark, is emitted in the forward direction, meaning it continues in roughly the same direction as the initial quark. As a result, the multiplicity of forward jets peaks at one for the tHq production (see Figure 6.18b). In contrast, the backgrounds tend to not present any forward jet. The separation distributions for these two variables are shown in Figure C.7.

There are variables such as E_T^{miss} (Figure 6.19b) that are not used in the $\text{BDT}(tHq|_{\text{OS}})$ despite providing a great separation power between the tHq and the general background. The reason to do so is that this variable is only separating tHq from $Z + \text{jets}$ but not from $t\bar{t}$, and it dominates the training process. As a result, if the E_T^{miss} is used, the model would not be able to properly separate $t\bar{t}$ from tHq .

Regarding the $\text{BDT}(t\bar{t}|_{\text{OS}})$, the invariant mass of the two light leptons (see Figure 6.19a) is the distribution that performs better to separate $t\bar{t}$ from $Z + \text{jets}$. This due to the peak in the value of the Z -boson mass, where most $Z + \text{jets}$ events are located. As can be seen in Figure 6.17b, the second highly ranked variable is the E_T^{miss} (see Figure 6.19b), in which $Z + \text{jets}$ processes are separated from $t\bar{t}$ because the latter has more neutrinos in the final state and, hence, larger E_T^{miss} .

Finally, according to Figure 6.17c, the $\text{BDT}(tHq|_{\text{SS}})$ exploits the transverse momentum of the light lepton originated from the top-quark decay to separate the signal (see Figure 6.20a). Although variables such as the $\delta\eta$ distance between the ℓ_{top} and the leading-light jet (see Figure C.37 in Appendix C) appear to be more discriminant, the $p_T(\ell_{\text{top}})$ is used more times by XGBoost to make splits in the BDT. With a similar importance to $p_T(\ell_{\text{top}})$, the $\Delta\eta$ distance between the forward and b -tagged jets provides good separation between signal a background. These two jets tend to be more geometrically distant in the signal than for the backgrounds. Also, in the main backgrounds of this channel there are no forward jets originated in the hard-scattering process.

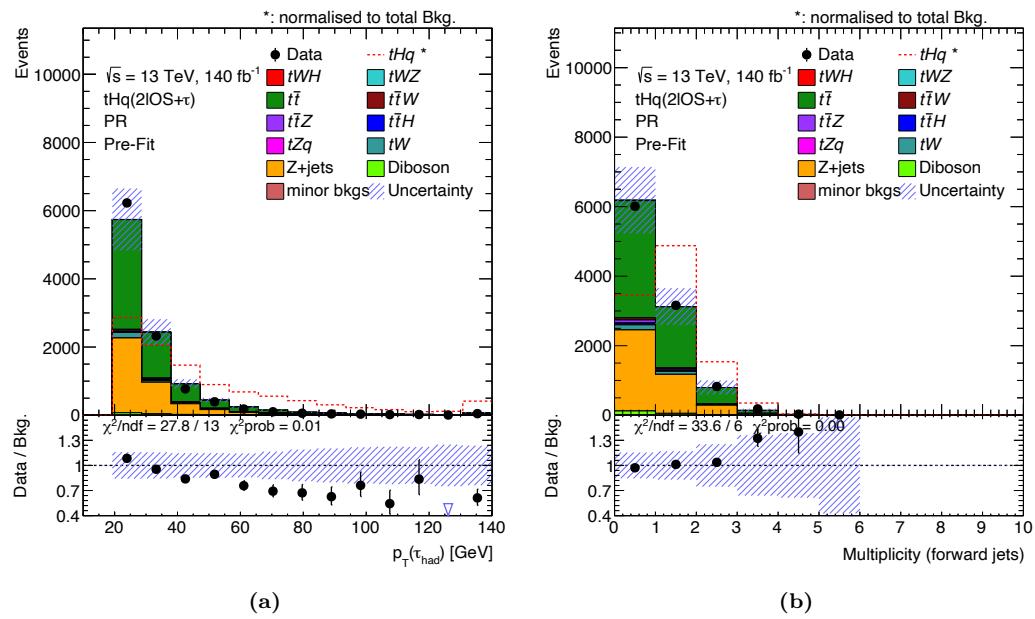


Figure 6.18: The two input variables with highest separation power for the BDT($tHq|os$). These are the (a) p_T of the τ_{had} and (b) the multiplicity of forward jets. The uncertainty bands include the statistical and systematic uncertainties and the lower panel presents the ratio between the collected data and the MC-simulated samples. The discrepancy between data and MC is due to the misidentification factors discussed in Section 6.5. Additionally, the χ^2 measures the agreement between real data and MC-simulated-event samples.

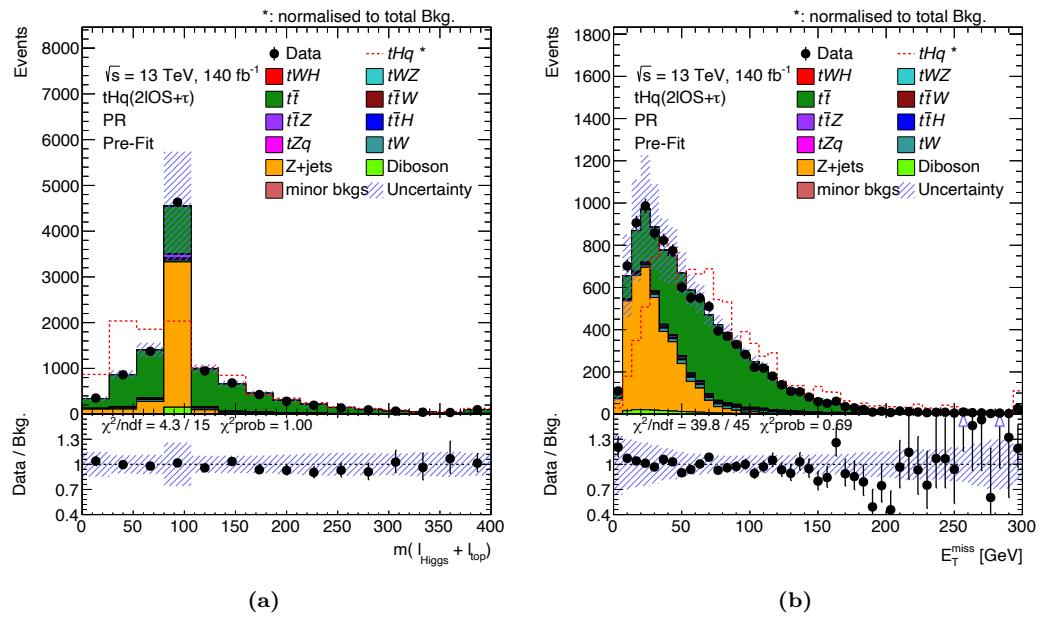


Figure 6.19: The two input variables with the highest separation power for the $\text{BDT}(t\bar{t}|_{\text{OS}})$. There are the invariant mass of the two light leptons (a) and the E_T^{miss} (b). The uncertainty bands include the statistical and systematic uncertainties and the lower panel presents the ratio between the collected data and the MC simulation. Additionally, the χ^2 measures the agreement between real data and MC-simulated-event samples.

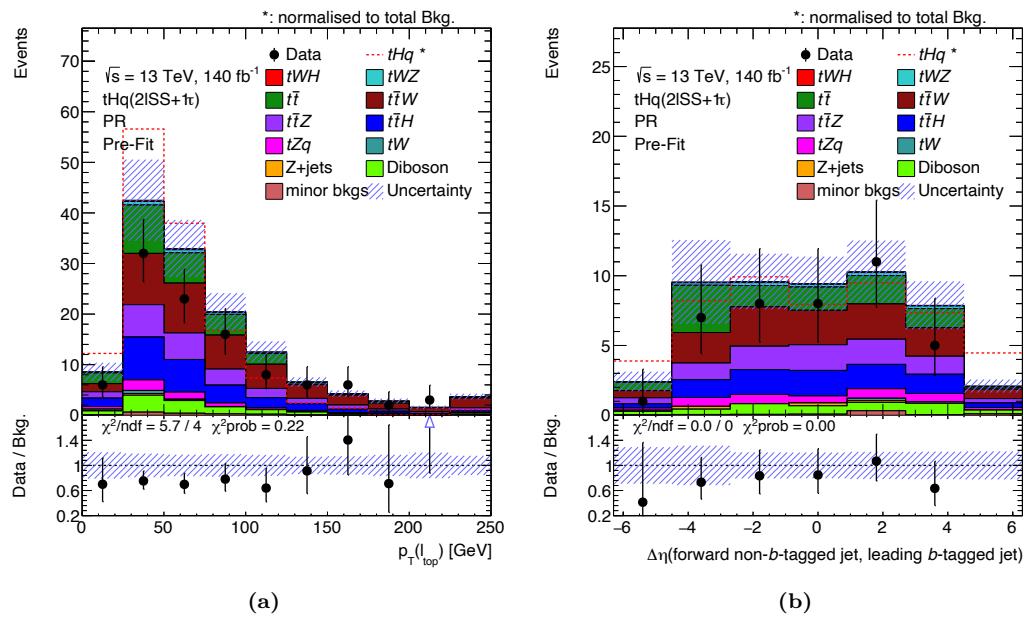


Figure 6.20: The two input variables with the highest separation power for the $\text{BDT}(tHq|\text{ss})$. The distribution for $p_T(l_{\text{top}})$ is presented in (a) and while (b) shows the distribution for the $\Delta\eta$ between the forward non- b -tagged jet and the leading b -tagged jet. The uncertainty bands include the statistical and systematic uncertainties and the lower panel presents the ratio between the collected data and the MC simulation. Additionally, the χ^2 measures the agreement between real data and MC-simulated-event sample.

6.6.2.3 Hyperparameter optimisation

The hyperparameters are the parameters that define the structural aspects of the ML model and the training process itself. The hyperparameters are not learned from the data during the training but are set prior to this training process. An in-depth discussion about the hyperparameters of the employed BDTs is given in Appendix B.5.

Selecting appropriate hyperparameters is crucial as it can significantly influence the performance of the model. The optimisation of the hyperparameters of the BDT for region definition is done using a genetic algorithm (GA) [294], which is described in detail in the Appendix B.5.1. Alternatively, a grid search has also been tested to find the optimal set of hyperparameters but the performance of the GA surpassed that of the grid search. The optimal set of hyperparameters found is presented for all three BDTs simultaneously in Table 6.22.

Hyperparameter	BDT($tHq \text{os}$)	BDT($t\bar{t} \text{os}$)	BDT($tHq \text{ss}$)
Maximum depth	4	4	4
Learning rate	0.1237	0.0334	0.04
Number of estimators	1500	1500	1500
Minimum child weight	0.52	0.077	0.026
Scale of positive weights	268.838	0.36	83.21
Neg. weight strategy	Only positive	Only positive	Absolute values

Table 6.22: Configuration of the hyperparameters used to manage the training of the three gradient BDTs employed for region definition. The rest of the hyperparameters are set to their default values. In Appendix B.5, these and the rest of hyperparameters are discussed.

6.6.2.4 Performance

The performance of the BDTs is evaluated through three main figures of merit, the ROC curve, the AUC and the logarithmic loss function (LogLoss). The ROC curves of the three region-definition BDTs are presented in Figure 6.21. The ROC is a graphical representation plotting the True Positive Rate against the False Positive Rate at various thresholds. Its AUC evaluates the overall separation power of the binary classification model. The third metric, the LogLoss, in contrast to AUC, measures the accuracy of a classifier by penalising the confidence of incorrect predictions. These metrics are described in Appendix B. Table 6.23 presents the AUC and the LogLoss for the BDTs for region definition. Since the k -folding method for cross-validation has also been used here, Table 6.23 reflects the mentioned metrics for the five folds separately. In that table can be seen that the performance does

not vary across the folds, meaning that the models are not overtrained. This ability of generalisation across folds can also be observed in Figure 6.22, where the BDT score for each fold is presented and compared to the rest of the folds. As can be seen in Figure 6.22, there is complete compatibility in three trained BDTs across their folds. For the BDT($tHq|_{OS}$) in Figure 6.22a most of the events are in the left-most part of the distribution. This is because the $t\bar{t}$ and $Z + \text{jets}$ events are grouped on the left side while the tHq process is uniform across the BDT score. In the case of the BDT($t\bar{t}|_{OS}$), the two peaks in Figure 6.22b correspond to the peaks of $t\bar{t}$ (right peak) and $Z + \text{jets}$ (left peak). Finally, for the BDT($tHq|_{SS}$) in Figure 6.22c, most of the events are centred in the distribution but the tHq signal is peaking on the left part.

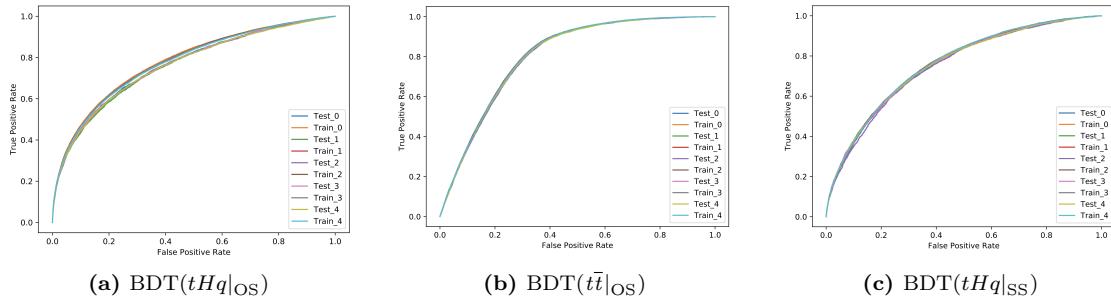


Figure 6.21: ROC for the all folds of the three BDT models. For each figure five pairs of ROC curves are visible. Each pair corresponds to a fold and, within a fold, the pair addresses the train and test samples separately.

Fold	BDT($tHq _{OS}$)		BDT($t\bar{t} _{OS}$)		BDT($tHq _{SS}$)	
	AUC	LogLoss	AUC	LogLoss	AUC	LogLoss
0	0.655	0.1960	0.7439	0.4378	0.6685	0.306
1	0.655	0.1998	0.7431	0.4381	0.6702	0.307
2	0.657	0.1978	0.7446	0.4374	0.6705	0.309
3	0.658	0.1955	0.7461	0.4374	0.6717	0.311
4	0.662	0.1973	0.7477	0.4391	0.6723	0.311
Mean	0.657 ± 0.003	0.1972 ± 0.0015	0.7452 ± 0.0015	0.4379 ± 0.0006	0.6706 ± 0.0013	0.308 ± 0.002

Table 6.23: AUC and LogLoss for the five folds of the BDTs used for region definition. In the last row, the mean is presented with the corresponding standard deviation. The average is the metric used as overall for the model.

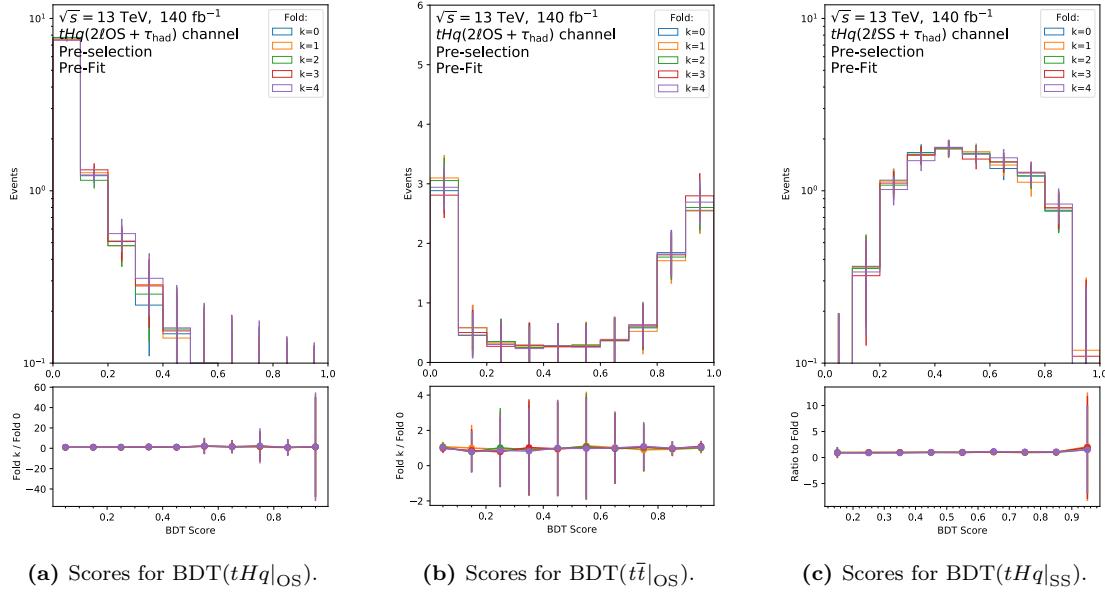


Figure 6.22: In the upper panel of each figure, the profile of the BDT is plotted for all five folds simultaneously. The uncertainty in each bin of the upper panel corresponds to the weighted standard deviation. In the lower panel, the ratio of event yields between the first fold (Fold 0) and all the other is presented. Both panels share the same x-axis. For all folds within a BDT, the distributions are compatible within the statistical uncertainty. This indicates that the model generalises well and it is not affected by the specific set of events that are used for train in each fold. The profiles are created using the test samples only. Note the logarithmic scale on (a) and (c).

6.6.3 Signal Region

More and more stringent requirements are made to eliminate background events, signal events are also lost. Therefore, there is a trade-off background rejection against signal acceptance. Since the tHq data is scarce, the event selection is a highly non-trivial process that requires attention. The region definition is not done only by looking at the S/B ratio and the significance, but also taking into account the behaviour of the fit (described in Section 6.8). The final conditions defining the regions of the phase space used in this analysis are described through Sections 6.6.3.1 and 6.6.3.2 as well as Section 6.6.4.

6.6.3.1 SR in the $2\ell \text{ OS} + 1\tau_{\text{had}}$

Achieving a high signal purity region in the $2\ell \text{ OS} + 1\tau_{\text{had}}$ channel is challenging due to the S/B ratio being only 0.018% at PR level. The $\text{BDT}(tHq|_{\text{OS}})$, as presented in Section 6.6.2, enhances both the S/B ratio and significance by setting

a minimum score threshold. Figure 6.23 illustrates the progression of these metrics and the remaining tHq events relative to the minimum $\text{BDT}(tHq|_{\text{OS}})$. This figure shows that a higher $\text{BDT}(tHq|_{\text{OS}})$ cut increases the S/B ratio and significance, peaking at approximately 0.8. However, aggressive cuts in the BDT score can lead to a shortage of events in the SR, causing instability in the fit. Consequently, a more conservative cut of $\text{BDT}(tHq|_{\text{OS}}) \geq 0.3$ is employed to define the SR in the $2\ell \text{OS} + 1\tau_{\text{had}}$ channel. The distribution of the $\text{BDT}(tHq|_{\text{OS}})$ discriminant is presented in Figure 6.24.

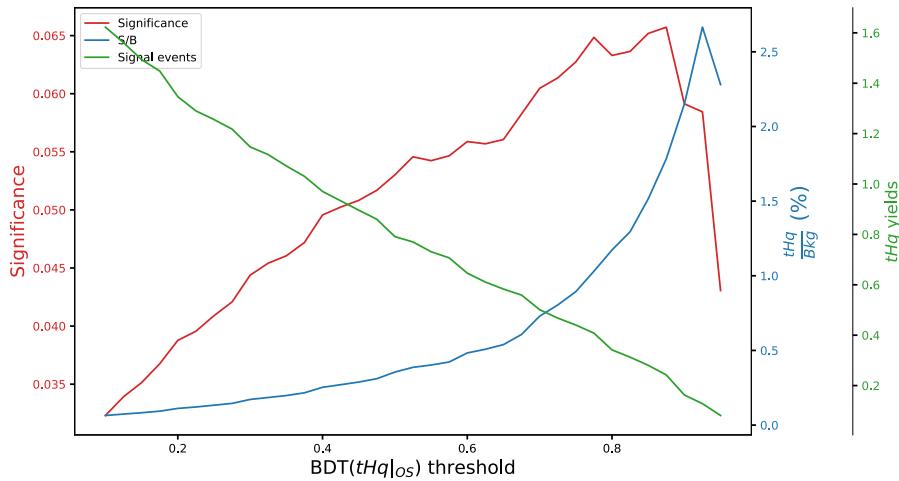


Figure 6.23: Significance (red), S/B ratio (blue), and number of tHq events (green) as function on the lower cut on $\text{BDT}(tHq|_{\text{OS}})$.

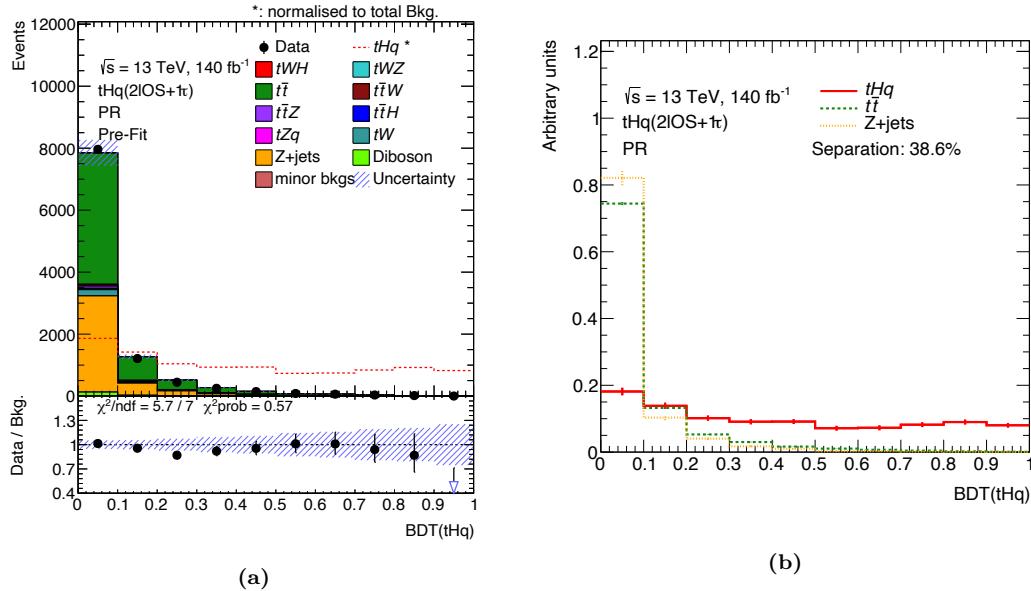


Figure 6.24: Distribution of the discriminant $\text{BDT}(tHq|_{\text{OS}})$ at PR in the $2\ell \text{ OS} + 1\tau_{\text{had}}$ channel. The dotted-red line in (a) shows the normalised tHq signal. The separation plot (b) presents the normalised distributions of the tHq signal and the two main backgrounds. While the tHq process has a relatively uniform distribution, the background processes peak at leftmost part. This allows to cleanse the vast majority of the background while conserving most of the signal events.

6.6.3.2 SR $2\ell \text{ SS} + 1\tau_{\text{had}}$

When considering $2\ell \text{ SS} + 1\tau_{\text{had}}$ events, all background contributions are significantly reduced compared to the $2\ell \text{ OS} + 1\tau_{\text{had}}$ case. Nevertheless, the dominant background remains $t\bar{t}$, followed by the $t\bar{t}X$ processes, as indicated in Table 6.20. This table presents the event yields of the MC simulated data samples after applying the preselection requirements (see Section 6.6.1).

The signal selection for the $2\ell \text{ SS} + 1\tau_{\text{had}}$ channel is accomplished cutting on the BDT presented in Figure 6.25. The evolution of the S/B ratio, significance and number of remaining tHq events according to the lower cut on $\text{BDT}(tHq|_{\text{SS}})$ is presented in Figure 6.26. This figure also suggests a strict cut in the BDT in order to define a pure SR. However, to reduce the instabilities in the fit due the small statistics in the SR, the final threshold is looser; $\text{BDT}(tHq|_{\text{SS}}) \geq 0.40$.

6.6.4 Background dedicated regions

The background-enriched regions are used to control the behaviour of the background processes. These can either be CR or VR, and its role on the fit calculations

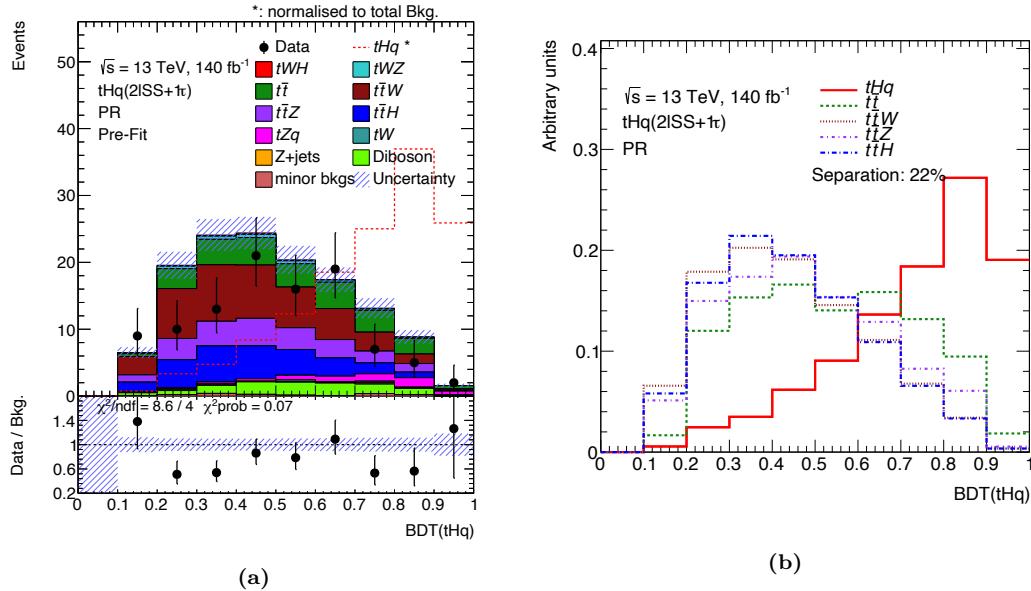


Figure 6.25: Distribution (a) and separation plot (b) of the discriminant $\text{BDT}(tHq|_{\text{SS}})$ at PR in the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel. The dotted-red line in (a) shows the normalised tHq signal. The separation plot presents the normalised distributions of the tHq signal and the main background processes. Here it can be appreciated that the all $t\bar{t}X$ processes ($t\bar{t}W$, $t\bar{t}H$ and $t\bar{t}Z$) have the same profile, being very difficult to separate from each other.

is explained in Section 6.8.1. Sections 6.6.4.1 and 6.6.4.2 explore the details of the definition of these regions.

6.6.4.1 Background-enriched regions in the $2\ell \text{OS} + 1\tau_{\text{had}}$ channel

Two background dedicated regions are defined in the $2\ell \text{OS} + 1\tau_{\text{had}}$ channel: one for $t\bar{t}$ and another for $Z + \text{jets}$. In this case, both regions are considered as CRs for the fit. The distribution of the $\text{BDT}(t\bar{t}|_{\text{OS}})$ is presented at PR in Figure 6.28.

The conditions that define the SR, CR($t\bar{t}$), and CR($Z + \text{jets}$) in the $2\ell \text{OS} + 1\tau_{\text{had}}$ analysis are summarised in Table 6.25. The event yields per process corresponding to these regions of the phase space are presented in Table 6.25.

6.6.4.2 Background-enriched regions in the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel

The primary backgrounds in the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel consist of the $t\bar{t}$, $t\bar{t}W$, $t\bar{t}Z$, and $t\bar{t}H$ processes. Ideally, one would define CRs specifically tailored for each of these processes. However, due to limited statistics training ML-based methods to target these processes separately becomes exceedingly challenging.

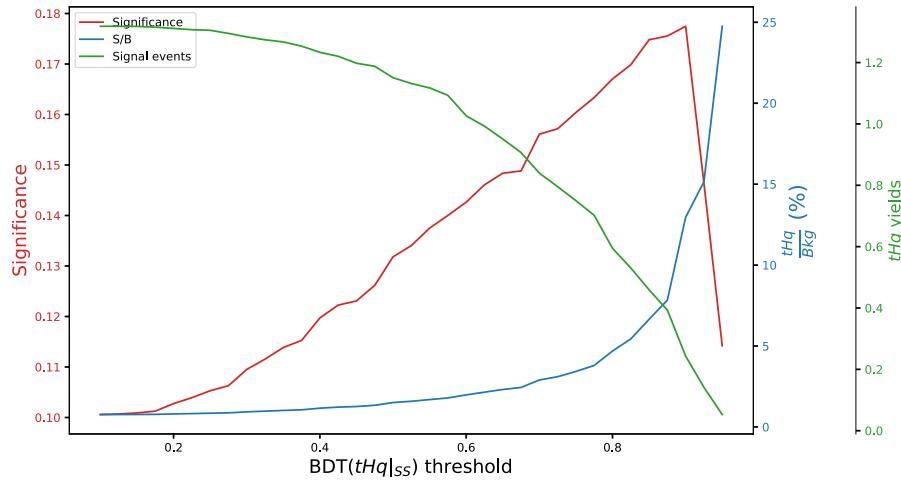


Figure 6.26: Significance (red) and S/B ratio (blue), and tHq event yields (green) as function on the cut on $\text{BDT}(tHq|_{\text{ss}})$.

Region	BDT score	Ambiguity cut
SR	$\text{BDT}(tHq _{\text{os}}) \geq 0.40$	yes
CR($t\bar{t}$)	$\text{BDT}(tHq _{\text{os}}) < 0.40$ $\text{BDT}(t\bar{t} _{\text{os}}) \geq 0.50$	yes
CR($Z + \text{jets}$)	$\text{BDT}(tHq _{\text{os}}) < 0.40$ $\text{BDT}(t\bar{t} _{\text{os}}) < 0.50$	yes

Table 6.24: Definition of the analysis regions in the $2\ell \text{ OS} + 1\tau_{\text{had}}$ channel. The ambiguity cuts refer to the ECIDS and to the [Electron Ambiguity Tool](#).

As an alternative approach, a feasible option is to combine the $t\bar{t}W$, $t\bar{t}Z$, and $t\bar{t}H$ productions into a unified category termed $t\bar{t}X$. This consolidation enables the separation of $t\bar{t}$ and $t\bar{t}X$ in the phase space region orthogonal to the SR. Although BDTs targeting $t\bar{t}$ and $t\bar{t}X$ have been trained, the scarcity of statistics remained an obstacle to achieving favourable outcomes.

An alternative to the use of BDTs is performing requirements (cuts) on kinematic variables. To implement this approach, the initial step entails exploring all $2\ell \text{ SS} + 1\tau_{\text{had}}$ distributions within the PR\SR. Subsequently, variables that demonstrate effective $t\bar{t}$ discrimination can be identified and selected for further analysis. In this context, the variable H_T emerges as a potent discriminator between $t\bar{t}$ and $t\bar{t}X$. Using H_T it is possible to define a CR for $t\bar{t}$ and another for the $t\bar{t}X$ processes using the conditions:

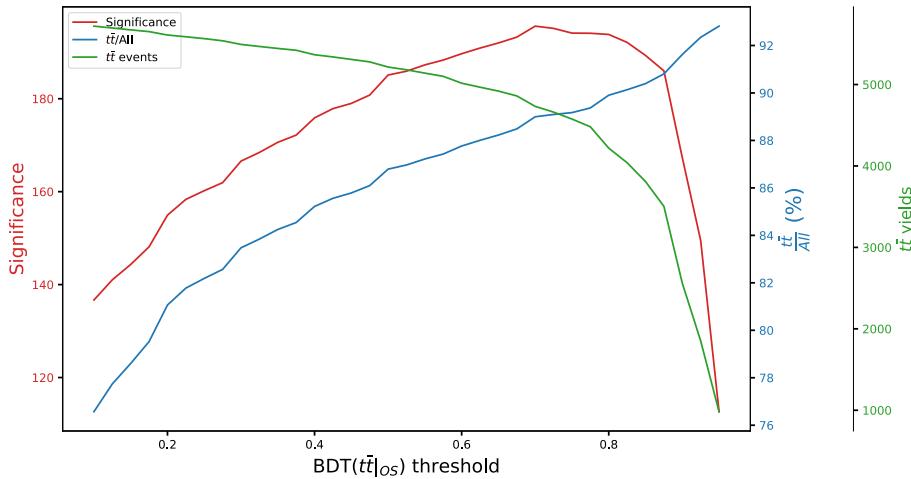


Figure 6.27: Significance (red) and $t\bar{t}$ purity (blue), and $t\bar{t}$ event yields (green) as function on the cut on $\text{BDT}(t\bar{t}|_{\text{OS}})$.

- $\text{CR}(t\bar{t}) = \text{BDT}(tHq|_{\text{OS}}) \geq 0.40 \text{ and } H_T \geq 260 \text{ GeV}$
- $\text{CR}(t\bar{t}X) = \text{BDT}(tHq|_{\text{OS}}) \geq 0.40 \text{ and } H_T > 260 \text{ GeV}.$

However, due to the lack of statistics in the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel, this analysis makes use of a single CR that groups all the backgrounds. This is referred to as $\text{CR}(\text{All Bkg})$. Studies are carried to evaluate the impact of performing the profile-likelihood fit with $\text{CR}(\text{All Bkg})$ on one side, and with $\text{CR}(t\bar{t})$ and $\text{CR}(t\bar{t}X)$ on the other side. The results with the single-CR method are more robust.

The region definition for the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel is summarised in Table 6.26 and its yields are given in Table 6.27.

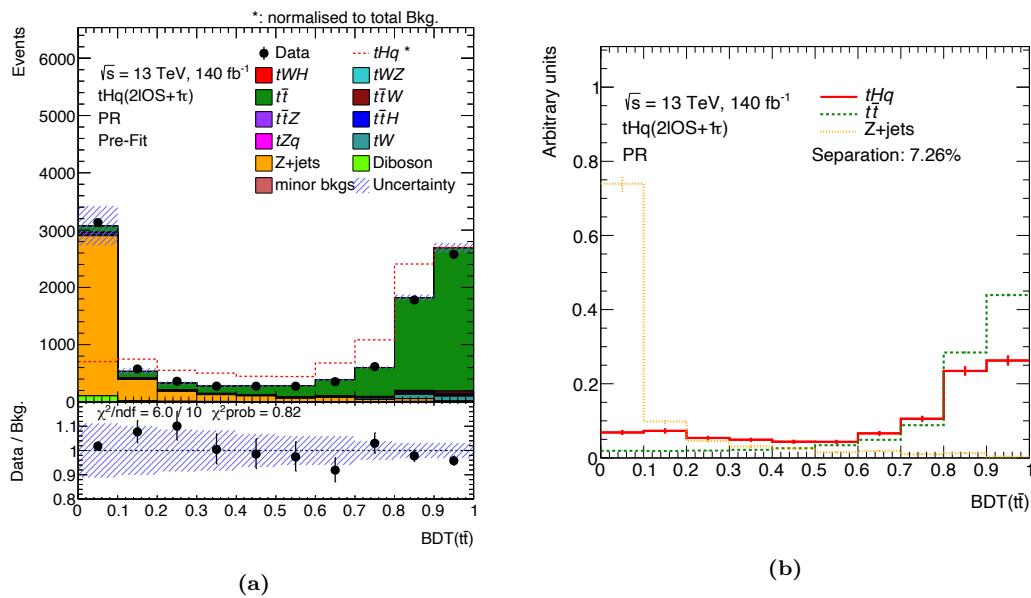


Figure 6.28: Distribution of the discriminant $\text{BDT}(t\bar{t}|_{\text{OS}})$ at PR. The dotted-red line in (a) shows the normalised tHq signal, which follows a similar distribution as $t\bar{t}$. The separation plot in (b) presents the normalised distributions for the tHq , $t\bar{t}$ and $Z + \text{jets}$ processes.

Process	SR	$\text{CR}(t\bar{t})$	$\text{CR}(Z + \text{jets})$
tHq	0.94 ± 0.15	0.7 ± 0.44	0.31 ± 0.16
tWH	0.40 ± 0.13	1.6 ± 0.9	0.39 ± 0.22
$t\bar{t}$	230 ± 40	4880 ± 140	580 ± 50
$Z + \text{jets}$	68 ± 18	220 ± 40	3500 ± 400
$t\bar{t}H$	9.0 ± 2.0	45 ± 9	7.1 ± 1.7
$t\bar{t}W$	5.5 ± 1.5	77 ± 19	9.8 ± 2.4
$t\bar{t}Z$	15.4 ± 3.0	56 ± 11	67 ± 15
tWZ	2.1 ± 1.2	5.8 ± 3.4	11 ± 6
tZq	7.4 ± 1.3	8.5 ± 1.3	24 ± 44
tW	9.3 ± 1.9	202 ± 30	46 ± 7
Diboson	12 ± 7	26 ± 15	140 ± 120
minor bkg	0.9 ± 1.8	6 ± 4	7 ± 4
Total background	360 ± 50	5530 ± 170	4400 ± 400
Data	347	5370	4504
S/B or purity (%)	0.261	88.5	79.7
Significance	0.050	118.0	83.5

Table 6.25: Pre-fit yields of the $2\ell \text{OS} + 1\tau_{\text{had}}$ channel in SR, $\text{CR}(t\bar{t})$ and $\text{CR}(Z + \text{jets})$. The uncertainties include the statistical uncertainties as well as all systematic uncertainties. For the SR the S/B ratio is presented. For the $\text{CR}(t\bar{t})$ and $\text{CR}(Z + \text{jets})$ present, respectively, the fraction of $t\bar{t}$ and $Z + \text{jets}$ events with respect the total MC event sample. The significances correspond the target process of each region.

Region	BDT score	Ambiguity cut
SR	$\text{BDT}(tHq) \geq 0.40$	Yes
CR(All Bkg)	$\text{BDT}(tHq) < 0.40$	Yes

Table 6.26: Definition of the three regions of the phase space used in the fit of the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel. The ambiguity cuts refer to the ECIDS and to the Electron Ambiguity Tool.

Process	SR	CR(All Bkg)
tHq	1.1 ± 0.4	0.08 ± 0.07
tWH	0.60 ± 0.25	0.35 ± 0.21
$t\bar{t}$	17.6 ± 2.7	7.2 ± 2.9
$Z + \text{jets}$	0.46 ± 0.25	0.27 ± 0.10
$t\bar{t}H$	14.1 ± 0.8	11.1 ± 0.9
$t\bar{t}W$	23 ± 5	19 ± 4
$t\bar{t}Z$	13.2 ± 0.7	7.9 ± 0.7
tWZ	1.6 ± 1.0	1.0 ± 0.6
tZq	4.9 ± 0.9	0.47 ± 0.13
tW	1.0 ± 0.8	0.30 ± 0.14
Diboson	8 ± 4	2.2 ± 1.4
minor bkg	1.1 ± 0.64	0.7 ± 0.4
Total background	87 ± 9	50 ± 6
Data	70	32
S/B or purity (%)	1.264	89.38
Significance	0.118	11.68

Table 6.27: Pre-fit yields of the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel in the SR and the CR(All Bkg). For the SR the S/B and significance of the tHq process is presented. For the CR(All Bkg) the purity and significance of the $t\bar{t}, t\bar{t}W, t\bar{t}H$ and $t\bar{t}Z$ together is shown.

6.7 Systematic uncertainties

In a physics analysis, uncertainties refer to the limitations and potential variations in the measurements and calculations used to obtain results. These uncertainties provide a measure of the range within which the true value of a measured quantity is expected to lie. There are two types of uncertainties: statistical and systematic.

Statistical uncertainty arises from the inherent randomness or fluctuations in the finite data. It is a consequence of the finite size of the analysed dataset. These fluctuations are typically described by statistical methods, such as probability distributions, and are quantified by statistical measures like standard deviation or confidence intervals. Statistical uncertainties are fully uncorrelated between subsequent measurements, meaning that each measurement carries its own independent statistical uncertainty. This uncertainty applies to both the real data collected by the detector and the MC-simulated data events.

Systematic uncertainties encompass all sources of error or variation that are not directly due to the statistics of the data. Systematic uncertainties can occur at any point in the analysis chain. They are associated with various factors, including the measurement apparatus, experimental conditions, assumptions made in this analysis, theoretical models employed, object reconstruction, background-estimation techniques and MC simulations and many others. Systematic uncertainties are fully correlated between subsequent measurements, meaning that they affect the entire dataset consistently.

Unlike statistical uncertainties, which are inherent in the data, systematic uncertainties are associated with the methodology and procedures used in this analysis. These uncertainties can have a significant impact on the final results and must be carefully evaluated and accounted for. Therefore, they require detailed investigations and studies to quantify their effects.

In a physics analysis, it is essential to include both statistical and systematic uncertainties in the overall uncertainty estimation. Statistical uncertainties are typically incorporated into the inference method used to extract results. However, the inclusion of systematic uncertainties and their propagation through the statistical analysis is a more complex task. It involves understanding the sources of systematic uncertainties, quantifying their magnitudes, and considering their correlations. Systematic uncertainties are usually evaluated through dedicated studies and variations in input parameters, alternative models, or comparison with independent measurements.

Properly accounting for uncertainties in a physics analysis is crucial for robust and reliable scientific conclusions. By quantifying and considering both statistical and systematic uncertainties, researchers can assess the precision, accuracy, and limitations of their measurements, compare their findings with theoretical predictions, and provide a comprehensive understanding of the physics processes under investigation.

For this specific analysis, the systematic uncertainties are categorised into two main groups: theoretical or modelling uncertainties and experimental uncertainties. Each group captures different sources of variation and potential biases in the measurement process. These systematic uncertainties are further discussed and presented throughout the rest of the section.

6.7.1 Symmetrisation of systematic uncertainties

The “up” and “down” variations refer to different directions in which a systematic uncertainty is varied. Sources of systematic uncertainties are often represented by a central value, which is the nominal value used in the analysis, and two variations: an “up” variation and a “down” variation. The former corresponds to an upward shift or increase in the systematic effect, while the latter variation corresponds to a downward shift or decrease.

For systematic uncertainties that have both the “up” and “down” variations, any asymmetries are preserved unless they are determined to be the result of statistical fluctuations in the underlying MC templates. In such cases, the content of each bin is adjusted to represent a variation of

$$\text{uncertainty} = \frac{|\text{up}| + |\text{down}|}{2}.$$

Then, the resulting uncertainty is symmetrised. In the case of systematic uncertainties where only a one-sided variation is provided, the variation is mirrored around the mean value in each bin. An example of one-sided systematics are the modelling uncertainties derived from an alternative generator. These symmetrisations allow for a clearer interpretation of the potential deviations or pulls (fitted $\theta \neq 0$) and constraints (fitted $\sigma_\theta \neq 1$), facilitating a more meaningful analysis of the results (see Section 6.8.1 for details about the θ).

6.7.2 Theoretical uncertainties

Theoretical or modelling uncertainties are inherent in the calculations and simulations used to predict physical observables in particle physics experiments. These

uncertainties arise from the approximations and assumptions made in theoretical models and the limitations of computational tools. Understanding and quantifying these uncertainties is crucial for interpreting experimental results and assessing the robustness of theoretical predictions.

In this section, the theoretical or modelling uncertainties that are considered in this analysis are discussed. These uncertainties arise from various sources, including the choice of MC generators, scale variations, PDFs, parton shower and hadronisation models, higher-order QCD corrections, and non-perturbative effects. Each source of uncertainty contributes to the overall theoretical uncertainty and must be carefully evaluated

To quantify these uncertainties, different theoretical predictions and variations in model parameters are compared to assess their impact on the results. This allows to estimate the accuracy and reliability of the theoretical calculations.

Modelling uncertainties are evaluated in three ways: comparing the nominal prediction to an alternative prediction; varying the internal parameters of the nominal simulation; or varying the predicted cross-section within the theoretical uncertainty.

In the rest of the section, I describe each source of modelling uncertainty, discussing the evaluation methods and their impact on the analysis results.

- **tHq modelling:** To assess the systematic uncertainty due to the choice of `MADGRAPH5_AMC@NLO 2.6.2. +PYTHIA` as nominal generator, alternative MC samples for the tHq signal events are produced using the `MADGRAPH5_AMC@NLO 2.8.1` generator at NLO with the `NNPDF3.0NLO_nf4` PDF set, and interfaced with `HERWIG 7.1.6` (instead of `PYTHIA`) using the `MMHT2014NNLO [173]` PDF set and the `HERWIG 7.1 [194, 199]` tune. The scales μ_R and μ_F are the same as for the nominal tHq event sample.
- **$t\bar{t}$ modelling:** The systematics regarding the modelling of the $t\bar{t}$ samples are the most impactful among the theoretical uncertainties.

For the ISR and FSR simulated with `POWHEG +PYTHIA`, new simulations are done from variations of several parameters. First, the μ_R and μ_F are independently adjusted by scaling them down by a factor of 0.5 and up by a factor of 2. Additionally, both μ_R and μ_F together are simultaneously varied by applying the same scaling factors of 0.5 and 2. The variations on the scale in the showering are done with the `Var3c` up or down variations of the A14 tune, which corresponds to the variation of α_s for ISR in the A14 tune [249]. These new simulations are then compared to the nominal ones. On the one hand, for the modelling of the ISR, four independent variations are defined.

- The first two variations consist of varying the μ_R and μ_F scales in the simulation of the hard-scattering process and the PS. These two parameters are varied independently by a factor 0.5 and 2 with respect to the nominal value in the simulation.
- The third variation is modifies the scale in the showering by varying the Var3c eigentune of the A14 tune [295].
- The final ISR variation is on the h_{damp} parameter of HERWIG. The nominal value of 1.5 times the mass of the top quark is varied up to $3 \times m_t$.

The uncertainty from the FSR simulation is calculated by using alternative event weights within the nominal $t\bar{t}$ event sample. These alternative weights are obtained by varying the FSR renormalisation scale (FSR α_s) for emissions from the PS. Specifically, it is varied by a factor of 0.5 (downward) and 2 (upward) compared to its nominal value.

An uncertainty is also attributed to the choice of the POWHEG approach to perform the matching between the hard-scatter and the PS. It is estimated comparing the POWHEG +HERWIG 7.1.3 prediction with the AMC@NLO +HERWIG 7.1.3 simulation.

The uncertainty associated with the choice of the hadronisation model and the other non-perturbative aspects of the PS are evaluated comparing the nominal sample with POWHEG +HERWIG 7.2.1.

The predicted $t\bar{t}$ cross-section is affected by the scale uncertainty, the PDF+ α_s uncertainty and the uncertainty on the top-quark mass. The uncertainties in the cross-section due to the PDF and α_s are calculated using the PDF4LHC15 prescription [170] with the MSTW2008NNLO [177, 296], CT10NNLO [297, 298] and NNPDF2.3LO PDF sets in the 5FS, and are added in quadrature to the effect of the scale uncertainty. Assuming a top quark mass of 172.5 GeV, the considered uncertainty is $985^{+23}_{-35}(\text{scale})^{+41}_{-41}(\text{PDF}+\alpha_s)^{+27}_{-26}(\text{mass})$ pb. The three uncertainty components of the $\sigma(t\bar{t})$ are considered uncorrelated. An overall $\pm 6\%$ is used to vary the $t\bar{t}$ cross-section.

- **$t\bar{t}H$ modelling:** The impact of the $t\bar{t}H$ modelling is estimated through the evaluation of ISR, FSR, PS and hadronisation model. For ISR and FSR, μ_R and μ_F are varied by factors of 0.5 and 2, including adjustments to the Var3c eigentune. The effects of hadronisation and other non-perturbative parton shower aspects are evaluated by comparing POWHEG +HERWIG 7 to the nominal model. Uncertainties from the technique to match the hard-scatter

and the PS are estimated by comparing the nominal POWHEG +PYTHIA 8 to AMC@NLO +PYTHIA 8 simulations.

Finally, regarding the cross-section, the uncertainty is $^{+5.8\%}_{-9.2\%}$ (scale) + 3.6%(PDF+ α_s). The two uncertainty components of the $\sigma(t\bar{t}H)$ are considered uncorrelated.

- **Single top quark modelling:** To assess the uncertainty due to the generator choice, MADGRAPH5_AMC@NLO 2.6.2 is used for producing alternative samples.

The approach employed for assessing the ISR and FSR modelling for the three single-top-quark processes (t -channel, tW -channel and s -channel) is the same as in $t\bar{t}H$. The Var3c parameters, μ_R , and μ_F are varied. The impact on the hadronisation model is obtained by comparing the nominal to the POWHEG +HERWIG 7.1.6. For the matching method, the nominal is compared with AMC@NLO +PYTHIA8 for the tW -channel, and POWHEG +HERWIG 7.1.6 with AMC@NLO +HERWIG 7.1.6 for tW -channel and s -channel.

Finally, a 5% uncertainty on the theoretical cross-section of single top-quark is considered in this analysis.

- **$t\bar{t}W$ and $t\bar{t}Z$ modelling:** The effects of varying the \mathcal{M} , PS and hadronisation models are assessed through comparison between the primary simulation samples and alternative generator-produced samples.

To evaluate the impact of the choice of generator in for $t\bar{t}W$, the nominal (SHERPA 2.2.10) is compared to MADGRAPH5_AMC@NLO 2.3.3 and a uncertainty on the cross-section of $^{+12.9\%}_{-11.5\%}$ (scale) + 3.4%(PDF+ α_s) is obtained.

For $t\bar{t}Z$ the nominal simulation (MADGRAPH5_AMC@NLO 2.3.3) is compared to the SHERPA 2.2.0 LO prediction to derive the modelling uncertainty. The predicted $t\bar{t}Z$ cross-section uncertainty is $^{+9.6\%}_{-11.3\%}$ (scale) + 4%(PDF+ α_s). The two uncertainty components of the $\sigma(t\bar{t}Z)$ are considered uncorrelated.

- **$Z + \text{jets}$ modelling:** When it comes to modelling of $Z + \text{jets}$, only the theoretical uncertainty on the predicted cross-section is considered. The considered normalisation uncertainty is 35%. This is the result of adding in quadrature the theory uncertainty (4% for $W + \text{jets}$ and 5% for $Z + \text{jets}$ and dibosons) and an additional 24% per jet, according to the Berends–Giele scaling [299].
- **Other background modelling:** For the other backgrounds, the theoretical uncertainties on the predicted cross-section are:
 - For $W + \text{jets}$ a 40% uncertainty is considered.

- For tZq , the uncertainty on $\sigma_{\text{NLO}}^{\text{pred}}$ is $^{+7.7\%}_{-7.9\%}$ (scale) + 0.9%(PDF+ α_s) with uncorrelated components. For this process, the uncertainty due to ISR and to higher/lower parton radiation is evaluated following exactly the same prescription as for the t -channel process.
 - For tWZ an uncertainty of 50% is applied. This is done in the same manner as for the $Z + \text{jets}$.
 - For the diboson processes a 24.5% (i.e. 5% for 0 b -jets and 24% for 1 b -jets) is considered [299]. This is calculated as for the $Z + \text{jets}$ or the tWZ processes.
 - For rare top quark and other Higgs boson processes a 50% uncertainty on the cross-section is used.
- **PDF uncertainty:** The uncertainties associated with the PDFs are evaluated using either the PDF4LHC15 or the NNPDF30 sets. These sets consists of various nuisance parameters depending on the considered process [170]. Internal reweighting in the nominal simulation signal or main background samples is used. It is reweighted to the PDF4LHC15 or NNPDF30 PDF and its uncertainty set, and the symmetrised uncertainties are propagated to the measurement of distributions used in the template fit.

In the past, to account for higher and lower parton radiation, the μ_R and μ_F scales were varied by factors of 0.5 and 2. The variations used were $\{\mu_R, \mu_F\} = \{0.5, 0.5\}, \{1, 0.5\}, \{0.5, 1\}, \{1, 1\}, \{2, 1\}, \{1, 2\}, \{2, 2\}$. The final uncertainty is estimated by taking the envelope of all the uncertainties associated with each variation, as it is recommended by the Physics Modelling Group. In order to be aligned with the rest of the analyses, rather than use this approach, the up and down variations are added as described above.

6.7.3 Experimental uncertainties

Experimental uncertainties play a crucial role in particle physics experiments as they arise from the measurement process itself. In the context of ATLAS analyses, these uncertainties primarily stem from detector-related factors and encompass various aspects, such as the limitations of the measurement apparatus, calibration procedures, and the efficiency of reconstructing physical objects within the detector.

This section is focussed the experimental uncertainties considered in this analysis. These uncertainties arise from multiple sources, and their evaluation involves rigorous procedures to ensure the accuracy of the measurements. Some common sources of experimental uncertainties include:

- **Luminosity:** For each data-taking year there is an uncertainty in the integrated luminosity collected by the ATLAS detector, as Table 6.3 shows [179, 244]. These uncertainties are partially correlated between years and it is applied to each MC simulated process to scale them to match the expected number of events at the given luminosity for each year.
- **Pile-up re-weighting:** The events of the MC simulation samples are re-weighted to match the observed distribution of the average number of interactions per bunch-crossing in data [300]. The $\langle\mu\rangle$ is presented in Figure 6.2. The pile-up uncertainty accounts for the differences in $\langle\mu\rangle$ between the simulated and real data. The value of this uncertainty is obtained by re-scaling the $\langle\mu\rangle$ value in data by 1/0.99 and 1/1.07 around the nominal scale factor of 1/1.03 [301].
- **Jet energy scale:** The jet energy scale (JES) calibration corrects the energy and direction of the jets to match that of the jets reconstructed at particle level [302]. The JES associated uncertainty is obtained from test-beam data, LHC collision data and MC simulations. Events with a vector boson and additional jets are used to calibrate jets in the central region. In total, there are 30 independent NPs, each with an up/down variation to address the JES systematic uncertainty. These account for effects related to the pile-up, the jet flavour-composition, single-particle response, the p_T and η dependence and the effects of PS leaving the calorimeters (Punch through).
- **Jet energy resolution:** The jet energy resolution (JER) refers to the ability of the experiment to measure the energy of the jets. A smearing model corresponding to 13 nuisance parameters is used to address the JER uncertainty. The JER is measured separately for data and MC using the two in-situ techniques described in References [302, 303]. A JER uncertainty is defined as the quadratic difference between the JER for data and MC. The associated uncertainty is evaluated by smearing the energy of the jets in the MC-simulated samples by their residual differences and the changes in shapes and normalisations of the final discriminant are compared to the default predictions. To propagate the uncertainty in the p_T resolution, for each jet from the simulated sample, a random number (r) is used. It is generated from a Gaussian distribution centred at zero and with a root mean square equal to the quadratic difference between the fractional p_T resolution with the smearing tool and the nominal one. The four-momentum of the jet is scaled by a factor $1 + r$ and, since the simulated jets cannot be under-smeared, by definition the resulting uncertainty on the normalisation and shape of the final discriminant is one-sided. The JER uncertainty is symmetrised.

- **Jet vertex tagger:** The uncertainties associated to the JVT (see Section 6.3.5) are based on the residual contamination from pile-up jets after pile-up suppression and the MC generator choice [272, 304].
- **Misidentified τ_{had} :** The systematics associated with the τ_{had} identifications are presented in Table 5.2 for both the LLH- and BDT-based methods. This uncertainty depends on the chosen working points.
- **Misidentified τ_{had} rates:** To address the uncertainty associated to the method of determining the probabilities of misidentified τ_{had} , the Template Fit method is compared to the counting method. Details about these two methods are presented in Section 6.5. For the counting method, the uncertainty arises from the global normalisation and for the Template Fit it comes from the quark-to-gluon composition.
- **τ_{had} energy scale:** As well as for the jet, the τ_{had} energy has to be calibrated. This is done in four steps: The first step adjusts the calorimeter measurement right after the reconstruction. The second step subtracts the energy contribution from the pile-up. In the rest of the steps, other effects such as missing clusters are compensated by scaling the true energy of the τ . A number of systematics are related to the calibration of the τ_{had} energy.
- **Heavy- and light-flavour tagging:** The efficiency of the flavour-tagging algorithm is measured for each jet flavour using data and simulation control samples. This assessment yields correction factors to adjust the tagging rates in the simulations. For b -tagged jets, these factors, along with their uncertainties, are estimated from data incoming from di-leptonic $t\bar{t}$ events [236, 304]. For c -jets, the factors stem from jets produced by W boson decays in $t\bar{t}$ events [237]. Light-flavour jet corrections are based on di-jet events [238]. Uncertainties related to b - and c -tagging efficiencies, which vary depending on jet p_{T} , are meticulously assessed, accounting for bin correlations [236]. Additional uncertainties arise when extrapolating the b -tagging efficiency measurement from the p_{T} region used to determine the correction factors to higher p_{T} areas. In total, 19 NPs are considered across light-, b -, and c -jets.
- **Electron efficiency:** The electron efficiency scale factors correct the differences in the reconstruction efficiency between the real and simulated data. For electrons, these scale factors are measured with a “tag-and-probe” method in $Z \rightarrow e^+e^-$ and $J/\psi \rightarrow e^+e^-$ events. The information on the correlation of the different components of the systematic uncertainties is provided for all efficiency measurements. The default uncertainty correlation model provides one up/down variation for a reduced set of uncertainties [211, 305].

- **Muon efficiency:** Similarly to the case of the electrons, scale factors for muons are obtained from $Z \rightarrow \mu^+\mu^-$ and $J/\psi \rightarrow \mu^+\mu^-$ events. These are applied to correct for the differences between data and MC in the muon identification and isolation efficiencies [306] (see Section 6.3.3). Uncertainties on these scale factors are provided by the muon CP group and applied as up/down variations of the nominal scale factors for each component. Additional uncertainties are considered for the “track-to-vertex association” muon scale factors [306, 307]. They are evaluated in the same way as the other components (identification, isolation) of the muon scale factors. In total, six NPs are used.
- **Electron energy scale and resolution:** The accuracy of the electron momentum scale and resolution in MC simulations is verified through the examination of reconstructed mass distributions of $Z \rightarrow e^+e^-$ and $J/\psi \rightarrow e^+e^-$ events. Additionally, E/p studies using $W \rightarrow e\nu$ events are used. When comparing data and MC samples, small discrepancies are observed and corrections are applied to account for this unbalance. These corrections are applied using the tools provided by the E/gamma CP group. All the effects are considered fully correlated in η and they are summed in quadrature to provide up/down variation for a reduced set of uncertainties [211, 305].
- **Muon momentum scale and resolution:** Corrections for momentum scale and resolution have to be applied for MC-simulated muons. The uncertainties of these are obtained varying four NPs related to the ID and the MS. Additional uncertainties are considered to account for the charge-dependent scale correction (“sagitta bias”) applied on data. A more detailed description can be found in Reference [306, 307].
- **E_T^{miss} soft term:** Uncertainties on the scale and resolution are specifically applied to the “soft-track” (“soft term”) on the E_T^{miss} , which cannot be associated to any of the reconstructed and calibrated objects (“hard term”). These uncertainties are derived from the level of agreement between data and MC of the p_T balance between the hard and soft E_T^{miss} components [277]. The scale and resolution uncertainties of E_T^{soft} are treated as separate NPs.

6.8 Fit results

The determination of the absolute normalisations of signal and background yields, taking into account higher-order theoretical predictions, is performed through a fit to the data. This fit aims to normalise the event yields and validate the modelling in both the signal and control regions. By performing the fit

the signal strength ($\mu_{tHq}^{2\ell+1\tau_{\text{had}}}$) and the background normalisation factors (k_p) are obtained¹⁴.

A profile-likelihood-binned-fit method is employed in the SRs and CRs to constrain the systematic uncertainties as *nuisance parameters* (NPs). Importantly, a simultaneous fit is performed considering all regions since no single region exclusively contains a single process. Cross-contamination from multiple backgrounds, as well as the signal, occurs in the CRs. The software package employed to perform the fit is TRExFitter.

In this section, the fit procedure is detailed and the results of the tHq search are presented. The inclusion of systematic uncertainties as constrained NPs, the treatment of signal and background normalisations, and the simultaneous fit to all regions are explained. Subsequently, the signal strength and normalisation factors are presented and discussed in Section 6.8.1.

The different types of fit are described in Section 6.8.2 and the strategy to select the binning is briefly commented in Section 6.8.3. The fit results using the MC-only (Asimov) dataset are presented in Sections 6.8.4 and 6.8.5.

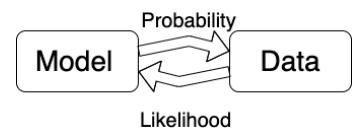
Finally, the results using the collected data are discussed in Sections 6.8.6 and 6.8.7.

6.8.1 Likelihood fit

The likelihood fit is a statistical technique used to estimate the parameters of a model by maximising a likelihood function. The likelihood function, denoted as $L(\vec{x}, \mu, \vec{\theta})$, is defined as the probability of observing a certain set of data (\vec{x}) given a model with NPs ($\vec{\theta}$) and signal strength (μ). In the case of binned distributions, such as histograms, the likelihood function can be expressed as the product of the probabilities of observing the observed and expected number of entries in each bin.

The likelihood function has the form:

$$L(\vec{x}, \mu, \vec{\theta}) = \mathcal{P}(\vec{x}|\mu, \vec{\theta}) = \prod_i \mathcal{P}(x_i|\mu, \vec{\theta})$$



¹⁴The signal strength is defined as the ratio between the observed and expected production cross-sections $\mu_{tHq} = \frac{\sigma^{\text{obs}}(tHq)}{\sigma^{\text{pred}}(tHq)}$. The normalisation factor is the same but for the fitted background process $k_p = \frac{\sigma^{\text{obs}}(p)}{\sigma^{\text{pred}}(p)}$.

where i runs over the data points. $\mathcal{P}(x_i|\vec{\theta})$ is the probability density function for the data point x_i given the model defined by $\vec{\theta}$.

The likelihood function is maximised to obtain the estimated values of the NPs $\vec{\theta}$ and μ that best fit the observed data. This process is known as the maximum likelihood estimation, $(\hat{\mu}, \vec{\theta}) = \text{argmax}_{(\mu, \vec{\theta})} \{L(\vec{\theta})\}$. It is important to note that the likelihood function itself does not provide a direct probability interpretation of the parameter values; it is a measure of the goodness-of-fit between the model and the observed data.

Since particle physics experiments are counting experiments, the probability density function follows the Poissonian statistics. The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event.

Fit for binned distributions

The employed fit is a shape analysis, meaning that the shape of the kinematic distributions is used to find the parameter of interest (POI). The kinematic variables used for the fit in this analysis are presented in histograms, which are binned distributions. For binned distributions, the likelihood function can be written as

$$\begin{aligned} L(\vec{n}|\mu, \vec{\theta}) &= \prod_{i \in \text{bins}} \mathcal{P}(n_i^{\text{obs}}|n_i^{\text{exp}}(\mu, k_p \vec{\theta})) \\ &= \prod_{i \in \text{bins}} \mathcal{P}(n_i^{\text{obs}}|\mu \cdot s_i^{\text{exp}}(\vec{\theta}) + \sum_{p \in \text{Bkg}} k_p \cdot b_i^{\text{exp}}(\vec{\theta})). \end{aligned} \quad (6.1)$$

Here, i runs over the bins of the histogram, n_i^{obs} and n_i^{exp} are the observed and expected number of entries in the bin i . The predicted signal entries in the bin i are $\mu \cdot s_i^{\text{exp}}$ while for particular background process “ p ” are $k_p \cdot b_i^{\text{exp}}$. Therefore, the number of events in i is:

$$n_i^{\text{exp}} = \mu \cdot s_i^{\text{exp}}(\vec{\theta}) + \sum_{p \in \text{Bkg}} k_p \cdot b_i^{\text{exp}}(\vec{\theta}),$$

with the index p running over the different background processes. While the parameter k_p normalisation factor is typically set to the unity for most background processes, it can also be let to float as a free parameter in fit. Later, in Section 6.8.2.1, it is explained why only the normalisation of $t\bar{t}W$ in the $2\ell\text{SS} + 1\tau_{\text{had}}$ is the only one fitted in this search. The result of the fit is the best estimate of the POIs, which in this analysis is the $\mu_{tHq}^{2\ell+1\tau_{\text{had}}}$ but it could also include different k_p . In other words, the $\mu_{tHq}^{2\ell+1\tau_{\text{had}}}$ is the one that maximises $L(\vec{n}|\mu, \vec{\theta})$ in Equation 6.1.

Implementation of uncertainties

The statistical model described in Equation 6.1 provides a method to derive the POI. However, it does not account for the effect of the uncertainties. To factor in these uncertainties, one can introduce a set of additional parameters within the statistical construct to simulate the systematics influence. These parameters are referred as NPs. There are NPs due to the reconstruction imperfections, others are connected to theoretical calculations and some NPs are also coming from the statistical uncertainty of a MC-simulated dataset.

The effect of each systematic uncertainty is encapsulated by generating a modified distribution of the observable, which is then compared against the nominal distribution. For every NP, two alternative distributions of the fitted variable can be built using ± 1 standard deviations variations of the associated systematic uncertainty:

$$\theta = \theta^0 \pm \Delta\theta.$$

Here θ are the varied NPs, θ^0 the nominal values and $\Delta\theta$ the variation around the nominal value. The construction of these varied distributions is discussed in detail in Section 6.7. The deviation between the nominal and varied distributions is parametrised by a singular NP. The NPs are evaluated via dedicated measurements.

Each NP is included as a penalty term in the likelihood expression in Equation 6.1 using a Gaussian distribution for each of them. Therefore, the profile likelihood function to be maximised is:

$$L(\vec{n}|\mu, \vec{\theta}) = \prod_{i \in \text{bins}} \left[\mathcal{P}(n_i^{\text{obs}}|\mu \cdot s_i^{\text{exp}}(\vec{\theta}) + b_i^{\text{exp}}(\vec{\theta})) \times \prod_{j \in \text{syst}} \mathcal{G}(\theta_{j,i}^0|\theta_{j,i}, \Delta\theta_{j,i}) \right],$$

where j runs over the different NPs and \mathcal{G} is the Gaussian distribution function. The $\theta_{j,i}^0$ is the nominal value of NP j on the bin i . Its variation $\Delta\theta_{j,i}$ is set, by convention, to one standard deviation. The i index runs over all bins in all the distributions considered for the fit.

Besides the NPs associated to the systematic uncertainties, other NPs can included in the fitting procedure in the same way as the POI. Another element that is included in the fit as a NP is the effects on the statistical uncertainty. Independent MC-statistical uncertainties are considered in each bin and the NP associated to these statistical uncertainties are known as *gammas* (γ). The gammas are fully decorrelated among different bins and follow Poissonian distributions. Taking the γ -NPs into account, the resulting likelihood function is expressed as:

$$L(\vec{n}|\mu, \vec{\theta}) = \prod_{i \in \text{bins}} \left[\mathcal{P}(n_i^{\text{obs}}|\mu \cdot s_i^{\text{exp}}(\vec{\theta}) + b_i^{\text{exp}}(\vec{\theta})) \times \prod_{j \in \text{syst}} \mathcal{G}(\theta_{j,i}^0|\theta_{j,i}, \Delta\theta_{j,i}) \times \prod_{s \in \gamma} \mathcal{P}(\theta_{s,i}^0|\theta_{s,i}, \Delta\theta_{s,i}) \right]. \quad (6.2)$$

The effect of the systematic uncertainties can be decomposed according to whether they are universal across all bins or they impact the samples differently from bin to bin. The first is called the normalisation effect and second is the shape effect. For instance, the NPs describing the cross-section of a background sample do affect this sample in all bins the same and only affect the normalisation while NPs related to the reconstruction might affect each bin separately (i.e. shape effect).

In summary, the fitting process can be characterised as the optimisation problem involving the maximisation of a multi-dimensional likelihood expression. The outcome of this procedure yields specific values for the μ , k_p and θ_j . The results concerning the NP can present two distinct types of variations. First, the post-fit value of the NP can deviate from its nominal θ_j^0 a phenomenon referred as *pulling*. Secondly, its uncertainty could be less than ± 1 standard deviations from its nominal value, a condition identified as *constraint*.

6.8.2 Fit strategy

The strategy of the fit refers to the choice of the configuration for the fitting procedure. This includes the decision of which k_p are allowed to vary in the calculations, as well as the designation of which regions defined in Section 6.6.4 are used as CR or VRs. The POIs of each fit are discussed in Section 6.8.2.1. Section 6.8.2.2 details the approach used to avoid biases by not examining real data in signal-enriched regions. Afterwards, in Section 6.8.2.3, the different type of fits are discussed. Finally, Section 6.8.2.4, describes the processes of eliminating the contributions from systematic uncertainties that do not significantly impact the analysis.

The choice of distributions for the fit and its binning are also considered to be part of the fit strategy and are described later in Section 6.8.3.

6.8.2.1 Parameters of interest

The POIs in the fit are in the signal strength of the tHq signals and the normalisation factors of the backgrounds, k_p . In the binned-likelihood function in Equation 6.2, the k_p factors do not appear explicitly but they are included in the expected number of background events in bin as $b_i^{\text{exp}} = \sum_{p \in \text{Bkg}} k_p \cdot b_{p,i}^{\text{exp}}$. Here, $b_{p,i}^{\text{exp}}$ is the expected number of events of the background p in the bin i .

As introduced in Section 6.8.1, k_p is fixed to one for most of processes. In this analysis there are several processes that are let to float as free parameters of the fit. For the $2\ell \text{OS} + 1\tau_{\text{had}}$ some discussion has arisen regarding the extraction of the

$k_{t\bar{t}}$ and $k_{Z+\text{jets}}$ normalisation factors in the fit calculation. Since the backgrounds due to misidentifications have already been fitted, as discussed in Section 6.5, the original proposal was to not include $k_{t\bar{t}}$ and $k_{Z+\text{jets}}$ to avoid double fitting. However, due to the large mismodelling¹⁵ of the backgrounds, the normalisation for the $t\bar{t}$ and $Z + \text{jets}$ processes is calculated in order to bring stability to the fit.

Table 6.28 shows the different approaches explored in the configuration of the $2\ell \text{OS} + 1\tau_{\text{had}}$ fit. The first column of the table explores whether to use the regions defined in Section 6.6.4.1 as CR or VRs. The difference between CR and VR is detailed in Section 6.8.2.3. The second column of Table 6.28 refers to whether or not the $k_{t\bar{t}}$ and $k_{Z+\text{jets}}$ are free-floating parameters of the fit. Finally, the option of not using the misidentification scale factors is also explored.

	$t\bar{t}$ and $Z + \text{jets}$	Free parameters	Corrected weights
Baseline	CR	$\mu_{tHq}, k_{t\bar{t}}, k_{Z+\text{jets}}$	✓
Alternative 1	VR	μ_{tHq}	✓
Alternative 2	CR	μ_{tHq}	✓
Alternative 3	CR	$\mu_{tHq}, k_{t\bar{t}}, k_{Z+\text{jets}}$	✗

Table 6.28: Different fit configurations tested for the $2\ell \text{OS} + 1\tau_{\text{had}}$ channel. All of these have been explored (with blinded data) to check how the setup of the fit would influence the results. Baseline refers to the one presented in this thesis.

The “Alternative 1” one does not yield good results because the SR does not have enough statistics to have a stable fit. To solve the problem of the low statistical sample in the fit, the “Alternative 2” includes the background-enriched regions but the discrepancy between the data and the MC-simulation samples is so large that it would result in extreme values of the $\mu_{tHq}^{2\ell \text{OS} + 1\tau_{\text{had}}}$. This is why the “Baseline” option is the one implemented. Alternatively, to avoid the issue with the misidentification scale factors, the weights without corrections could be used and the misidentification rates could be corrected in the same fit that is used to derive the signal strength. This approach is referred to as “Alternative 3”, and the results testing this option are not very promising.

Regarding the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel, the different strategies that have been tested are presented in Table 6.29. The one presented on this thesis is using only a CR consisting on the region of the phase space orthogonal to the SR. Additionally, the use of CRs dedicated to the $t\bar{t}$ and $t\bar{t}X$ processes have been used so that the

¹⁵The mismodelling mentioned arises from an issue in the derivation and application of the weights corresponding to the misidentification scale factors. It affects both $2\ell + 1\tau_{\text{had}}$ channels and constitutes one of the major problems in this analysis. It is hoped that this issue can be resolved in future versions of the NTuples.

fit had more ability to control the backgrounds separately but these options are less stable than the baseline. Only the one referred to as “Alternative 2” in the Table 6.29 provided results comparable to these of the “Baseline” approach.

	CRs	Free k_p
Baseline	All bkgns	$k_{t\bar{t}, t\bar{t}X}$
Alternative 1	$t\bar{t}, t\bar{t}X$	$k_{t\bar{t}}, k_{t\bar{t}W}$
Alternative 2	$t\bar{t}, t\bar{t}X$	$k_{t\bar{t}}, k_{t\bar{t}X}$
Alternative 3	$t\bar{t}, t\bar{t}X$	$k_{t\bar{t}}, k_{t\bar{t}W}, k_{t\bar{t}H}, k_{t\bar{t}Z}$

Table 6.29: Different fit configurations tested for the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel. All of these have been explored with blinded data to check how the setup influence the results.

6.8.2.2 Blinding

In formulating the analysis strategy, it is crucial to avoid examining data regions which are anticipated to be rich in tHq -signal events. This practice, known as *blinding*, plays a vital role in safeguarding against the influence of bias in results, which can occur if researchers are swayed by statistical variations observed within the data. Consequently, unless specified otherwise, the determination of the strategy (e.g. building of the BDT models, region definition, region-binning tuning, etc) relies entirely on blinded data. In this research, the data point in a bin is only shown if the expected signal fraction is smaller than 0.3% and all bins in the SRs were fully blinded.

Even though the plots presented in this thesis are not blinded, all distributions remained blinded while carrying this analysis. Only in Sections 6.8.6 and 6.8.7 the search is done with unblinded data.

6.8.2.3 Fit types

The configuration of the fits allows to set up three different types of data regions:

- **Signal region:** A signal-enriched region that drives the signal extraction, i.e. the μ_{tHq} , but adds negligible constraints on the background parameters (k_p).
- **Control region:** A background-enriched region with negligible signal contamination. These regions are used in the fit calculations to derive the POIs. Typically, the CR of a background-process p is used to control it by obtaining

its k_p factor. Note that this p -background can either be a single process or a collection of processes. For instance, the $k_{t\bar{t}W}$ in Table 6.29 targets a single process while the $k_{t\bar{t}X}$ or $k_{t\bar{t}X}$ control the production of several processes simultaneously.

- **Validation region:** A region in which the background and signal levels are predicted, but which are not used in the fit to constrain the POIs. The VRs serve to validate the extrapolation from the fit results.

The fit is based on orthogonal signal and control regions. This way all the CRs and SRs are statistically independent, can be modelled by separate profile fits, and thus be combined into one simultaneous fit. The background parameters are predominantly constrained by CRs, probably with large statistics, which in turn reduces the impact of their uncertainties in the SR.

Given these types of regions, there are three types of fit configurations that are used in the analysis:

- Using Asimov dataset in all regions: The purpose of performing the fit using only Asimov data is to evaluate the sensibility of the experiment and to set the expected limits in the measurement. The results using the Asimov dataset are presented in Section 6.8.4
- Background fit in the CRs: This fit is performed using only the CRs for the fit, and with the blinding threshold activated. The background-only hypothesis consists of setting the strength of the signal process to zero ($\mu_{tHq} = 0$). The purpose of the CR-only–background-only fit is to estimate the backgrounds in the SRs and VRs without assumptions on the signal model. Here, only the CRs are used to constrain the fit parameters neglecting any potential tHq contribution.
- Using blinded data in CRs and SR and the MC-simulated backgrounds only: After performing the CR-only–background-only fit, the fit can be extended to the SRs in the CRSR–background-only fit. Here the SR remains blinded only in for distribution plots but the data is used in profile-likelihood fit. In contrast to the CR-only–background-only fit, the real data in the SR is used in the fit calculations. This allows to check how the parameters of the previous fit are affected by including the SR. All the different configurations presented in Tables 6.28 and 6.29 are explored to this extent.
- Full-data fit using the CRs and SR: This is the last step in the analysis. Once all the already-described fits have been carried out, and no problems are detected, one shall fit the data in all regions without any assumption about

the normalisation factors or the signal strengths. First, this is done using blinded data and, finally, with unblinded data. In Sections 6.8.6 and 6.8.7, the profile-likelihood-fit over all the regions with unblinded data is presented for the two considered channels. Note that only one fit configuration from Table 6.28 and one from Table 6.29 are fitted with unblinded data.

6.8.2.4 Pruning

In Section 6.8.1 is explained that the systematic uncertainty can impact the fit through normalisation and/or shape effects. If a NP has an overall little impact on the fit, it can be ignored from the calculations. This process of removing the systematics that do not affect much the fit is named *pruning*. By applying the pruning over the NPs, the fit procedure is less consuming computationally and instabilities due to the large number of NPs are avoided.

The impact of each NP is checked separately for the shape and normalisation. If the normalisation impact of a given systematic uncertainty is smaller than 0.5% in a particular region and for a sample, the normalisation uncertainty of that NP is removed from the fit for that region and that event sample. Regarding the shape impact, it is evaluated by normalising both the nominal and the varied distributions and checking bin by bin the change. If the variation is smaller than 1% for each bin, the shape impact is not considered. If a NP has negligible shape and normalisation contributions, it is dropped for the given region and event sample. If a systematic uncertainty is removed across all regions and every event sample, it is entirely excluded from the likelihood model.

6.8.3 Binning optimisation and distributions for the fit

The profile-likelihood-binned-fit procedure described in Section 6.8.1 is performed over binned distributions to obtain the POIs. The binning of these distributions play a role in the result of the fit. Note that the binned distributions are the same through all the steps of the fit i.e. Asimov, CR-only–background-only, and full data fit.

Binning refers to the process of defining the size of the bins in a histogram. While the distributions of the variables are often presented using equally-sized bins (see plots in Appendix C as an example), this is not a requirement. In fact, when performing the binned likelihood fit fits, using equally-sized bins may not be the optimal choice. A finer binning means better S/B and S/\sqrt{B} in the most important bins for the fit. On the one hand, the more bins there are, the more information

is retained about the shape of the distribution. On the other hand, too many bins can lead to a small statistical significance of some bins. To mitigate statistical fluctuations in specific bins, the binning process can be optimised. The optimisation aims to enhance the separation of the tHq signal from the background and minimise bins with large statistical errors. The automatic-binning-Transform-D algorithm examines the original distribution, initiating from the bin with the highest BDT score. It then merges bins until a certain fraction of signal and background events is achieved. This merging threshold is determined by the function Z :

$$Z = z_b \frac{n_b}{N_b} + z_s \frac{n_s}{N_s},$$

where n_s and n_b denote the number of signal and background events in the merging bin, respectively. Similarly, N_s and N_b represent the total counts of signal and background events. The parameters z_s and z_b can be adjusted as needed. A bin is established once Z reaches 1 or exceeds it. The parameters z_s and z_b govern the maximum portion of signal and background events in a bin, respectively, subject to the condition $z_s + z_b = N_{\text{bins}}$.

6.8.3.1 Fitted distributions in the $2\ell \text{ OS} + 1\tau_{\text{had}}$

The distribution of the $\text{BDT}(tHq|\text{OS})$ discriminant is fitted to extract $\mu_{tHq}^{2\ell \text{ OS}+1\tau_{\text{had}}}$ in the SR. Regarding its binning, after performing several tests, the best results are achieved with $z_s = 1$ and $z_b = 2$ with the Transform-D algorithm and some fine tuning by hand. The SR is presented in Figure 6.29a.

Meanwhile, $\text{BDT}(t\bar{t}|\text{OS})$ and H_T are used for the two CRs dedicated to the $t\bar{t}$ and $Z + \text{jets}$, respectively. Figures 6.29b and 6.29c presents these two distributions which are used to extract the normalisation factors $k_{t\bar{t}}^{2\ell \text{ OS}+1\tau_{\text{had}}}$ and $k_{Z+\text{jets}}^{2\ell \text{ OS}+1\tau_{\text{had}}}$.

6.8.3.2 Fitted distribution in the $2\ell \text{ SS} + 1\tau_{\text{had}}$

In the $2\ell \text{ SS} + 1\tau_{\text{had}}$ the distribution for the SR is the $\text{BDT}(tHq|\text{SS})$ and preliminary binning has been selected using the transform-D algorithm for automatic with parameters $z_s = z_b = 2$. Afterwards, some fine tuning using by hand is applied to merge bins with low statistics. The SR is fitted in the $\text{BDT}(tHq|\text{SS})$ distribution. The background-enriched region is fitted on the H_T distribution. The choice of H_T is motivated by the fact that this variable allows to discriminate between $t\bar{t}$ and $t\bar{t}X$ in the scenario where these two processes have its own free-floating k_p . However, due to the low statistical sample present in this channel, a single CR is used, CR(All Bkg). The fitted distributions are presented in the Figure 6.30.

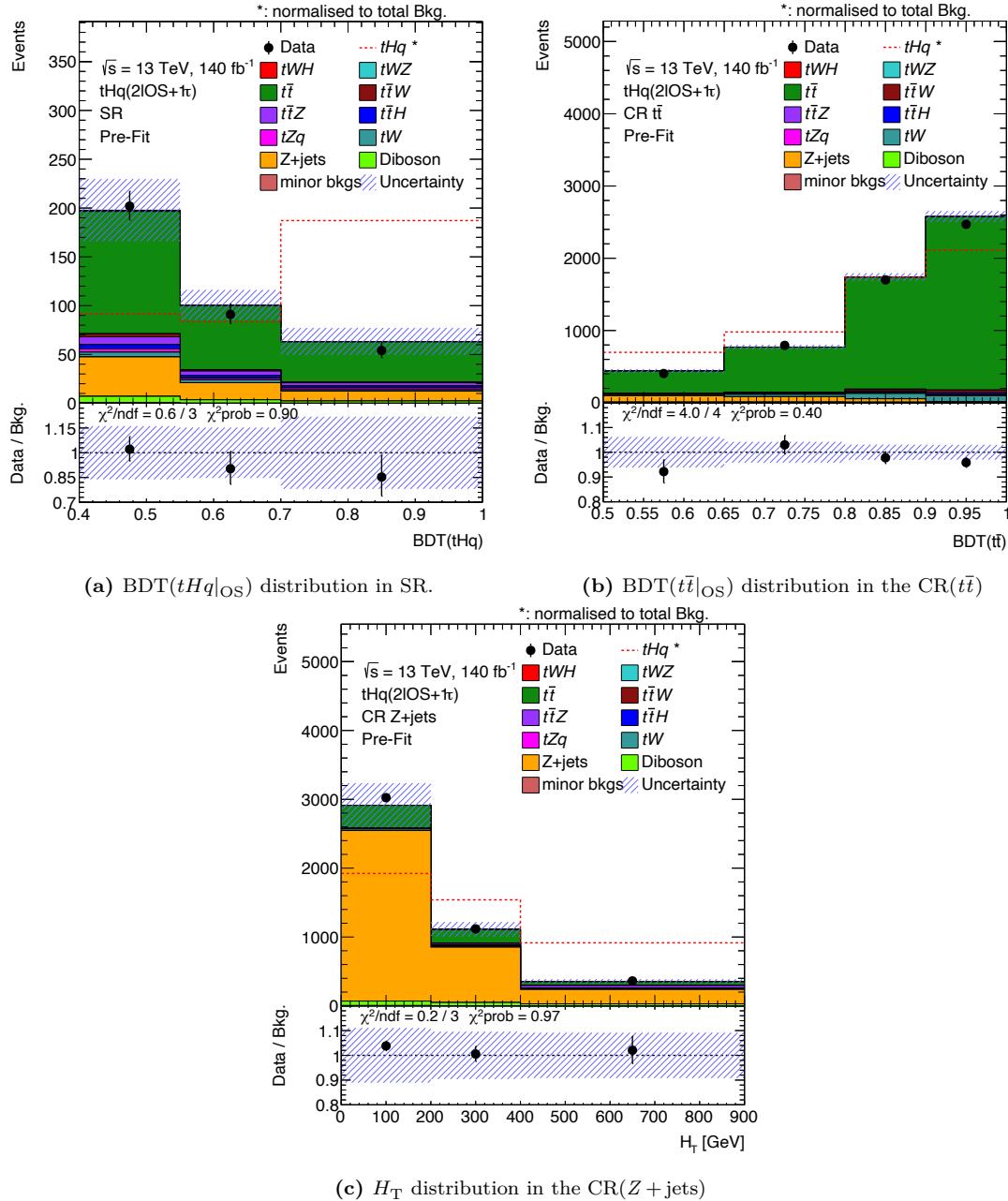


Figure 6.29: Pre-fit distributions used for the (a) tHq SR, and CR of the (b) $t\bar{t}$ and (c) $Z + \text{jets}$ processes in the binned-profile-likelihood fit of the $tHq(2\ell \text{ OS} + 1\tau_{\text{had}})$ channel. Note that, while the data is being showed in (a), all the fit strategy is defined with blinded data in the SR. The dotted-red line corresponds to the normalised tHq signal. The statistical and systematic uncertainties in the MC-simulated samples are represented with the dashed line and. The error bars correspond to the statistical uncertainty of the data.

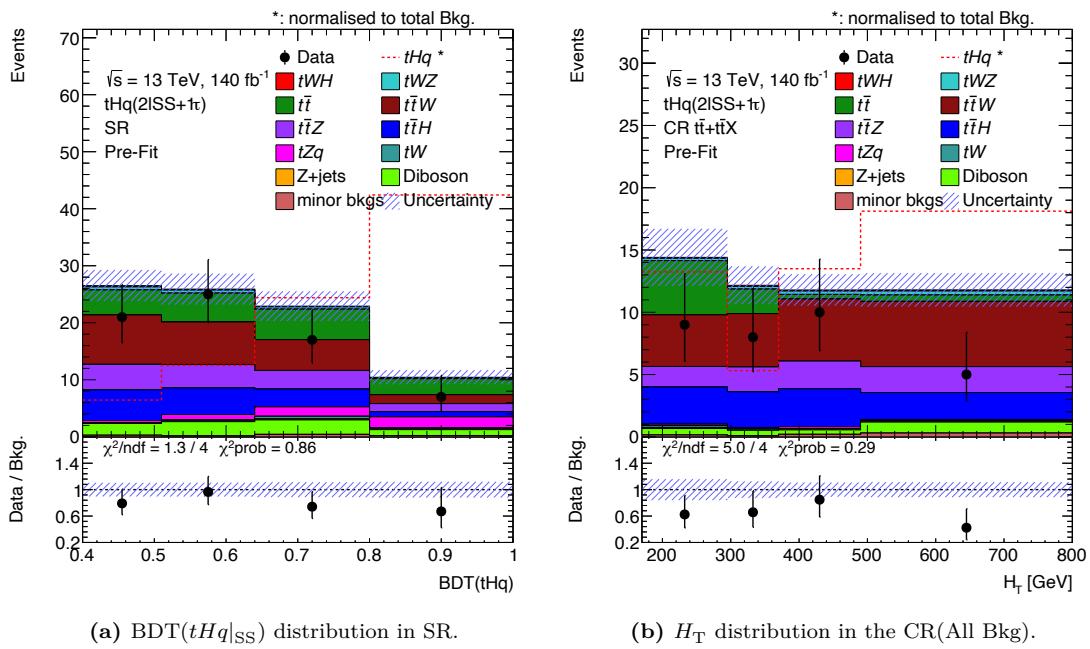


Figure 6.30: Pre-fit distributions used for the (a) tHq SR and (b) CR(All Bkg) in the binned-profile-likelihood fit of the $tHq(2\ell \text{ SS} + 1\tau_{\text{had}})$ channel. Note that, while the data is being showed in (a), all the fit strategy is defined with blinded data in the SR. The dotted-red line corresponds to the normalised tHq signal. The statistical and systematic uncertainties in the MC-simulated samples are represented with the dashed line and. The error bars correspond to the statistical uncertainty of the data.

6.8.4 Asimov-hypothesis fit in the 2ℓ OS + $1\tau_{\text{had}}$

The Asimov fit consists of using only the MC-generated events as if these were the collected data. Asimov datasets are built as binned datasets, in which the n_i^{obs} in each bin is set to the n_i^{exp} for the chosen model parameters. In the Asimov hypothesis, the signal strength in Equation 6.2 is set to the unity. In this same expression, the different NPs are set to their nominal values ($\theta = \theta^0$).

Since the μ_{tHq} and the k_p are fixed, in the Asimov fit one can extract information about the impact of the different uncertainties in the final results. In other words, the Asimov fit allows to estimate the median experimental sensitivity of a search or measurement as well as fluctuations about this expectation [292]. That sensitivity and the expected-upper limit on the production cross-section are the results presented in Section 6.8.4.2.

6.8.4.1 Nuisance parameters in the 2ℓ OS + $1\tau_{\text{had}}$ Asimov fit

The pruning procedure to remove the non-impactful NPs is described in Section 6.8.4.4. The results of the pruning are presented in Section E.1 of Appendix E. There it can be seen that most of the NPs are completely pruned or that shape contribution is dropped.

The Asimov setup is employed to assess the influence of systematic uncertainties. It is anticipated that the majority of the uncertainties of the NPs will lie between the ± 1 standard deviations ($\pm 1\sigma$). However, there may be instances where the fit process uncovers additional information about a NP within the fit dataset, leading to a constraint of that NP. From the NPs that survived the pruning, only the ones shown in Figure 6.31 presented constraints. Note how the black line denoting the relative uncertainty on the NP is smaller than the width of the green band in Figure 6.31. The green band represents ± 1 standard deviations from the nominal value of the NP.

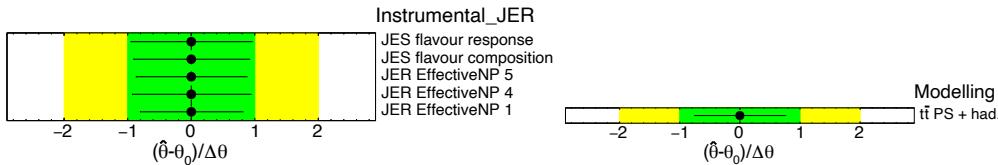


Figure 6.31: Pull plots of the constraint NPs under the Asimov hypothesis in the 2ℓ OS + $1\tau_{\text{had}}$ channel. Each NP is shown as the relative change from its nominal value. The green and yellow areas represent the $\pm 1\sigma$ and $\pm 2\sigma$ deviations from the nominal value of the NP, respectively. The points represent the best-fit value for the NP and the uncertainty bars represent the post-fit uncertainty.

Apart from the systematic uncertainties, it is necessary to evaluate the uncertainty associated with the γ s of the bins used for the fit. These γ s accounting for the statistical uncertainty of the MC dataset in the $2\ell \text{OS} + 1\tau_{\text{had}}$ Asimov fit are presented in Figure 6.32.

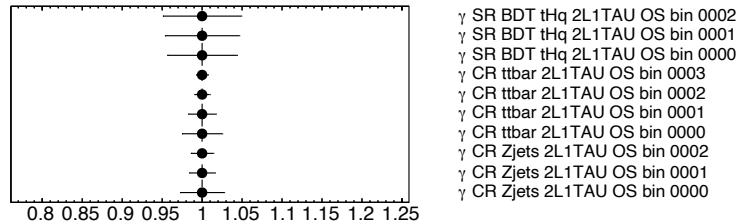


Figure 6.32: The γ s of all bins used in the $2\ell \text{OS} + 1\tau_{\text{had}}$ channel under the Asimov hypothesis. The γ s are associated to the bins in the SR from left to right and numbered starting with 0.

The correlations play an important role in the profile likelihood fit, being responsible for the majority of the reduction of the post-fit uncertainty. As Figure 6.33 shows, there are no high correlations in the $2\ell \text{OS} + 1\tau_{\text{had}}$ channel. The largest correlation is of a 35.7% between the k_{Z+jets} normalisation factor and the JES pileup ρ topology NP.

The impact of the NPs by groups is presented in Table 6.30, where it can be seen that theoretical uncertainties are the dominant contribution to the total systematic error. The most important theoretical uncertainty is that of the $t\bar{t}$ NLO generator, followed by the $t\bar{t}$ FSR. The NPs associated with the PDFs also rank high in this table. This is due to the contribution of the PDF uncertainty of the tHq production. It makes sense that, in terms of determining the signal strength of the tHq , the modelling of the tHq process plays an important role in the error. Additionally, leftmost panel in Figure E.4 hints the impact of the tHq PDFs uncertainty by not dropping completely almost any of its associated NPs in the SR. This systematic error of ± 7.5 is the largest among all PDF uncertainties in this analysis.

The influence of the different NPs over the results is ranked in Figure 6.34 according to its impact on the determination of μ_{tHq} . The impact is calculated by fixing the specific NP to its nominal value and varying the others upwards or downwards. Then, the value of the impact is the μ_{tHq} obtained in each of these four configurations minus the μ_{tHq} obtained in the nominal fit.

The top-ranked NP is the MC uncertainty in the last bin of the SR distribution. It can be seen that there are not many events in the last bin in Figure 6.29a and, hence, it makes sense to have this γ in the first place of the ranking. The other highly ranked NPs are related to the generation of the $t\bar{t}$ events. This could be expected since $t\bar{t}$ is the main background in the $2\ell \text{OS} + 1\tau_{\text{had}}$ channel and its

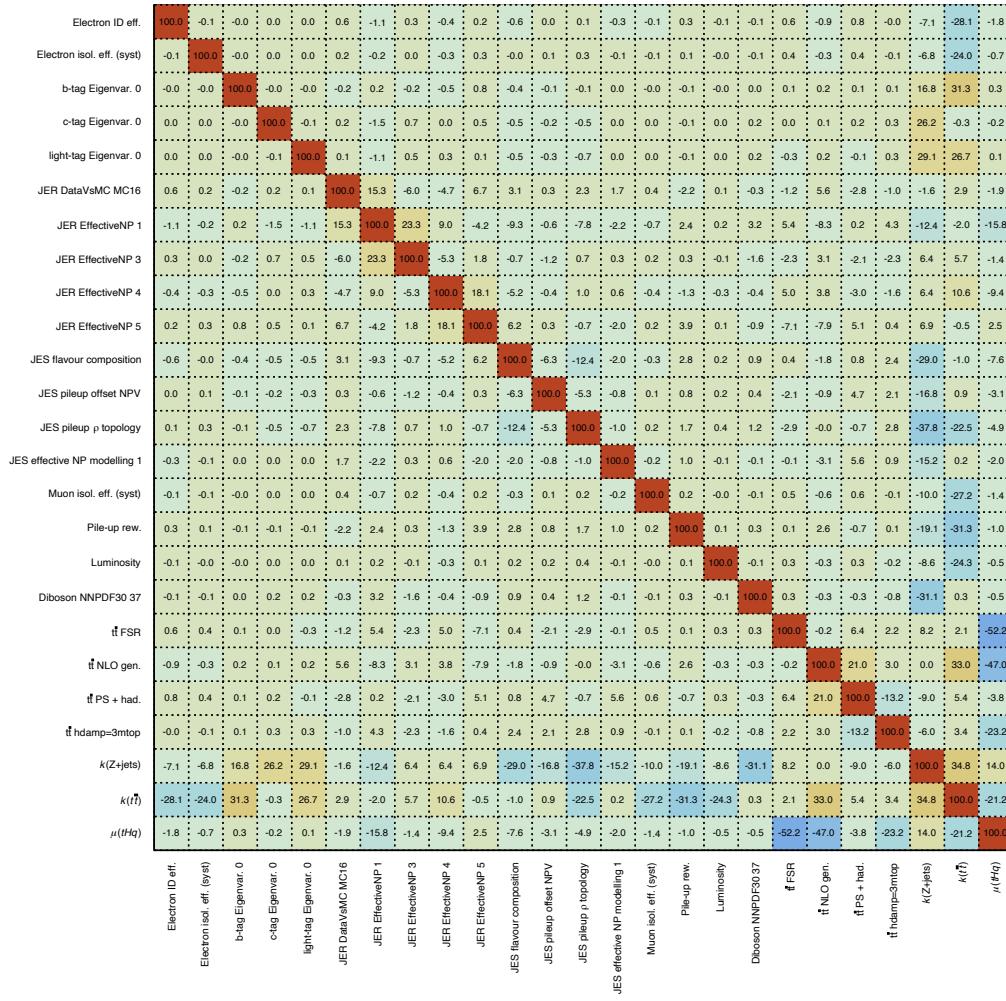


Figure 6.33: Correlation between the different NPs and the POIs in the 2ℓ OS + $1\tau_{\text{had}}$ in the Asimov hypothesis. Only the NPs with at least one correlation above 15% are shown.

presence is not only confined to its dedicated CR($t\bar{t}$) but it is also the dominant process in the SR (see Table 6.25).

6.8.4.2 Asimov-fit results 2ℓ OS + $1\tau_{\text{had}}$

The expected sensitivity to the signal strength and the normalisation factors in the 2ℓ OS + $1\tau_{\text{had}}$ channel is:

$$\Delta \mu_{tHq} = {}^{+31.52}_{-32.66} \text{ (tot.)} {}^{+15.87}_{-14.85} \text{ (stat.)} \quad (6.3)$$

$$\Delta k_{t\bar{t}} = \pm 0.04 \text{ (tot.)} \pm 0.02 \text{ (stat.)} \quad (6.4)$$

$$\Delta k_{Z+jets} = \pm 0.10 \text{ (tot.)} \pm 0.02 \text{ (stat.)}. \quad (6.5)$$

Uncertainty source	Grouped impact
MC uncertainty	± 7.845
Modelling	
Theoretical uncertainties	± 11.917
PDF uncertainty	± 7.895
Experimental	
Instrumental	± 4.917
Flavour tagging	± 0.439
JES/JER	± 10.851
NormFactors	± 10.493
Total systematic uncertainty	± 28.511

Table 6.30: Impact of different groups of systematic uncertainties in the measurement of μ_{tHq} in the $2\ell \text{OS} + 1\tau_{\text{had}}$ channels under the Asimov hypothesis. The impact of each group of uncertainties is computed by performing a fit where the NPs in the group are fixed to their best-fit values, and then subtracting the resulting uncertainty on the μ_{tHq} in quadrature from the nominal fit. The line “MC uncertainty” refers to the statistical uncertainties in the MC events (see “gammas” in Section 6.8.1). The other theoretical uncertainties refer to the uncertainties on the cross-section, the ISR/FSR, alternative models, and the other elements described in Section 6.7.2. The total uncertainty is different from the sum in quadrature of the different components due to correlations between NP built by the fit.

The total uncertainty (tot.) includes statistical and systematic effects. The statistical uncertainty alone (stat.) is also shown, separately. The overall uncertainty on μ_{tHq} is dominated by the contribution of the systematic uncertainties. The systematic uncertainty ($^{+27.23}_{-29.09}$) is larger than the statistical one ($^{+15.87}_{-14.85}$). Nevertheless, an improvement in data statistics would improve the sensitivity over μ_{tHq} .

The fit described above focuses on measuring in the SM tHq process. However, as it is discussed in Section 2.3.3.3, the cross-section of the tH production would increase by a factor of 10 in the CP-violating scenario in which the Yukawa coupling between the top quark and the Higgs boson has the opposite sign with respect to the SM predictions. This increase in the cross-section would make the tHq observable with the dataset analysed in this thesis. The $y_t = -y_t^{\text{SM}}$ is referred inverted Yukawa coupling hypothesis. It is explored by substituting the tHq and tWH MC event samples by an alternative simulation in which $y_t = -y_t^{\text{SM}}$. By doing so, the following expected sensitivity to the tHq process in the $2\ell \text{OS} + 1\tau_{\text{had}}$ channel is obtained:

$$\Delta\mu_{tHq,y_t=-y_t^{\text{SM}}} = {}^{+23.64}_{-19.88} \text{ (tot.)} {}^{+5.36}_{-5.10} \text{ (stat.)}.$$

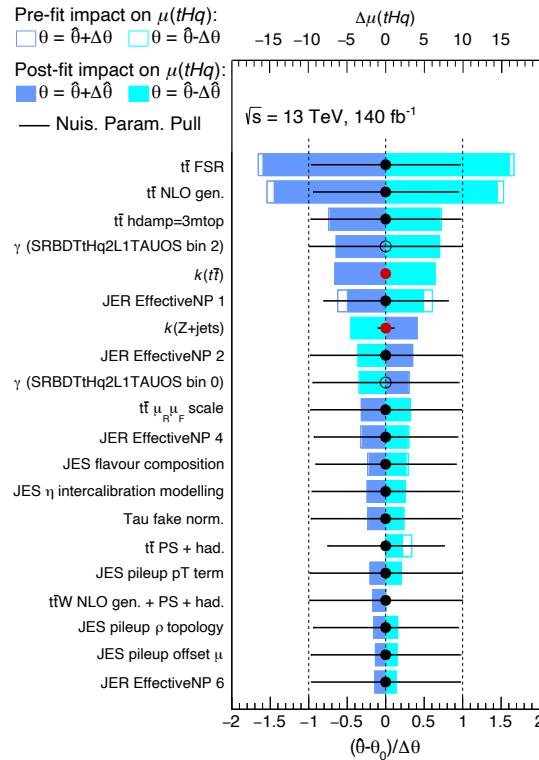


Figure 6.34: Ranking of the most impactful NPs on the Asimov fit in the 2ℓ OS + $1\tau_{\text{had}}$ channel. The NPs are sorted by their impact on the μ_{tHq} in decreasing order. The blue and cyan boxes refer to the upper x-axis and show the impact on μ_{tHq} . The empty rectangles show the pre-fit impact and filled the post-fit. Moreover, the NPs values and their uncertainties are also included as dots and lines, respectively. The uncertainty of the NPs is measured with the lower x-axis.

6.8.5 Asimov-hypothesis fit in the 2ℓ SS + $1\tau_{\text{had}}$

The steps in the fit of the 2ℓ SS + $1\tau_{\text{had}}$ channel are the same as is the channel in which the light leptons have opposed electrical charge. First, the profile-likelihood fit is performed under the Asimov hypothesis. Under this hypothesis the MC-samples are used instead of the data collected by the ATLAS detector. Therefore, in each bin is satisfied that $n_i^{\text{obs}} = n_i^{\text{exp}}$.

As mentioned in Section 6.8.3.2, the 2ℓ SS + $1\tau_{\text{had}}$ fit is performed over two profiles. Apart from the tHq signal, the backgrounds containing a top-quark-antiquark pair are normalised simultaneously, i.e., a single normalisation factor, $k_{t\bar{t}, t\bar{t}X}$, is used to control $t\bar{t}$, $t\bar{t}W$, $t\bar{t}H$ and $t\bar{t}Z$.

6.8.5.1 Nuisance parameters in the $2\ell \text{SS} + 1\tau_{\text{had}}$ Asimov fit

The pruning of NPs in the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel is presented in Section E.2. For instance, in this appendix can be seen that all NPs related to the PDF of the $t\bar{t}$ and $t\bar{t}W$ generation are pruned from the analysis. In that section can be seen that most of the NPs are dropped completely. The normalisation of the NPs is kept sometimes and, for a few NPs, the entire information or only the shape is kept. All NPs that survived the pruning process are almost completely free of constraints. Figure 6.35 shows how the $t\bar{t}$ NLO generator NP is slightly constrained.

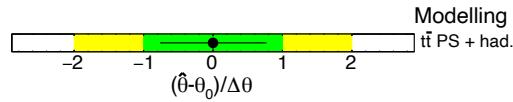


Figure 6.35: Pull plot for the $t\bar{t}$ NLO generator NP under the Asimov hypothesis in the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel. Each NP is shown as the relative change from its nominal value. The green and yellow areas represent the $\pm 1\sigma$ and $\pm 2\sigma$ deviations from the nominal value of the NP, respectively. The points represent the best-fit value for the NP and the uncertainty bars represent the post-fit uncertainty.

The correlation matrix for the NPs in the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel is presented in Figure 6.36. No significant correlation is present in the matrix. The normalisation $k_{t\bar{t}, t\bar{t}X}$ 35.7% anticorrelation with the signal strength is the largest element of Figure 6.36.

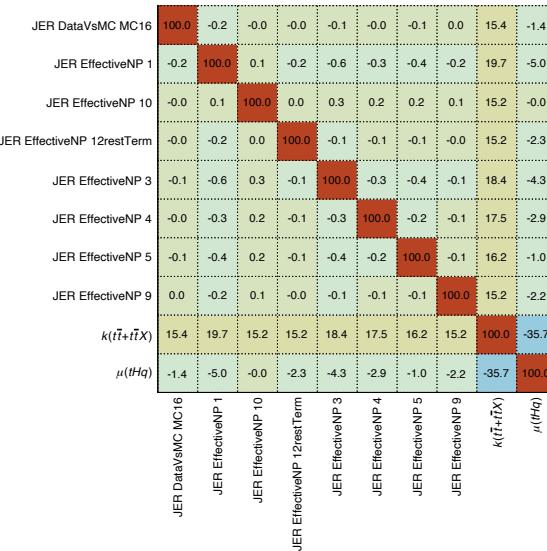


Figure 6.36: Correlation between the different NPs and the POIs in the $2\ell \text{SS} + 1\tau_{\text{had}}$ in the Asimov hypothesis. Only the NPs with at least one correlation above 15% are shown.

To account for the statistical uncertainty of the MC dataset in each of the bins defined for $2\ell \text{SS} + 1\tau_{\text{had}}$ the gamma γ s are used. These are presented are presented in Figure 6.37 for the Asimov fit. As can be seen there, all γ uncertainties are smaller than 0.1 on both sides.

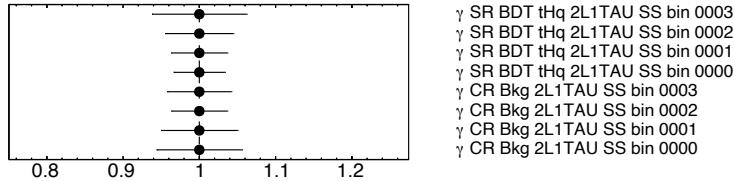


Figure 6.37: The γ s of all bins used in the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel under the Asimov hypothesis. The γ s are associated to the bins in the SR from left to right and numbered starting with 0.

Table 6.31 ranks the NPs by groups based on their influence on the sensitivity of $\mu_{tHq}^{2\ell \text{SS} + 1\tau_{\text{had}}}$. It can be seen that the largest group is that of the JES and JER systematics. It is followed by the normalisation factor of the $t\bar{t}$, $t\bar{t}W$, $t\bar{t}H$ and $t\bar{t}Z$ backgrounds. The PDF uncertainty is small when compared to the one in $2\ell \text{OS} + 1\tau_{\text{had}}$ channel. The ranking of the most relevant NPs is presented in Figure 6.38. There, it is shown that the normalisation of the $t\bar{t} + t\bar{t}X$ processes is the most impactful element on the tHq fit. This makes sense since, as Figure 6.30 suggests, the background processes have to be scaled significantly to fit the data in the CR and this affects the distributions in the SR and, hence, the measurement of μ_{tHq} . The next most significant contribution is the statistical significance rightmost bin of the SR. As Figure 6.30a shows, this bin has the largest concentration of tHq events but also the smallest number of events.

6.8.5.2 Asimov-fit results $2\ell \text{SS} + 1\tau_{\text{had}}$

The expected sensitivity to the tHq signal strength and the $k_{t\bar{t},t\bar{t}X}$ normalisation factor in the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel is:

$$\Delta\mu_{tHq} = \pm 6.51(\text{tot.}) \pm 6.06(\text{stat.}) \quad (6.6)$$

$$\Delta k_{t\bar{t},t\bar{t}X} = \pm 0.16(\text{tot.}) \pm 0.12(\text{stat.}) \quad (6.7)$$

The statistical error is the dominant component in the total $\mu_{tHq}^{2\ell \text{SS} + 1\tau_{\text{had}}}$ uncertainty. The sensitivity to the tHq production improves significantly with respect to the $2\ell \text{OS} + 1\tau_{\text{had}}$ channel (Equation 6.3). Regarding the uncertainty on the background, it is larger than for the $2\ell \text{OS} + 1\tau_{\text{had}}$ channel. This is partially explained by the fact that the background contribution is remarkably reduced in the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel. Although, the majority of the uncertainty over $k_{t\bar{t},t\bar{t}X}$ is due to the systematic contributions.

Uncertainty source	Grouped impact
MC uncertainty	± 1.376
Modelling	
Theoretical uncertainties	± 1.262
PDF uncertainty	± 0.837
Experimental	
Instrumental	± 0.793
Flavour tagging	± 0.163
JES/JER	± 2.877
NormFactors	± 2.725
Total systematic uncertainty	± 4.614

Table 6.31: Impact of different groups of systematic uncertainties in the measurement of μ_{tHq} in the $2\ell \text{SS} + 1\tau_{\text{had}}$ channels under the Asimov hypothesis. The impact of each group of uncertainties is computed by performing a fit where the NPs in the group are fixed to their best-fit values, and then subtracting the resulting uncertainty on the μ_{tHq} in quadrature from the nominal fit. The line “MC uncertainty” refers to the statistical uncertainties in the MC events (see “gammas” in Section 6.8.1). The other theoretical uncertainties refer to the uncertainties on the cross-section, the ISR/FSR, alternative models, and the other elements described in Section 6.7.2. The total uncertainty is different from the sum in quadrature of the different components due to correlations between NPs built by the fit.

In the same way that the result in Equation 6.8.5.2 is obtained, the expected sensitivity to the tHq process with inverted Yukawa coupling in the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel is:

$$\Delta\mu_{tHq, y_t = -y_t^{\text{SM}}} = \pm 1.72(\text{tot.}) \pm 1.30(\text{stat.}) .$$

The procedure to perform the fit with $y_t = -y_t^{\text{SM}}$ coupling is the same as for the SM studies.

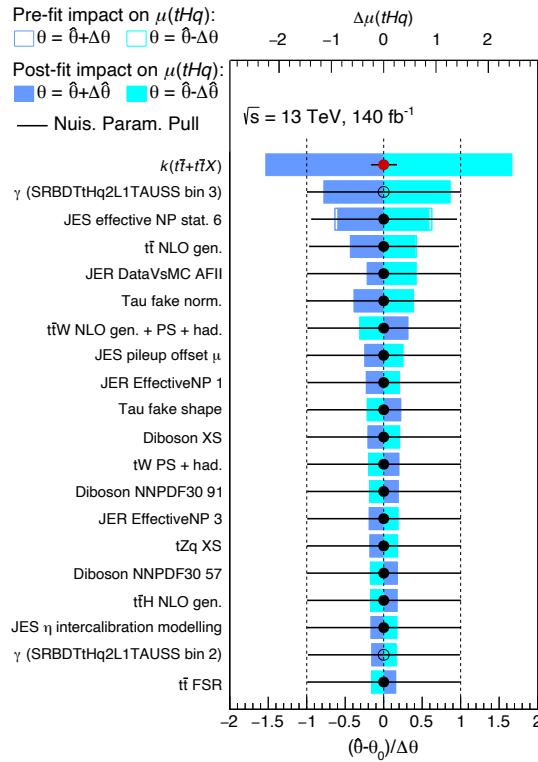


Figure 6.38: Ranking of the most impactful NPs on the Asimov fit in the 2ℓ SS + $1\tau_{\text{had}}$ channel. The NPs are sorted by their impact on the μ_{tHq} in decreasing order. The blue and cyan boxes refer to the upper x-axis and show the impact on μ_{tHq} . The empty rectangles show the pre-fit impact and filled the post-fit. Moreover, the NPs values and their uncertainties are also included as dots and lines, respectively. The uncertainty of the NPs is measured with the lower x-axis.

6.8.6 Full-data fit results in the 2ℓ OS + $1\tau_{\text{had}}$ channel

In the case of the fit to the data, there are no ad-hoc conditions applied to the signal strength, the normalisation factors or the NPs. Therefore, in contrast to the Asimov fit the values of μ_{tHq} and the k_p can diverge from one, and the θ s can be different from their θ^0 s (i.e. pull).

6.8.6.1 Nuisance parameters in the 2ℓ OS + $1\tau_{\text{had}}$ full-data fit

When the data is included in the fit, the NPs can be pulled. In Figure 6.39 only the pulled NPs are presented. The pulls and constraints of all other NPs in the CR-only–background-only fit and in the unblinded-full fit do not present any unexpected behaviour. The largest pull corresponds to the NP of the simulation of

PS and hadronisation for the $t\bar{t}$ processes. The generation of the hard scattering in $t\bar{t}$ also presents the second largest pull.

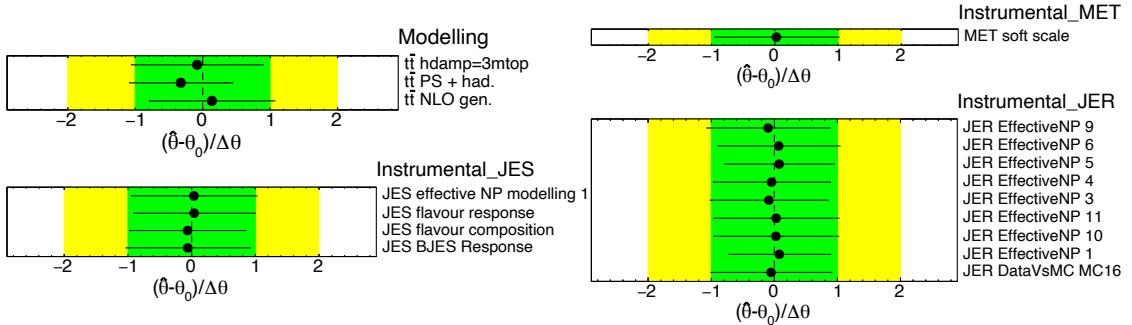


Figure 6.39: Pulled NPs in the unblinded-data-full fit of the 2ℓ OS + $1\tau_{\text{had}}$. The rest of NPs do not present any significant pull or constraint from its nominal value.

The observed pulls can be ranked according to their significance. The pull significance is a quantitative measure of the compatibility between data and constraints. This is particularly important to spot nuisance parameters with a non-zero value preferred by data but are not constrained.

For each NP, it is computed as Equation 6.8 expresses. The most significant pulls are presented in Figure 6.40. There is no significant pull.

$$\text{pull significance} = \frac{|\text{post fit value}|}{\sqrt{1 - (\text{postfit error})^2}}. \quad (6.8)$$

When unblinding the data, the pulls can appear not only on the NPs but also in the γ factors. These are shown in Figure 6.41. There it can be seen that no significant pulls are present in the γ factors and, while some γ drift from 1, their are still compatible with the SM when their uncertainties are taken into account.

The correlation matrix for all the NPs is shown in Figure 6.42. In this figure, two elements stand out with -52.8% and 44.9% of anticorrelation with μ_{tHq} . These correspond, respectively, to the $t\bar{t}$ FSR and $t\bar{t}$ NLO generator. Note that the normalisation factors $k_{Z+\text{jets}}$ and $k_{t\bar{t}}$ are 35.4% correlated.

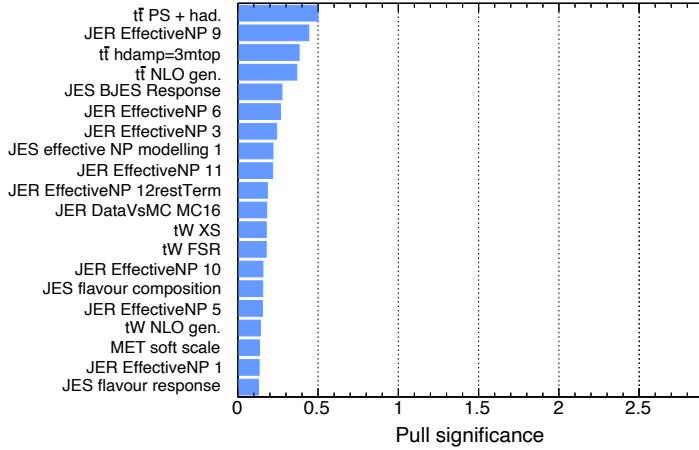


Figure 6.40: The 20 largest pull significances for the 2ℓ OS + $1\tau_{\text{had}}$ in an ordered manner, with the most significant pull shown first.

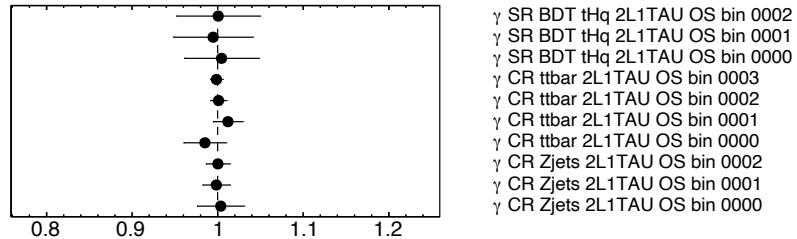


Figure 6.41: The γ_s s of all bins used in the unblinded-data-full fit of 2ℓ OS + $1\tau_{\text{had}}$ channel. The γ_s s are associated to the bins in the SR from left to right and numbered starting with 0.

6.8.6.2 Results of the 2ℓ OS + $1\tau_{\text{had}}$ full-data fit

The result of the profile-likelihood-binned fit in the 2ℓ OS + $1\tau_{\text{had}}$ channel produces the following signal strength and normalisation factors:

$$\mu_{tHq} = -22.11^{+29.22}_{-34.08} (\text{tot.})^{+14.76}_{-13.72} (\text{stat.}) \quad (6.9)$$

$$k_{t\bar{t}} = 0.97 \pm 0.03 (\text{tot.}) \pm 0.02 (\text{stat.}) \quad (6.10)$$

$$k_{Z + \text{jets}} = 1.03 \pm 0.11 (\text{tot.}) \pm 0.02 (\text{stat.}) . \quad (6.11)$$

The total uncertainty (tot.) includes statistical and systematic effects. The statistical uncertainty (stat.) is also shown, separately. The obtained results are compatible with the SM. The $t\bar{t}$ and $Z + \text{jets}$ normalisations are hardly scaled, being close to the SM prediction. In the case of $k_{t\bar{t}}$, the result is compatible with the SM within one standard deviation. Due to its small uncertainty, $k_{Z + \text{jets}}$ is not compatible with the SM within one standard deviation. Regarding μ_{tHq} , a negative factor

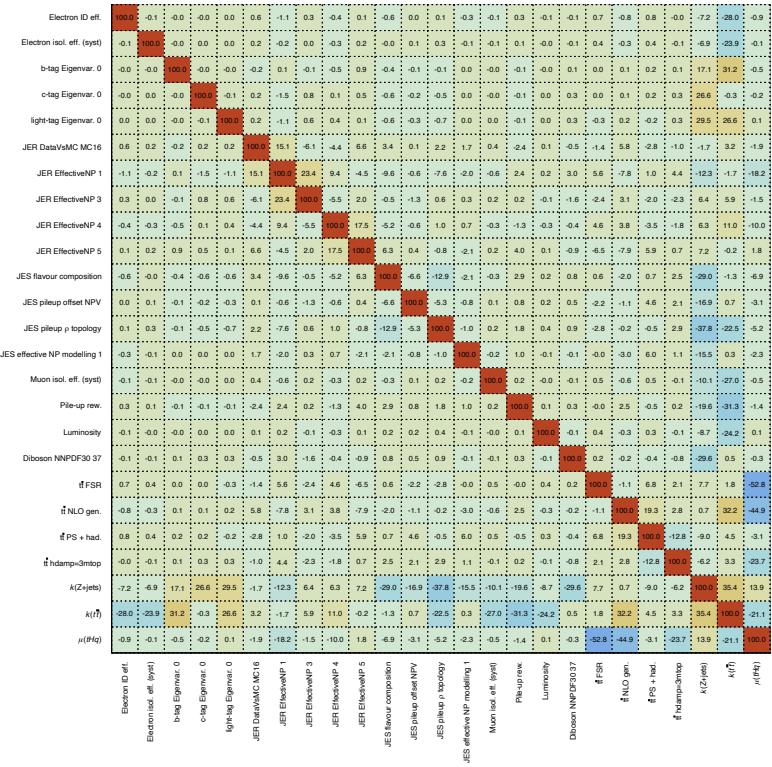


Figure 6.42: Correlation between the different NPs and the POIs in the $2\ell \text{OS} + 1\tau_{\text{had}}$ channel in the unblinded-full fit to data. Only the correlations of the NPs and normalisation factors with at least one correlation greater than 15% are shown. Correlations are represented by a positive number and anticorrelations by a negative one.

of -22.11 is applied to the tHq process. Of course, a negative factor and, hence, a negative number of events does not have any physical meaning. However, the uncertainty covers the compatibility with the SM within one standard deviation. The large statistical uncertainty in the μ_{tHq} result (the $^{+14.76}_{-13.72}$ in Equation 6.9) could be reduced if the statistical sample was larger and, hence, our understanding of the tHq normalisation would be better.

For the inverted Yukawa coupling hypothesis, the measured signal strength is:

$$\mu_{tHq, y_t = -y_t^{\text{SM}}} = -10.7^{+11.6}_{-39.4} (\text{tot.})^{+5.0}_{-4.8} (\text{stat.}).$$

The effect of the full-data fit in the $2\ell \text{OS}$ channel is presented in Figure 6.44 where the binned distributions are shown after the fit and can be compared to the pre-fit ones in Figure 6.29c. As expected, the χ^2 increases after the fit, meaning that the data/MC agreement has been improved. Also, note how the uncertainty bands are now thinner.

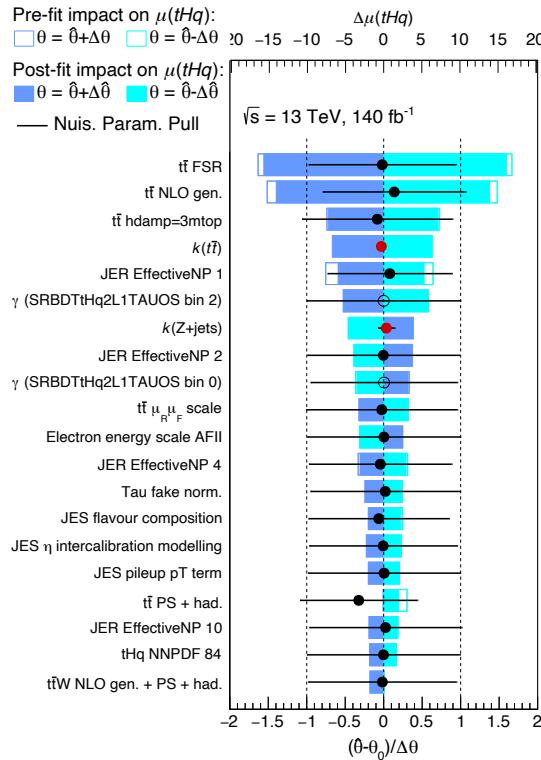


Figure 6.43: Ranking of the most impactful NPs on the fully-unblinded-data fit in the 2ℓ OS + $1\tau_{\text{had}}$ channel. The NPs are sorted by their impact on the μ_{tHq} in decreasing order. The blue and cyan boxes refer to the upper x-axis and show the impact on μ_{tHq} . The empty rectangles show the pre-fit impact and filled the post-fit. Moreover, the NPs values and their uncertainties are also included as dots and lines, respectively. The uncertainty of the NPs is measured with the lower x-axis.

In addition to the signal strength, an upper bound is also determined through a one-sided statistical test. This test is contingent upon the value of μ_{tHq} , and the upper limit is derived using the CLs method [308] to establish a confidence level of 95% (CL 95%). The interpretation of the signal strength, along with the upper limit, can also be articulated in terms of the production cross-section of the signal process as follows:

$$\sigma^{\text{obs}} = \mu_{tHq} \times \sigma^{\text{pred}},$$

where σ^{pred} is the value sign of the production cross-section predicted by the theory. For the 2ℓ OS + $1\tau_{\text{had}}$ channel, the expected and observed upper limits on the signal strength are presented in Table 6.33. This means that in only 5% of all statistical fluctuation cases, the signal strength would take a value higher or equal to 47.5 that is obtained.

Uncertainty source	Grouped impact
MC uncertainty	± 7.136
Modelling	
Theoretical uncertainties	± 24.041
PDF uncertainty	± 9.624
Experimental	
Instrumental	± 5.408
Flavour tagging	± 0.455
JES/JER	± 11.377
NormFactors	± 10.527
Total systematic uncertainty	± 28.543

Table 6.32: Impact of different groups of systematic uncertainties in the measurement of μ_{tHq} in the $2\ell \text{OS} + 1\tau_{\text{had}}$ channel with the profile-likelihood-binned-fit with unblinded data. The impact of each group of uncertainties is computed by performing a fit where the NPs in the group are fixed to their best-fit values, and then subtracting the resulting uncertainty on the μ_{tHq} in quadrature from the nominal fit. The line “MC uncertainty” refers to the statistical uncertainties in the MC events. The other theoretical uncertainties refer to the uncertainties on the cross-section, the ISR/FSR, alternative models, and the other elements described in Section 6.7.2. The total uncertainty is different from the sum in quadrature of the different components due to correlations between NPs built by the fit.

Observed	Expected	$1\sigma \text{ CL}_{95}$	$2\sigma \text{ CL}_{95}$
47.5	61.2	[41.95, 93.48]	[30.56, 148.6]

Table 6.33: Expected upper limit for the μ_{tHq} in the $2\ell \text{OS} + 1\tau_{\text{had}}$ channel. The third and fourth columns correspond to the 1σ and 2σ confidence interval on the expected upper limit.

To estimate how much the fit can constrain tHq cross-section under the $y_t = -y_t^{\text{SM}}$ hypothesis, the expected upper limit is calculated in Table 6.34. The observed limit obtained with the fit to the data is presented in the same table.

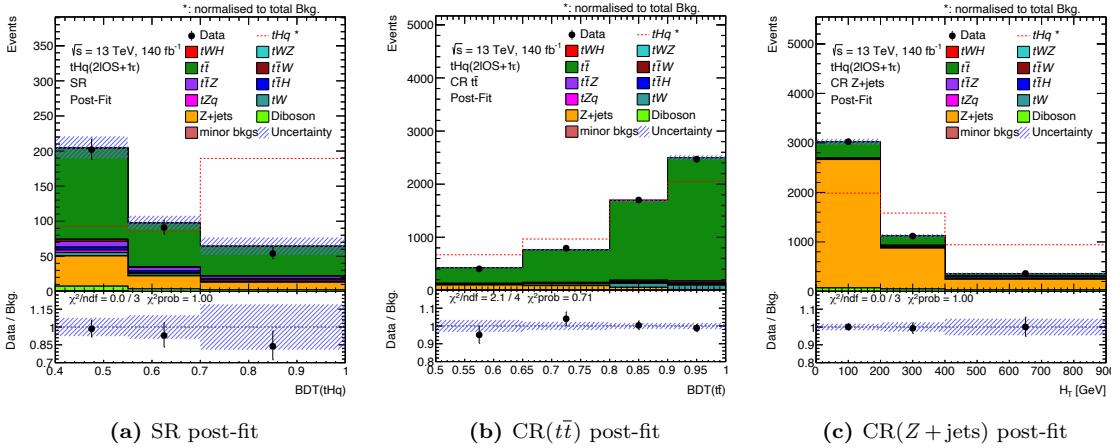


Figure 6.44: Binned distributions in the 2ℓ OS + $1\tau_{\text{had}}$ channel after the fully-unblinded fit. The real and simulated data events are shown using the following distributions: $\text{BDT}(tHq|_{OS})$ for the SR, $\text{BDT}(t\bar{t})$ for the $\text{CR}(t\bar{t})$, and H_T for the $\text{CR}(Z + \text{jets})$. The uncertainty bands include the statistical and all the systematic sources. The lower panels show the ratio between real and simulated background data events. The χ^2/ndf and the probabilistic χ^2 are included to measure the agreement between real and simulated data events.

Observed	Expected	$1\sigma \text{ CL}_{95}$
77.95	119.6	[46.97, 284.9]

Table 6.34: Expected upper limit for the μ_{tHq} in the 2ℓ OS + $1\tau_{\text{had}}$ channel under the inverted Yukawa coupling hypothesis. The third column correspond to the 1σ confidence interval.

6.8.7 Full-data fit results in the 2ℓ SS + $1\tau_{\text{had}}$ channel

In the same manner as for the 2ℓ OS + $1\tau_{\text{had}}$ channel in Section 6.8.6, the profile-likelihood-binned-fit to the 2ℓ SS + $1\tau_{\text{had}}$ channel with unblinded data is presented in this section. The limits to the tHq production derived in this section are more stringent since the 2ℓ SS + $1\tau_{\text{had}}$ channel is more sensitive than the other sub-channel.

6.8.7.1 Nuisance parameters in the 2ℓ SS + $1\tau_{\text{had}}$ full-data fit

When unblinding the data, some pulls are observed in the NPs (Figure 6.45) while no significant pulls take place for the γ factors (Figure 6.46). The largest pulls are one of the JES effective NPs and the one accounting for the generation of the PS and hadronistaion of the $t\bar{t}$ process. The most significant pulls according to Equation 6.8 are presented in Figure 6.47. Three out of the four most relevant

pulls correspond to the τ -energy scale. This parameters refer to the calibration of the energy of the τ_{had} .

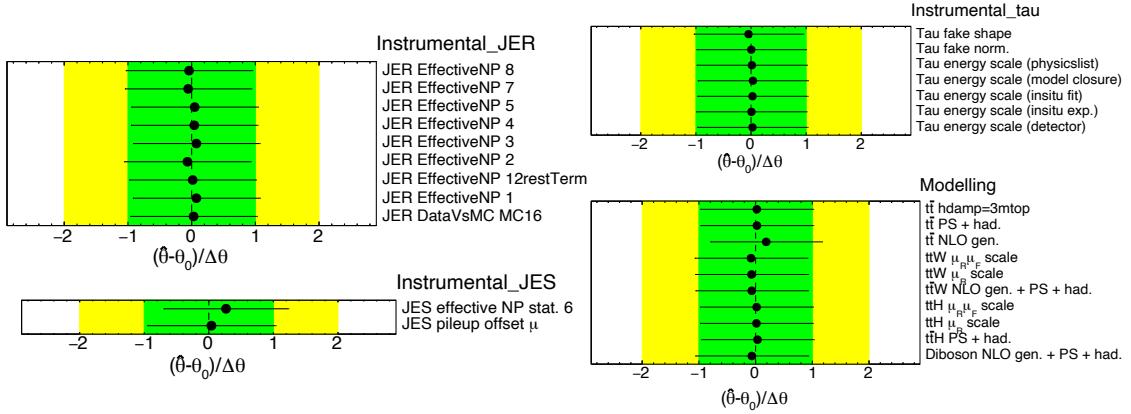


Figure 6.45: Pulled NPs in the unblinded-data-full fit of the $2\ell \text{SS} + 1\tau_{\text{had}}$. The rest of NPs do not present any significant pull or constraint from its nominal value.

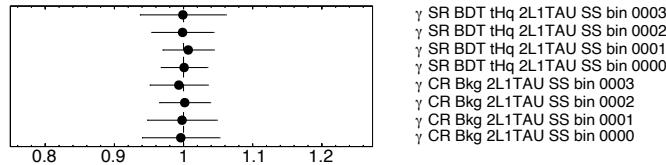


Figure 6.46: The γ s of all bins used in the unblinded-data-full fit of $2\ell \text{SS} + 1\tau_{\text{had}}$ channel. The γ s are associated to the bins in the SR from left to right and numbered starting with 0.

The correlations in the unblinded fit are presented in Figure 6.48. While none of the elements of this matrix is large, the most relevant of these is the anticorrelation of 34.4% between the $k_{t\bar{t}, t\bar{t}X}$ and $\mu_{tHq}^{2\ell \text{SS} + 1\tau_{\text{had}}}$. This leads to the most highly ranked impacted on the derivation of the μ_{tHq} (see Figure 6.49). The impact of the systematic uncertainties pretend in groups is shown in Table 6.35, where it can be seen that the largest groups of systematics are the JES and JER ones as well as the uncertainty from the $k_{t\bar{t}, t\bar{t}X}$. It is worth to mention that a bug in the implementation of the JER systematic uncertainties which results in an over-estimation of the uncertainty in certain regions of the phase space was found recently. The cause of this is that the same seed is used to smear jets in MC when correcting for Data/MC differences in the nominal calibration and when applying the uncertainty in one direction. This can account for the JER NPs impact in the fit.

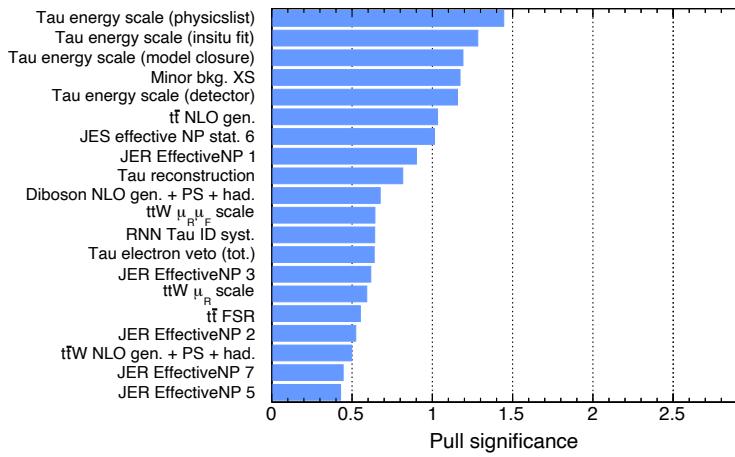


Figure 6.47: The 20 largest pull significances for the $2\ell \text{SS} + 1\tau_{\text{had}}$ in an ordered manner, with the most significant pull shown first.

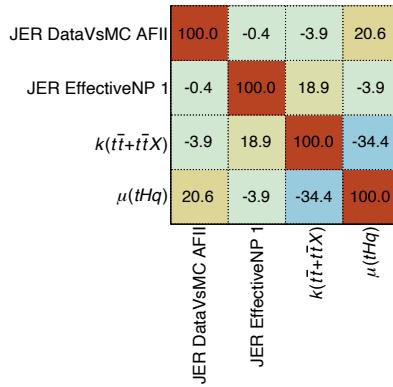


Figure 6.48: Correlation between the different NPs and the POIs in the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel in the unblinded-full fit to data. Only the correlations of the NPs and normalisation factors with at least one correlation greater than 15% are shown. Correlations are represented by a positive number and anticorrelations by a negative one.

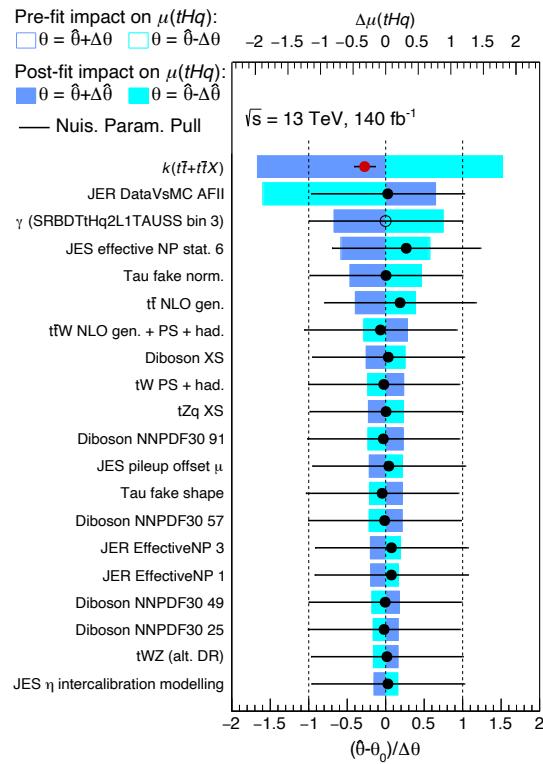


Figure 6.49: Ranking of the most impactful NPs on the fully-unblinded-data fit in the 2ℓ SS + $1\tau_{\text{had}}$ channel. The NPs are sorted by their impact on the μ_{tHq} in decreasing order. The blue and cyan boxes refer to the upper x-axis and show the impact on μ_{tHq} . The empty rectangles show the pre-fit impact and filled the post-fit. Moreover, the NPs values and their uncertainties are also included as dots and lines, respectively. The uncertainty of the NPs is measured with the lower x-axis.

Uncertainty source	Grouped impact
MC uncertainty	± 0.983
Modelling	
Theoretical uncertainties	± 1.105
PDF uncertainty	± 0.817
Experimental	
Instrumental	± 0.716
Flavour tagging	± 0.126
JES/JER	± 2.727
NormFactors	± 2.235
Total systematic uncertainty	± 4.081

Table 6.35: Impact of different groups of systematic uncertainties in the measurement of μ_{tHq} in the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel with the profile-likelihood-binned-fit with unblinded data. The impact of each group of uncertainties is computed by performing a fit where the NPs in the group are fixed to their best-fit values, and then subtracting the resulting uncertainty on the μ_{tHq} in quadrature from the nominal fit. The line “MC uncertainty” refers to the statistical uncertainties in the MC events. The other theoretical uncertainties refer to the uncertainties on the cross-section, the ISR/FSR, alternative models, and the other elements described in Section 6.7.2. The total uncertainty is different from the sum in quadrature of the different components due to correlations between NPs built by the fit.

6.8.7.2 Results of the $2\ell \text{SS} + 1\tau_{\text{had}}$ full-data fit

The signal strength and $k_{t\bar{t},t\bar{t}X}$ obtained as a result of the profile-likelihood-binned fit in the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel are:

$$\mu_{tHq} = -2.67 \pm 5.44(\text{tot.}) \pm 5.03(\text{stat.}) \quad (6.12)$$

$$k_{t\bar{t},t\bar{t}X} = 0.73 \pm 0.14(\text{tot.}) \pm 0.10(\text{stat.}) . \quad (6.13)$$

The $\mu_{tHq}^{2\ell \text{SS} + 1\tau_{\text{had}}}$ is scaled to a negative value. While a negative μ or k is clearly an unphysical solution, when taking into account the uncertainty, $\mu_{tHq}^{2\ell \text{SS} + 1\tau_{\text{had}}}$ is compatible with the SM prediction within one standard deviation. Another important point in the result of Equation 6.12 is that the leading uncertainty is the statistical one. Therefore, this channel could strongly benefit from and increase in the luminosity. For the normalisation factor of the backgrounds with a pair of top quarks (Equation 6.13), the SM prediction is at 1.93σ from the SM prediction. It is believed that the origin of this discrepancy comes from the scale factors correcting the misidentification rates (see Section 6.5). After the application of the template fit method, there is still a mismodelling of the distributions, being incompatible with the SM. There are two ways to fix this scenario. First, the application of the template fit method has to be improved to better describe the data. Second, the systematic uncertainties corresponding to the estimation of the fake rates should be able to account for the observed discrepancies. It must be also taken into account that here is a mild tension between the theory predictions on the $t\bar{t}W$ production and the experimental results (see Figure 2.1). Both ATLAS and CMS have shown some over-abundance in data compared to theoretical predictions [309, 310]. Since $t\bar{t}W$ is the most dominant background in the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel, the mentioned tension may play a role in the lack of agreement between data and MC.

For the inverted Yukawa coupling hypothesis, the measured signal strength is:

$$\mu_{tHq,y_t=-y_t^{\text{SM}}} = -0.3 \pm 1.2(\text{tot.}) \pm 1.1(\text{stat.}) .$$

Both the observed and expected upper limit for the $\mu_{tHq}^{2\ell \text{SS} + 1\tau_{\text{had}}}$ are obtained and presented in Table 6.36.

Regarding the limits on the inverted Yukawa coupling hypothesis, the $1\sigma \text{CL}_{95}$ and $2\sigma \text{CL}_{95}$ are calculated in Table 6.37.

Figure 6.50 presents the fitted distributions of the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel after performing the full-data fit. These can be compared to the prefit ones in Figure 6.30b. The simulated background is scaled by a factor $k_{t\bar{t},t\bar{t}X}$ in the post-fit distributions and, hence, the data/MC agreement improves. This is reflected by the improved χ^2 .

Observed	Expected	1σ CL ₉₅	2σ CL ₉₅
13.83	16.94	[10.33, 30.22]	[6.991, 52.24]

Table 6.36: Expected upper limit for the μ_{tHq} in the 2ℓ SS + $1\tau_{\text{had}}$ channel. The third and fourth correspond to the 1σ and 2σ confidence interval.

Observed	Expected	1σ CL ₉₅	2σ CL ₉₅
20.79	24.07	[7.478, 112.5]	[3.197, 342.4]

Table 6.37: Expected upper limit for the μ_{tHq} in the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel under the inverted Yukawa coupling hypothesis. The third and fourth columns correspond to the 1σ and 2σ confidence interval, respectively.

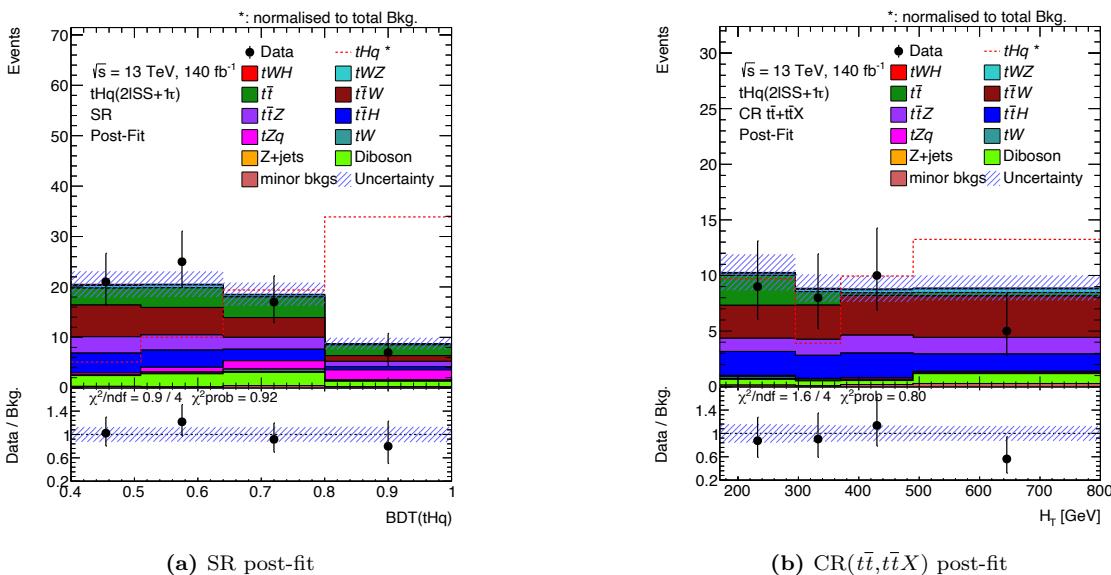


Figure 6.50: Binned distributions in the 2ℓ SS + $1\tau_{\text{had}}$ channel after the fully-unblinded fit. The real and simulated data events are shown using the following distributions: $\text{BDT}(tHq|_{\text{SS}})$ for the SR, and H_T for the CR($t\bar{t}, t\bar{t}X$). The uncertainty bands include the statistical and all the systematic sources. The lower panels show the ratio between real and simulated background data events. The χ^2/ndf and the probabilistic χ^2 are included to measure the agreement between real and simulated data events.

Chapter 7

Conclusion

Cuando el río es lento y se cuenta con una buena bicicleta o caballo sí es posible bañarse dos (y hasta tres, de acuerdo con las necesidades higiénicas de cada quién) veces en el mismo río.

—AUGUSTO MONTERROSO,
HERACLITANA (1978)

This thesis presents the study to measure the direct production of a Higgs boson in association with a single-top quark, focusing on final states with two light-flavour leptons and one hadronically-decaying- τ lepton, employing the ATLAS detector. Attending to the relative charge between the light lepton, this investigation is divided into two channels: $2\ell \text{ OS} + 1\tau_{\text{had}}$ and $2\ell \text{ SS} + 1\tau_{\text{had}}$.

The search for such a rare process is motivated by the intricate interplay between two fundamental particles: the Higgs boson and the top quark. On the one hand, the Higgs boson plays a critical role in our understanding of mass acquisition by particles through the Spontaneous Symmetry Breaking mechanism. On the other hand, the top quark is notable for being the most massive particle in the SM and the only one that decays before its hadronisation. Therefore, the Yukawa coupling between these two particles is expected to be the largest in the SM and it can be measured through this interaction. This measurement is central to the experimental

program of the LHC and it could hint at possible CP violation, influencing the tHq production cross-section.

In this thesis the theoretical foundation of physics of the top quark and Higgs boson are discussed, a review of the ATLAS detector and its performance is given, and the simulation chain and object reconstruction are described. Then, the search for the tHq production is carefully detailed.

This analysis uses proton-proton collisions at $\sqrt{s} = 13$ TeV from the ATLAS detector during Run 2 of the LHC with a total integrated luminosity of 140 fb^{-1} . The parton-level information is implemented and used to reconstruct the tHq process. The origin of the light-lepton is assessed via the use of BDT. Then the rate of misidentified particles is addressed using the template fit method to correct the MC yields. Afterwards, by using several BDTs, the SRs and CRs of the analysis are defined. Utilising these regions, a profile-likelihood-binned fit with Asimov data has been used to determine the sensitivity of the analysis. Then, the real data has been incorporated to derive the normalisation factors of the $t\bar{t}$ and $Z + \text{jets}$ productions in the $2\ell \text{ OS} + 1\tau_{\text{had}}$ channel. A normalisation factor has been derived as well for $t\bar{t}$, $t\bar{t}W$, $t\bar{t}H$ and $t\bar{t}Z$ all together in the $2\ell \text{ SS} + 1\tau_{\text{had}}$ channel. The profile-likelihood-binned fit with the entire dataset has been used to determine the signal strength of the tHq production in each channel. Figure 7.1 presents the results of the direct search for the tHq production for both the SM and the inverted Yukawa coupling. These results on μ_{tHq} are fully compatible with the SM. In regards to the inverted Yukawa coupling, the findings from our analysis do not provide sufficient grounds to reject this hypothesis outright. The statistical approach employed to assess the signal strength for the alternative cross-section may not be the most effective for testing this hypothesis, particularly because the negative μ_{tHq} ($y_t = -y_t^{\text{SM}}$) represents an unphysical outcome. Nevertheless, the observed deviation of approximately -1σ from the expected result under this hypothesis does not support the μ_{tHq} ($y_t = -y_t^{\text{SM}}$) scenario. The consistent preference for a cross-section lower than anticipated across both sub-channels further solidifies that this scenario is disfavoured.

The culmination of this analysis is the determination of limits on the tHq production cross-section. The observed and expected upper limits on the tHq cross-section are in Table 7.1 and Figure 7.2. For the expected upper limit the 1σ and 2σ variations are given. Both observed limits are compatible with the inverted Yukawa coupling hypothesis. The direct comparison with the results presented in Table 2.1 is complex since those target tH and this thesis tHq . For instance, the most recent exclusion limit [105] is similar to the one obtained in the $2\ell \text{ SS} + 1\tau_{\text{had}}$ channel. The Reference [105] limit_{95CL} is $14.6(\text{obs})19.3^{+9.2}_{-6.0}(\text{exp})$ while the one in Table 7.1 is $13.83(\text{obs})16.94^{+13.28}_{-6.61}(\text{exp})$.

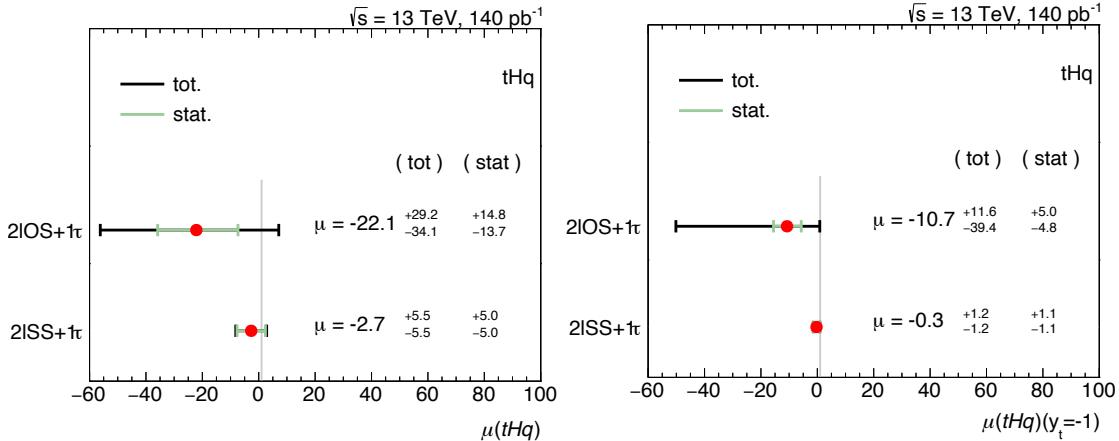


Figure 7.1: Signal strength values for the both the $2\ell OS + 1\tau_{had}$ and the $2\ell SS + 1\tau_{had}$ channels under (a) the SM and (b) the inverted Yukawa coupling scenarios. The total uncertainty (tot) includes statistical and systematic effects. The statistical uncertainty (stat) is also shown, separately.

μ_{tHq} upper limit				
Channel	Observed	Expected	1σ CL ₉₅	2σ CL ₉₅
$2\ell OS + 1\tau_{had}$	47.51	61.2	[41.95, 93.48]	[30.56, 148.6]
$2\ell SS + 1\tau_{had}$	13.83	16.94	[10.33, 30.22]	[6.991, 52.24]
$2\ell OS + 1\tau_{had} (y_t = -y_t^{\text{SM}})$	77.95	119.6	[46.97, 284.9]	—
$2\ell SS + 1\tau_{had} (y_t = -y_t^{\text{SM}})$	20.79	24.07	[7.478, 112.5]	[3.197, 342.4]

Table 7.1: Observed and expected upper limits for the μ_{tHq} for both $2\ell + 1\tau_{had}$ channels derived with the CL method. For completion, limits under the inverted Yukawa coupling hypothesis are included as well.

Looking ahead, the prospects for enhancing the outcomes of this analysis are promising, driven by several anticipated developments:

- The upcoming high-luminosity Run 3 promises increased statistical significance of the analysis. This will be very beneficial since in the $2\ell SS + 1\tau_{had}$ channel (the one with the best sensitivity) the statistical uncertainty is the dominant. In the $2\ell OS + 1\tau_{had}$ the uncertainty due to the statistical sample is similar to the systematic uncertainty and more data events will undoubtedly enrich this search.
- The current techniques to estimate the rates of backgrounds due to misidentified objects will also benefit from an increase in the statistical sample since these are data-driven methods.

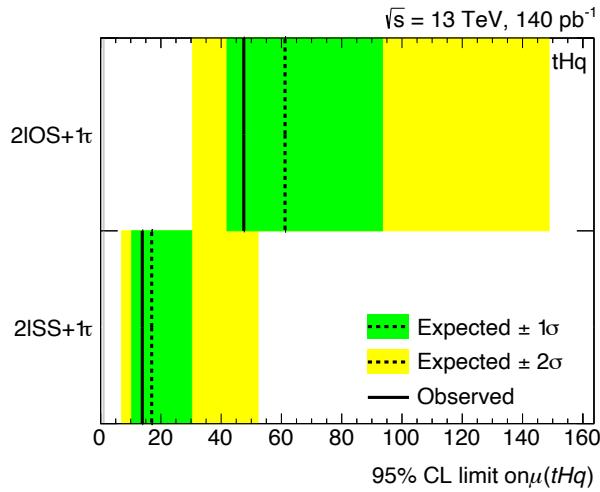


Figure 7.2: Observed and expected upper limits for the μ_{tHq} for both $2\ell + 1\tau_{\text{had}}$ channels. The green and yellow areas represent the 1σ (68% CL) and 2σ (95% CL) variations of the expected upper limit.

- As already mentioned, a significant challenge in the analysis presented in this thesis stems from the lack of agreement observed between the actual data and the MC-event samples. These discrepancies are attributed to limitations in the application of the template-fit method. Although the caveat in the implementation of the template is identified, a solution to it is not available at the time of composing this thesis. Concerning this obstacle, it may be beneficial to explore alternative methods to estimate the rates of the backgrounds based on misidentified objects.
- With more statistics we could explore splitting the regions used in the fit calculations according to the track multiplicity of the τ_{had} (1-prong or 3-prong). Since jets with different origins mimic differently the τ_{had} depending on its prongness, it can be beneficial to explore this classification when defining the regions used in the profiled-likelihood fit.
- The impact of the simulated number of events for certain samples is unfortunate. An improvement is expected from incorporating newer samples which have been simulated with more statistics. While the effect of low MC statistics could be mitigated more sensitive bins with less sensitive bins, after some tests, it is seen that the sensitivity of the analysis decreased by doing so.
- Currently, there is a mild tension between the simulation and the experimental results for $t\bar{t}W$. Therefore, progress in the simulation techniques or a better understanding of these processes will further refine the analysis. This

will be useful in the 2ℓ SS + $1\tau_{\text{had}}$ channel, where this is the second most important background.

Additionally, this analysis can be extended in several ways. First, the region definition can be improved by combining the BDTs described in this thesis with the NNs that are also being used by the ongoing ATLAS analysis. Secondly, the measurement of the tHq production in the $2\ell + 1\tau_{\text{had}}$ channel can be combined with the measurements in other channels. From the two sub-channels explored in this thesis, the 2ℓ SS + $1\tau_{\text{had}}$ one is the one that can add more sensitivity to future combinations with other tHq channels. Finally, incorporating the tWH processes in the signal would allow to perform a more complete tH search and, hence, the inverted-top-Yukawa-coupling hypothesis can be properly tested. This work is currently being done, although it is not included in this thesis. The tools and methods developed to analyse the tHq productions are the same that are employed for the inclusive tH measurement.

Appendix A

Details about the parton-level simulations

A.1 Software package for parton-level information

The TopPartons package performs the analysis of the parton-level truth information. For all particles except the ones in the final state, the information stored in the NTuples corresponds to the after FSR. For the final state particles, the ISR information is saved. For each particle, the PDG-ID, p_T , η , ϕ and mass are stored. Additionally, some of the τ related variables contain information on whether τ decays hadronically or leptonically. If the τ decays to a light lepton, it stores the PDG-ID of that lepton and if it decays hadronically it stores the PDG-ID of the W^- for τ^- and the W^+ for τ^+ .

Decay of the Higgs boson and the top quark

First of all, the code searches for the Higgs boson by looking at the PDG-ID of the truth particles. Then, it takes into account all the considered Higgs-boson decays and demands that it has exactly two children. The FSR state of the Higgs boson is stored and its children are studied. If the Higgs boson does not decay into $W^+ W^-$, $\tau^- \tau^+$ or ZZ , the event is discarded. To fill the truth variables it is also required that the top quark decays into Wb .

Spectator-quark ambiguity at NLO

fig:tHq:intro:diltauFeynmanDiagram The algorithm identifies the spectator quark,

i.e. the quark that comes from the hard-scattering process. At NLO, the spectator quark is not well defined from first principles at the parton level and, therefore, there would be no way that one can do a proper parton assignment with NLO generators.

To properly select the spectator light-quark, firstly the top quark is searched and then its parent found¹. The spectator quark is selected at parton level from among the light quarks after the QED and QCD radiations that are not products of the top-quark decay. In case of ambiguities arising from ISR, the spectator quark that minimises the p_T of the combined spectator-quark and top-quark system is chosen.

***b* quarks**

Afterwards, the second *b*-quark is identified as the one whose parent is a gluon since it is originated from the gluon splitting. The *b*-quark originated from the top-quark-decay system is saved as well. In Figure 6.14 the p_T and η distributions of these quarks are compared using the truth-level information.

Further decays

Finally, all the possible decays from the Higgs-boson children are considered. These are further decayed several times to explore all possible final states. The range of possible combinations is wide and the particular configuration that defines a final-state channel can be achieved in more than one way.

A.2 BR-based calculation for the tHq fractions

On the other side, the BR-based calculations are used to determine the fraction of each decay channel that is present in the $2\ell + 1\tau_{\text{had}}$. This predictions should be match the fractions obtained with the `TopPartons` package. This calculations are performed to validate the code not only for the $2\ell + 1\tau_{\text{had}}$ channels but also for the 3ℓ . By doing so, even a further confirmation of the proper functioning of the code is achieved. For the calculations, the BRs of the three Higgs-boson-decay modes that are considered are presented in Table A.1 [10]:

¹Note that in PYTHIA the intermediate quarks are left out so the parent of the top quark would be the gluon and up quark in the diagram of Figure 6.1. Therefore, the children of the gluon are the *b*-quark, the top quark, and spectator quark.

Decay	BR(%)
$H \rightarrow \tau\tau$	56.65
$H \rightarrow WW^*$	5.12
$H \rightarrow ZZ^*$	37.43

Table A.1: Probability of the Higgs boson to decay to each of its three considered decay channels. The percentages are normalised to 100%.

The decay products of each pair of Higgs-boson children are combined and the total decay fractions of these combinations are computed. To do so, decay ratios presented in Table A.2 are used [10]:

τ	BR(%)	W	BR(%)	Z	BR(%)
$\tau \rightarrow e\nu\nu$	14.82	$W \rightarrow e\nu$	10.71	$Z \rightarrow ee$	3.36
$\tau \rightarrow \mu\nu\nu$	14.39	$W \rightarrow \mu\nu$	10.63	$Z \rightarrow \mu\mu$	3.36
$\tau \rightarrow \text{Hadrons}$	64.74	$W \rightarrow \tau\nu$	11.38	$Z \rightarrow \tau\tau$	3.36
		$W \rightarrow \text{Hadrons}$	67.41	$Z \rightarrow \nu\nu$	20
				$Z \rightarrow \text{Hadrons}$	69.91

Table A.2: BR of each of the τ lepton, W and Z bosons. The percentages of each column are normalised to 100%. If the W or Z boson decay to a τ , they are further decayed into either a e/μ or τ_{had} .

For each of the Higgs-boson-decay modes, the possible final states are studied and its correspondent fractions are calculated in Table A.3. In the $H \rightarrow \tau\tau$ column can be seen that the $\tau_{\text{had}} + \tau_{\text{lep}}$ signature is more probable (45.6%) than the $\tau_{\text{lep}} + \tau_{\text{had}}$ one (12.4%). For $H \rightarrow WW^*$, the first two rows take part in the $2\ell + 1\tau_{\text{had}}$ signature when the τ_{had} comes from the t quark in the first item and the H boson in the second. For the ZZ^* , the $ZZ^* \rightarrow 2 \times e/\mu$ contributes to the $2\ell + 1\tau_{\text{had}}$ signature when the τ_{had} is produced in the top-quark system. When the top quark decays into a light lepton, the $ZZ^* \rightarrow e/\mu + \tau_{\text{had}}$ mode contributes to the $2\ell + 1\tau_{\text{had}}$ final state.

For the top-quark system only the $t \rightarrow W + b$ mode is taken into account and the W boson is further decayed until there are either hadrons, light leptons or a τ_{had} . Except when the W boson decays directly into hadrons, all modes contribute to the $2\ell + 1\tau_{\text{had}}$ channel.

Combining all the decay modes presented brings out a wide variety of final states with different probabilities. For the $2\ell + 1\tau_{\text{had}}$ final state, these results are summarised in Table 6.9. Additionally, from these calculations can be deduced that from all tHq events, only a 3.72% decay into a $2\ell + 1\tau_{\text{had}}$ final state. From

$H \rightarrow \tau\tau$	BR(%)	$H \rightarrow WW^*$	BR(%)	$H \rightarrow ZZ$	BR(%)
$\tau^+\tau^- \rightarrow e/\mu + e/\mu$	12.51	$WW^* \rightarrow e/\mu + e/\mu$	6.42	$ZZ^* \rightarrow 4 \times e/\mu$	0.52
$\tau^+\tau^- \rightarrow e/\mu + \tau_{\text{had}}$	45.63	$WW^* \rightarrow e/\mu + \tau_{\text{had}}$	3.73	$ZZ^* \rightarrow 2 \times e/\mu$	12.85
$\tau^+\tau^- \rightarrow \text{Hadrons}$	41.51	$WW^* \rightarrow e/\mu + \text{Hadrons}$	34.17	$ZZ^* \rightarrow e/\mu + \tau_{\text{had}}$	12.85
		$WW^* \rightarrow \text{Hadrons}$	55.92	$ZZ^* \rightarrow 4 \times \nu + \text{Hadrons}$	80.82
				$ZZ^* \rightarrow 2 \times \tau_{\text{had}}$	2.82
				$ZZ^* \rightarrow 3 \times e/\mu + \tau_{\text{had}}$	0.22

Table A.3: Probability of each final-state of the Higgs system. The probabilities of each column are normalised to 100%. This is calculated from the numbers in Table A.2.

these, in more than 80% of cases the τ_{had} is produced in the Higgs-boson-decay chain.

Appendix B

Boosted Decision Trees

A boosted decision tree, typically referred by its acronym BDT, is a supervised¹ machine learning (ML) technique used for classification. The analysis presented in this thesis uses several BDTs. Both the light-lepton-origin assignment (Section 6.4.2) and the processes separation (Section 6.6.2) are based on BDTs. This tool is applied in more scenarios within ATLAS. In the b -tagging, for instance, a BDT is trained to discriminate b -jets from light-jets [229]. During the development of this thesis, two software packages have been used to develop BDTs: TMVA [284] of ROOT [283] and XGBoost [293].

Section B.1 of this appendix explores the fundamental principles, training procedures, evaluation metrics, and practical applications of BDTs. While in Section B.4, the k -folding method for cross-validation is described. Section B.5 carefully describes the hyperparameters that control the training procedure. More details about the metrics and additional considerations are given in Section B.6. The information on the different BDTs presented in Chapter 6 is complemented in Section B.7. In this section, the metrics of the different folds are presented separately, the matrices displaying the linear correlation among variables are shown, and the evolution plots of the different training are introduced.

B.1 How does a BDT work?

A BDT is an ensemble of decision trees. Each decision tree is a map of possible results of related decisions. A decision tree takes a set of input features and splits input data recursively based on these features. This results in a tree structure

¹Supervised learning means that the data used in the training is labelled.

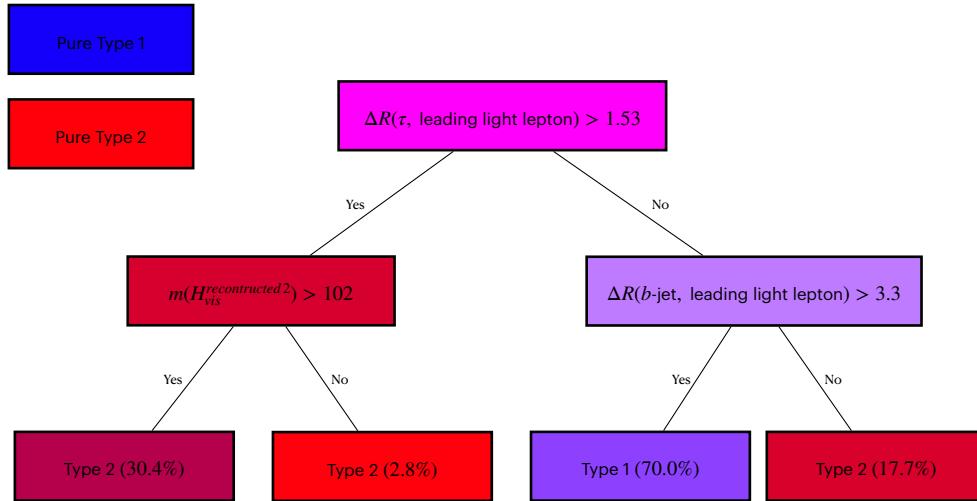


Figure B.1: Example of a decision tree with three nodes. This particular example corresponds to one of the trees in the BDT for the light-lepton-origin assignment (see Section 6.4.2). The colour of the boxes represents the purity of Type 1 or Type 2 events that arrive at each node. Repeated left/right (yes/no) decisions are taken on one single variable at a time until the classification takes place.

that resembles that of a flow chart with a decision or split at each node. The last level of the trees are the so called *leaves* (also known as terminal nodes) and each represents a class label. An example of a single tree can be seen in Figure B.1, where an event is classified in one of the two categories following a set of yes/no questions. In this work the BDTs employed are binary, i.e. they separate the dataset into two categories, but multiclassifier BDTs could be used as well. In fact for the region definition, multiclassifier BDTs have been tested but, since the result was not satisfactory, these are not used.

Boosting is a technique for turning numerous weak classifiers (trees in this case) into a powerful one. Each tree is created iteratively depending on the prior ones. The output of each tree, denoted as $h_t(\mathbf{x})$, is assigned a weight, w_t , proportional to its accuracy. The ensemble output or prediction is the weighted sum:

$$\hat{y}(\mathbf{x}) = \sum_t w_t h_t(\mathbf{x})$$

where t run over the trees. The \hat{y} is the model prediction of the true label, y . The goal of the boosting is to minimise a regularised objective function:

$$L(x) = \sum_i l(\hat{y}_i, y_i) + \sum_t \Omega(f_t) \quad (\text{B.1})$$

where $l(\hat{y}_i, y_i) = l(f(\mathbf{x}_i|\theta), y_i)$ is a differentiable convex loss function which measures the distance between the true label (y_i) and its prediction (\hat{y}_i) on the i^{th} sample. The term $\Omega(f_t)$ corresponds to the regularisation function, which penalises the complexity of the t^{th} tree, f_t . The θ in $f(\mathbf{x}_i|\theta)$ are the model parameters. For a BDT, θ would be the weights and biases. The \mathbf{x}_i are values for the input variables for the i^{th} sample, while y_i denotes the target variable real value. The regularisation term $\Omega(f_t)$ term helps to smooth the final learnt weights to avoid over-fitting.

The tree ensemble model in Equation B.1 cannot be optimised using traditional optimisation methods in Euclidean space. Instead, the model is trained in an additive way so that the objective function to minimise is:

$$L^{(t)} = \sum_i^n l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t). \quad (\text{B.2})$$

There are several types of boosting for BDTs. Some of the most common are AdaBoost, Gradient Boosting and XGBoost. The latter, which stands for “eXtreme Gradient Boosting”, is used for region definition and its details can be found in reference [293]. Boosting can significantly improve performance compared to that of a single tree and stabilise the response of the decision trees to fluctuations in the training sample.

B.1.1 Training

For a supervised ML algorithm, the process of training means to learn or determine optimal values for all the weights within the model. To achieve this, the algorithm takes the labelled data and fits the model. In the case of signal discrimination, for instance, the ML model takes the MC samples, where all the events are labelled. Within the data, the model also requires a set of variables that have some power to discriminate between the selected categories. The variables used for training are referred as input features. A condition on one discriminant variable is set on each node of the BDT to split the phase space into two parts. The training aims to find the optimal cut in each node so that after it the separation between the categories is maximised. This is done in a loop over all discriminating variables and trying to test as many as possible values for each cut (the default in TMVA is trying 20 values for each variable). The best cut is defined as the one that yields the highest splitting index difference between the parent node and the two child nodes (each weighted by the total number of events in the corresponding block). The splitting process continues until a stopping requirement is met.

Internal reweighting of events in the training sample

Occasionally, MC generators may provide event weights which may turn out to be extremely small or very high. To avoid artefacts, TMVA can renormalise the signal and background training weights internally so that their respective sums of effective (weighted) events are equal. By doing this, the performance of the BDT can be improved since some classifiers are sensitive to the relative amount of each category (Type 1/Type 2 or signal/background) in the training data. While for the lepton assignment, this renormalisation does not play an important role (the amount of Type 1 and Type 2 signal events is similar), for the tHq signal discrimination the signal sample in the training test requires reweighting.

Another case of internal reweighting is when negatively-weighted events are present in the training sample and the absolute value of the weight is used instead. More details about the problems associated with negatively-weighted events in the training sample of a BDT can be found in Section B.3.

B.1.1.1 Loss functions

The loss function, also referred to as the error function, is used to evaluate the quality of predictions by measuring the deviation between estimated values and their corresponding true values during a specific iteration of the model. It penalises prediction errors and plays a crucial role in supervised ML models.

TMVA loss function

The TMVA framework offers various options for loss functions that can be used for training and evaluating ML models. The choice of the loss function depends on the specific task and the nature of the problem being addressed. When utilising gradient boost, one commonly used loss function in TMVA is the squared-error loss, which measures the squared difference between the predicted value and the true value. It can be expressed as: $l(\hat{y}, y) = (\hat{y} - y)^2$.

For binary classification tasks, TMVA provides different loss functions that encapsulate the notion of misclassification. The most popular boosting method, AdaBoost, is based on exponential loss, $l(\hat{y}, y) = e^{(-\hat{y}y)}$ [311]. However, exponential loss lacks robustness in the presence of outliers or mislabelled data points, causing the performance of AdaBoost to degrade in noisy settings.

To address this weakness, the GradientBoost algorithm in TMVA allows for the use of other potentially more robust loss functions while maintaining the good out-of-the-box performance of AdaBoost. In TMVA, the implementation of Gradi-

entBoost for classification, the binomial log-likelihood loss function is employed:

$$l(\hat{y}, y) = \ln \left(1 + e^{-2\hat{y}y} \right).$$

XGBoost loss functions

The XGBoost library provides various loss functions that can be used for different purposes. For binary classification, as is the case in the BDTs presented in Section 6.6.2, the `logistic` loss function has been used. However, there are other available options, such as `logitraw` and `hinge`. These different loss functions allow flexibility in choosing the appropriate approach based on the specific requirements of the classification problem. For the tests involving multiclassifier BDTs, the `softprob` loss function was employed.

For the `logistic` loss function in XGBoost, the l term in Equation B.2 represents the logarithmic likelihood of the Bernoulli distribution. It can be formulated as follows:

$$l = y_i \log[\text{logistic}(\hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i))] + (1 - y_i) \log[1 - \text{logistic}(\hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i))]$$

where $\text{logistic}(\hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i))$ is the probability for the category. In an algebraically equivalent manner, it can be written as:

$$l = y_i[\hat{y}_i^{t-1} + f_t(\mathbf{x}_i)] - \log(1 + \exp(\hat{y}_i^{t-1} + f_t(\mathbf{x}_i)))$$

B.1.1.2 Overtraining

Being $f(\mathbf{x}|\theta)$ a ML model where x are the data points used as input and θ the tuneable parameters of the model. The function $f(\mathbf{x}|\theta)$ outputs the prediction of the model. The parameters θ of the model are tuned during the training process using a training set (\mathcal{T}). The true output (y) of the elements in \mathcal{T} . When successful, the training finds the θ that performs as good as possible on new, unseen, data.

For a given $f(\mathbf{x}|\theta)$ model, the training error, $\text{err}(\mathcal{T})$, is defined by [312]:

$$\text{err}(\mathcal{T}) = \frac{1}{N_t} \sum_{n=1}^{N_t} l(y_n, f(\mathbf{x}_n|\theta)) \quad (\text{B.3})$$

where N_t is the number of events used for the training and l the chosen loss function and \mathbf{x}_n and y_n are the features and label points in the training set. So, the error function measures the error of the model on a group of objects, whereas the loss function deals with a single data instance.

The $\text{err}(\mathcal{T})$ usually decreases as the number of training cycles increases, and it can begin to adapt to noise in the training data. When this happens the training error continues to decrease but the error on the data outside of the training set starts increasing, jeopardising the general performance of the model. This effect is the so called overfitting or overtraining.

In other words, overtraining occurs when a ML model can accurately predict training examples but is unable to generalise² to new data. When overtraining takes place, the ML model has learnt the details of the training data to an extent in which this knowledge do not reflect the behaviour of the test sample. This results in poor field performance.

Figure B.2 shows how an overtrained BDT evolves. In Figure B.2a can be seen that as the training of the BDT continues, the ability of the model to classify the events in \mathcal{T} (blue) improves while for the data in the test sample (orange), it does not. This means that the model is not generalising properly. With the plot of the loss function (Figure B.2b) can be seen how the error of the test data slightly increases while for the training samples is strongly reduced. These two evolution plots can be compared to the equivalent ones presented in Section B.7.4, where there is no overtrain.

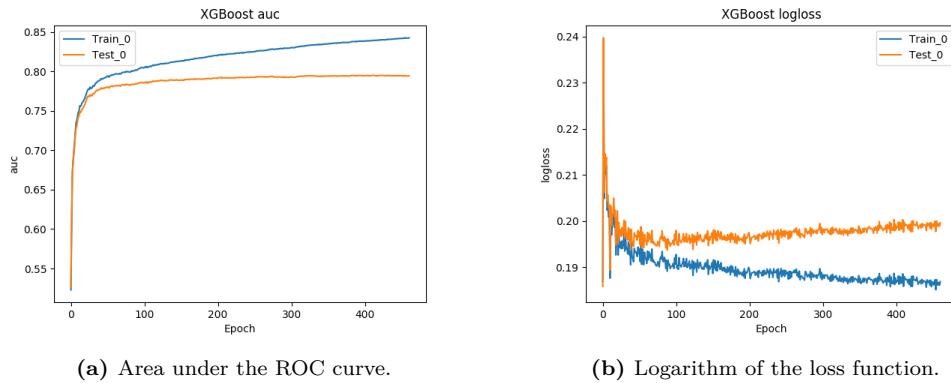


Figure B.2: Example of the evolution of the BDT metrics when overtraining occurs. The x-axis shows the training iteration. Observe how the curves for the train and test samples diverge as the training epochs advance.

Usually, overtraining is a result of too little data or data that is too homogenous. It arises when there are too few degrees of freedom because too many model parameters of an algorithm were adjusted to too few data points. Not all MVA methods are equally sensible to overtraining. While Fisher discriminant [313] hardly suffers

²By generalise is meant that the model recognises only those characteristics of the data that are general enough to also apply to some unseen data.

from it, BDTs usually present partial overtraining, owing to their large number of nodes. Nevertheless, for the BDTs some countermeasures can be applied to preserve the ability to generalise:

- Avoid testing the model on the data used for the training.
- The number of nodes in boosted decision trees can be reduced by removing insignificant ones (*tree pruning*). There are two types, pre-pruning and post-pruning
 - Pre-pruning: Refers to the early stopping of the growth of the decision tree.
 - Post-pruning: Allows the decision tree model to grow to its full depth, then removes the tree branches to prevent the model from overfitting.
- Cross validation is a powerful technique to use all the data for training at the same time that all the data for testing is employed while avoiding overfitting. This method is based in cleverly iterating the test and training split around and it is described on Section B.4.

B.2 Feature-selection optimisation

The optimisation of the list of input variables is divided into two distinct steps: Obtaining the ranking of the most discriminant variables and removing the features that present high correlations.

The initial step is performed by an iterative method based on the feature ranking provided by the XGBoost package. A benefit of the methods that provide a variable ranking is that it allows for interpretability of the results in terms of which physics objects are used. The XGBoost-ranking tool is based on references [314, 315] and it provides a score that indicates how useful each feature is in the construction of the BDTs within the model. It is calculated for a single decision tree by the amount that each attribute split point improves the performance measure, weighted by the number of observations the node is responsible for. After establishing an initial and preliminary set of input variables, each iteration involves the following steps:

1. Train the BDT using the current set of variables.
2. Rank the input variables according using the XGBoost-ranking tool and record various metrics such as the log-loss function and the AUC of the ROC curve.

3. Drop the lowest-ranked feature from the collection of training variables and go to the first step of the loop.

When all variables have been dropped out, the loop is completed. As a result, we get the information of what is the performance of the BDT model depending on which variables have been used. This allows to understand the actual impact of each variable in the BDT performance and the effect of removing it from the training. These steps are schematically illustrated in Figure B.3, where each iteration is depicted.

After doing this, the correlations among variables have to be studied and suppressed. The correlation studies for the variables used in BDTs for region definition are presented in Section B.7.3.

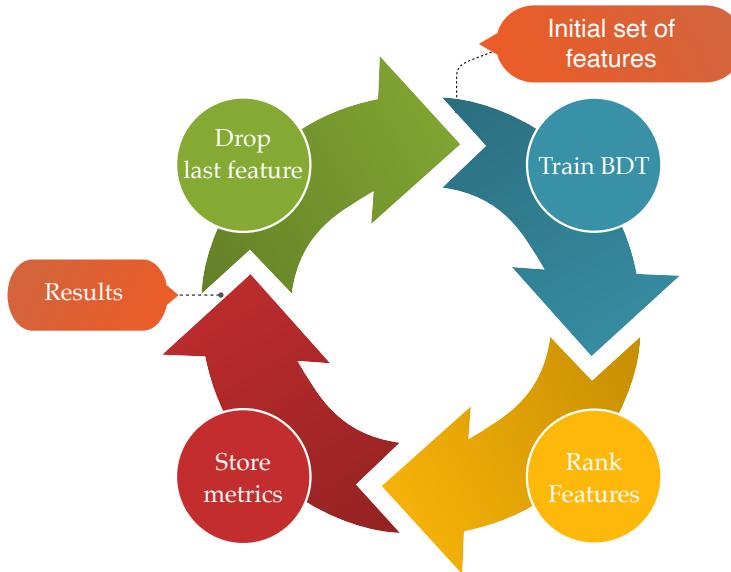


Figure B.3: Schematic view of the iterative algorithm method for feature-selection optimisation. The procedure is reaped until all features are dropped.

B.3 Treatment of negative weights

Negatively weighted events can be problematic when training a BDT for several reasons. Firstly, BDTs, typically operate under the assumption that data represents probabilities or frequency counts. Negative weights violate this assumption, as they do not have a natural probabilistic interpretation. Additionally, negative weights affect the stability and performance of the BDT by slowing the convergence model

or failing to generalise to new data. Negative weights can also distort the process of correcting the inaccuracies of the previous trees in the boosting iterations, leading to skewed decision boundaries.

While for the XGBoost there are no dedicated training options to address this problem, the algorithm can converge using negatively-weighted events. However, in the search presented in this thesis, it is preferred to avoid the risks involving the training with negative events and, hence, either the absolute value of the weight is used or only the events with negative weights are taken into account in the training process. Note that these two options are internal reweightings and, after the training, the event weights are the original ones.

For the TMVA case, there are specific options for dealing with events with negative weights. These are explored below in Section B.3.1.

B.3.1 Negative weights in BDT^{Lepton Assignment}

The TMVA library offers several possibilities to deal with the negatively weighted events. These are:

- InverseBoostNegWeights: It boosts with inverse boostweight. This option is not available for gradient boosting.
- IgnoreNegWeightsInTraining: This option is the one that is being used and as its name suggests, it removes the negatively-weighted events from the training sample.
- Pray: This option allows to use negative weights in the training but might cause problems with small node sizes or with the boosting. It was tested and the model could not achieve stability.
- PairNegWeightsGlobal: This option is still experimental. It takes the negatively weighted events and pairs them with the events with positive weights, annihilating both. When using this option the gradient BDT was not able to converge.

In the BDT^{Lepton Assignment}, the selected treatment ignores the negatively weighted events in the training. When testing the model, these weights are taken into account.

B.4 Cross validation and k -folding

Cross validation is a technique consisting of training several ML models on different subsets of the input data and evaluate these models on the complementary subset of the data. The goal of cross validation is to estimate the performance of a ML model. Cross validation can identify overfitting or recognise the failure of the model to generalise a pattern.

One particular method to cross validate is the k -folding. It is based on splitting the input data into $k \in \mathcal{N}$ equally-sized subsets. Each of these subsets is referred as a fold. With this procedure, the ML model is trained k times as Figure B.4 shows. For each train, $k - 1$ folds are used in the training set and the remaining non-used fold is the subset of data where the evaluation takes place. All folds are used once as test sample and $k - 1$ times in the training sample. There are no formal rules for the choice of k but typical values are 5 or 10 folds. As k gets larger, the difference in size between the training set and the resampling subsets gets smaller. As this difference decreases, the bias of the technique becomes smaller. In the analysis presented in this thesis a $k = 5$ is used for both BDTs.

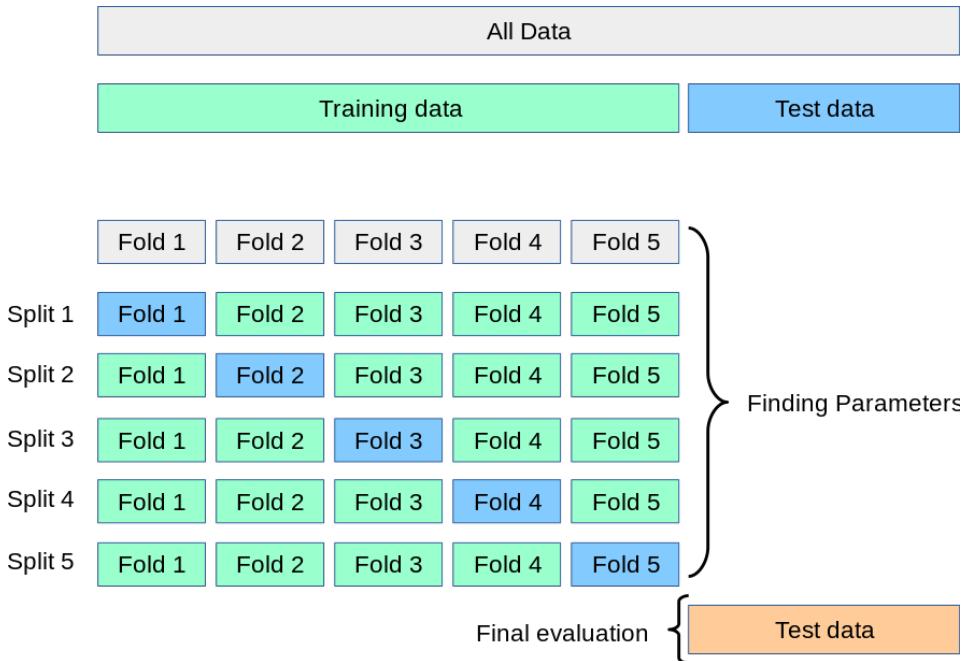


Figure B.4: Illustration of k -folding method for cross validation using 5 folds. For each model, the training folds are presented in green and the test fold in blue. The initial training/test separation is not performed in our case due to the scarcity of data.

The k -folding cross validation resample is of particular interest when the data available is limited because, by applying this method, all events are used in the training phase. It generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split (although train/test split is a particular case of k -folding using $k = 2$).

Note that when the score of the model is applied, each event gets the score that was assigned when it was used as test event. Not doing this would bias the model.

The expected error for a model $f(x|\theta)$ trained using k -folding is:

$$\text{err}(\mathcal{T}) = \frac{1}{k} \sum_k \text{err}(\mathcal{T}_k), \quad (\text{B.4})$$

where $\text{err}(\mathcal{T}_k)$ is the error as described in Equation B.3 for each splits test. As Equation B.4 shows, an increase in the number of folds would imply more models to average over and, hence, implying an improvement in the confidence of how consistent the $f(x|\theta)$ achieves a given level of performance. However, a larger k would also reduce the statistical strength of each fold.

When it comes to assigning an event to a fold, there are no particular rules besides trying to ensure that all the folds represent the general behaviour of your sample and that the size of all folds is the same. The best way to ensure this is by shuffling the dataset randomly. In this analysis, the function to split the data sample among the folds is using the event number. The event number is a categorical variable that assigns a number to each event and it has no physical meaning. Therefore, using this variable to split the data does not create any bias.

B.5 Hyperparameters

Hyperparameters is the term used to refer to the specifications that control³ the learning process of a ML algorithm.

A ML model is defined by its model parameters, which are set by the process of training. To reach some level of intelligence, the process of training a model involves selecting the optimal hyperparameters that the learning algorithm will use to learn the ideal model parameters that accurately map the input variables (\mathbf{x}) to the labels (y). The learning algorithm uses hyperparameters when learning, but these are not included in the resulting model.

³The prefix “hyper” suggest that these parameters are on a higher level that modulates the training process.

XGBoost

In XGBoost, the hyperparameters are classified in three categories: general, booster and task hyperparameters. For the work developed in this thesis, the parameters related to the boosting of the trees are the ones that have been optimised.

- General parameters: Refer to which booster its been used. The used option is based on linear functions. The general parameters also adapt the algorithm to the used device. Since the training of the BDTs takes place in ARTEMISA⁴ facility, it is set GPU.
- Learning task parameters: Decide on the learning scenario. Specify the learning task and the corresponding learning objective.
- Booster parameters: Control the performance of the selected booster. For trees, the most relevant are:
 - **Learning rate:** This tuning parameter determines the step size at each iteration while moving toward a minimum of a loss function. Figure B.5 shows the evolution per epoch for the loss function depending on the learning rate. This hyperparameter is also known as eta or shrinkage, and it ranges from 0 to 1.
 - **Minimum split loss:** Also known as *gamma* or Lagrangian multiplier. A node is split only when the resulting split gives a positive reduction in the loss function and the gamma gives the minimum loss reduction required to make a further partition on a leaf node of the tree. Therefore, the higher the gamma, the more conservative the algorithm will be. The minimum split loss ranges from 0 to ∞ . After testing several values for this hyperparameter, it is not being constrained.
 - **Minimum child weight:** It defines the minimum sum of weights of all observations required in a child. When the tree partition results in a leaf node with the sum of instance weight less than the value of this hyperparameter, the tree stops partitioning. Higher minimum child weight prevents the model from learning too specific relation. So, this is done to prevent overfitting. This tuneable parameter ranges from 0 to ∞ .
 - **Maximum tree depth:** It refers to the number of splits in each tree, which controls the complexity of the boosted ensemble. When `MaxDepth`

⁴ARTEMISA (ARTificial Environment for Machine learning and Innovation in Scientific Advanced computing) is a ML dedicated facility at IFIC. It is composed of several Intel Xeon Platinum CPUs and Tesla Volta GPUs that help to find the optimal configuration for ML algorithms.

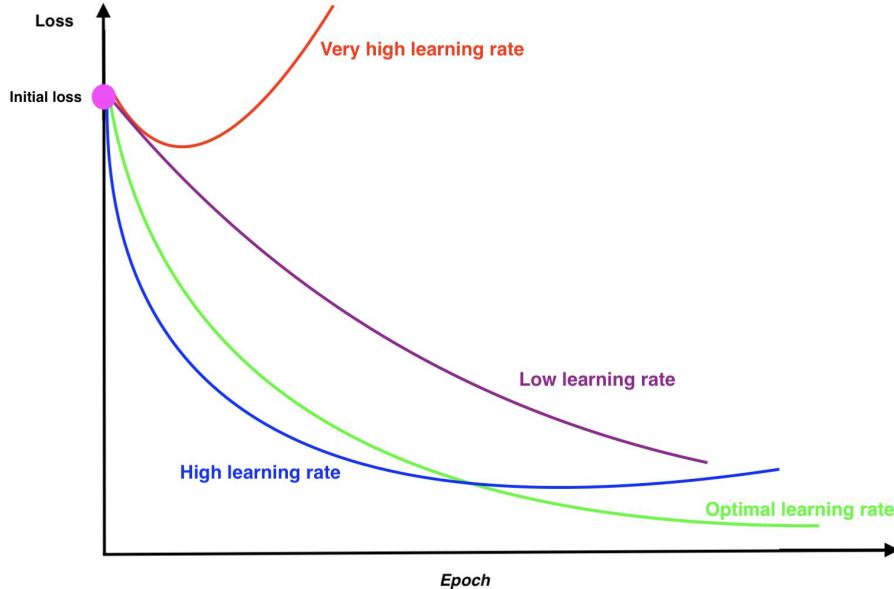


Figure B.5: Different loss-function curves versus iteration. Learning rate is one of the most important hyperparameters to adjust well during the ML model training. If it's high, it can cause the model to diverge. If it's too low it can slow down the training.

is set to an integer value greater than zero, the created cell tree will not be deeper than this integer. This hyperparameters controls de pre-pruning described in Section B.1.1.2. The maximum depth of a tree is an integer ranging from 1 to ∞ but it is rare to have trees with depth higher than 10 since XGBoost aggressively consumes memory when training a deep tree. In the BDTs for region definition, this hyperparameter is set to either 4 or 5.

- **Scale of positive weight:** When the categories are imbalanced, as it is the case for the signal and background in this analysis, the signal sample can be reweighted to have a larger impact. The scale of positive weight is the parameter by which the weights of the target process are multiplied. The typical value to consider is the fraction between positive instances (signal) and negative instances (background). When the BDT is targeting the identification of background processes this hyperparameter can be also used but it takes values closer to the unity.
- **Tree method:** It alludes to the tree-construction algorithm used in XGBoost. The method used here is the GPU implementation of a variant of the LightGBM [293, 316].

TMVA ROOT

This section complements the hyperparameter optimisation for the BDT for the lepton assignment that is described in Section 6.4.2.3. The TMVA-based BDTs allow to configure the several training hyperparameters. From those, the ones that have been explored are:

- **Number of trees:** Number of trees in the forest. The more trees, the more complex the model is and, hence, it can learn more. However, the complexity risk is that the BDT can learn the specifics of the training sample, i.e., overtraining.
- **Minimum size for each node:** Minimum percentage of training events required in a leaf node. The default for classification: 5%.
- **Number of cuts:** Control the number of cuts tested within a variable in order to find the optimal cut value for a node splitting.
- **Negative weight treatment:** Controls the approach for handling events with negative weights during BDT training. The TMVA library has options to include negative weights in the training by pairing them with positively weighted and “annihilating” both events. This strategy has been tested but in the end, removing the negative events provided the best performance.
- **BoostType:** Type of boosting algorithm. The options are
 - **AdaBoost:** This is the most popular type of boosting algorithm and it uses an exponential loss function. Its name comes from “adaptive boosting”. It consists of creating several weak trees, each of them adjusting what the previous one could not. This algorithm lacks robustness in the presence of outliers or mislabelled data points, which can happen in the lepton-origin-assignment scenario.
 - **Gradient boosting:** The TMVA implementation of the gradient boost uses the binomial log-likelihood loss function for classification. This algorithm attempts to overcome the problem presented by AdaBoost regarding the outliers or mislabelled data.
- **Use Bagged Boost:** If used, only a random subsample of the events is used for creating the trees at each iteration. The “bagged sample fraction” is the relative size of the bagged sample to the original size of the data sample.
- **Pruning:** Method used for removing statistically insignificant nodes in the BDT in order to counteract overtraining. For simple decision trees, it is beneficial to first grow the tree to its maximum size and then cut back,

rather than interrupting the node splitting at an earlier stage. However, for BDTs, the best performance is found when small trees (max. depth $\simeq 3$) are used rather than big trees with pruning.

The learning rate, maximum number of trees, and maximum depth that are explained above for the XGboost but these have also been tuned for TMVA.

B.5.1 Hyperparameter optimisation of a BDT with the genetic algorithm

A genetic algorithm (GA) is a search heuristic that mimics the process of natural selection to find solutions to optimisation problems [294]. It works by maintaining a population of potential solutions, and iteratively improving the population over time. Each element of the population is referred to as a chromosome and in our case is a collection of hyperparameters and, hence, a different BDT model once it is trained. Figure B.6 describes the steps of the GA for hyperparameter optimisation. These are:

1. Initialise a population of potential hyperparameter configurations (chromosomes). Each chromosome is a vector of values for the different hyperparameters of the BDT model.
2. Evaluate the fitness of each chromosome. For the BDT, this is done by evaluating the performance of a BDT trained with these hyperparameters. The fitness function used to measure the ability of the model is Zn:

$$Zn = \frac{1}{1 - AUC} \log(\text{LogLoss}) .$$

3. Select the best half of chromosomes as parents for the next generation. This means that half of the chromosomes with higher Zn is kept while the other is dropped.
4. Perform crossover and mutation to renew the population. This means that, from the chromosomes that are kept, new ones are created via two mechanisms:
 - Crossover: Combine the chromosomes of two parents to create new offspring. There is the possibility that a specific value of a hyperparameter is exchanged between two chromosomes. To favour diversity in the new chromosomes, the crossover rate has been set to 80%.

- Mutation: Randomly change one or more genes in a chromosome. For each hyperparameter in each new chromosome, the mutation probability is set to 8%. If a hyperparameter mutates, it is multiplied by a random number from a gaussian distribution with mean 1 and a standard deviation 0.05.

The crossover and mutation rates, as well as the mutation factor can be modified for each optimisation.

- Repeat steps 2-4 until a satisfactory solution is found.

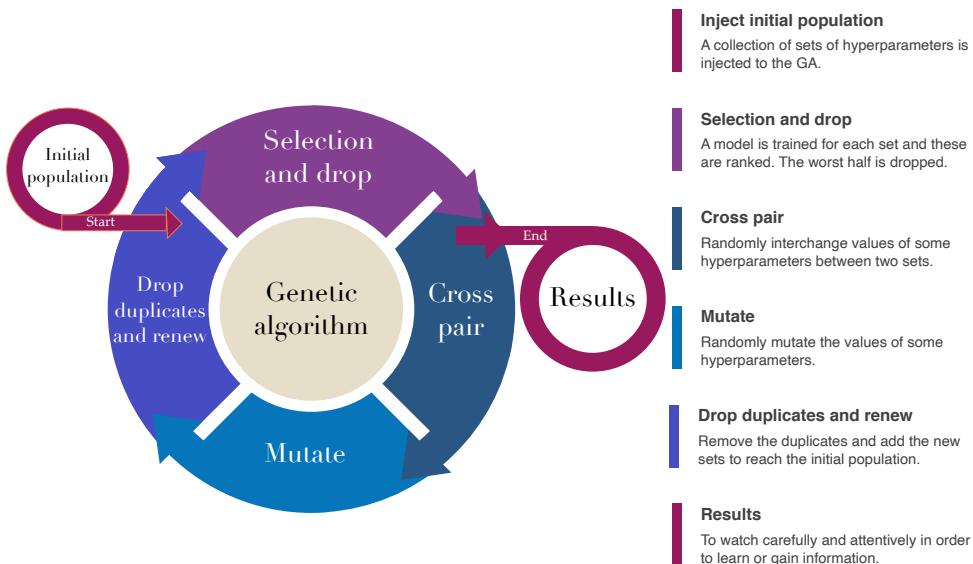


Figure B.6: Schematic view of the GA method for hyperparameter optimisation.

B.6 Other considerations about BDTs

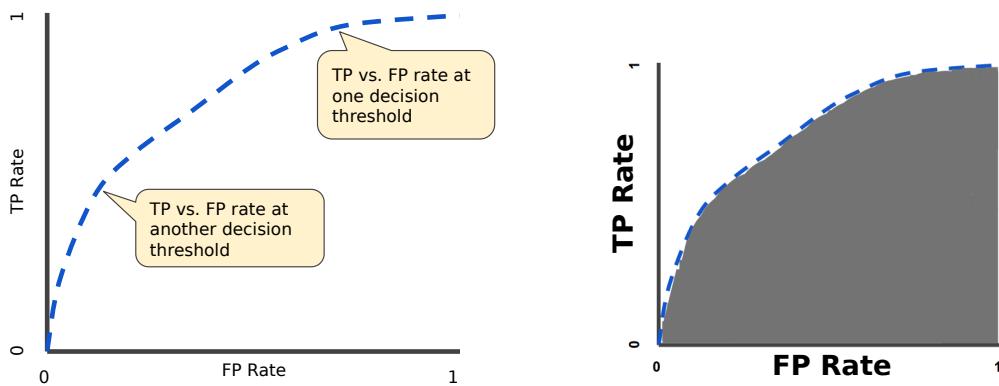
Receiver operating characteristic curve

The receiver operating characteristic curve (ROC) is a graphical plot used that is used to illustrate the ability of a binary classifier. It assesses the tradeoff between true positive (TP) and false negative (FP) rates as the parameters of the classification vary. This is depicted in Figure B.7a.

- True positive rate:** Also known as sensitivity. It is the possibility of a positive test conditioned on truly being positive. For instance, it is the probability for the BDT in Section 6.6.2 to identify a tHq event as such.

- **False positivity rate:** It can be calculated as $1 - \text{sensitivity}$. It refers to the possibility of a negative test given that it is truly positive. In the Section 6.6.2 BDT scenario, it would be the ability to classify a background event as if it was a tHq signal event.

The area under the curve (AUC) is a commonly used quantitative summary, it measures the bidimensional area under the ROC from $(0,0)$ to $(1,1)$ as Figure B.7b shows. The use of the AUC is convenient for several two reasons. Firstly, it is invariant with the scale because it does not measure absolute values but rates. Secondly, it is invariant with respect to the classification threshold and, hence, it evaluates quality of the classification model.



(a) Typical ROC curve. It shows that as the classification threshold decreases, more events are classified as positive, causing both the FP and FN rates to increase.

(b) The AUC varies from 0 to 1. While 0.0 corresponds to a model that always fails, a 1.0 means that the model is right a 100% of times.

Figure B.7: The ROC presents the TP vs the FN rate. The ROC analysis is related to cost/benefit interpretation of decision making.

Precision-Recall curves

While the ROC summarises the trade-off between the true positive rate and false positive rate for a predictive for different probability thresholds, there is another plot that helps with the diagnosis of the binary classification models; the Precision-Recall curves. These summarise the equilibrium between the true positive rate and the positive predictive value for a predictive model using different probability thresholds.

Typically, the use of ROC and precession-recall curves is such that the first type is used when there are roughly equal numbers of observations for each class and the second should be used when there is a moderate to large class imbalance.

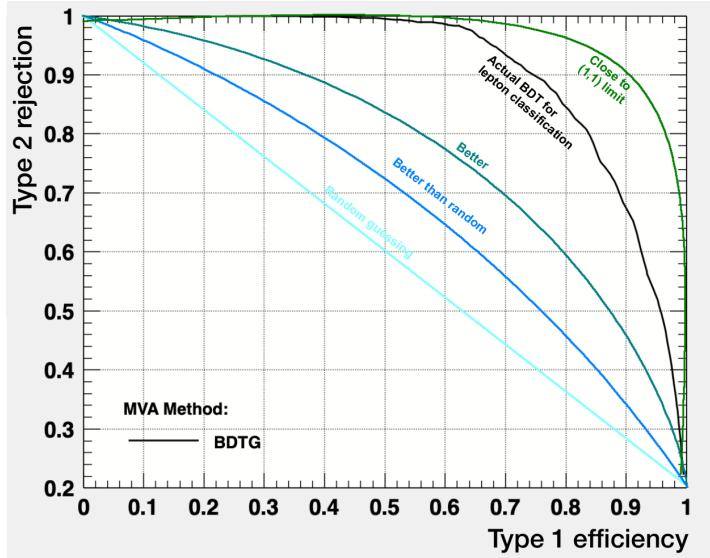


Figure B.8: Precision-recall curves for different models. The one in black corresponds to the $\text{BDT}^{\text{Lepton Assignment}}$ model. The larger is the area under the curve, the better is the model.

For both the ROC and the precision-recall curves, the larger the area under the curve, the better. Figure B.8 shows that the optimal classifier is the one in which the curve in the precision-recall plot is close to the (1,1) point.

Separation power

The separation power is a valuable metric for evaluating the performance of a variable or classifier in terms of its ability to distinguish the target process from other processes. The separation power $\langle S^2 \rangle$ of a classifier y can be quantified using the integral:

$$\langle S^2 \rangle = \frac{1}{2} \int \frac{(\hat{y}_S - \hat{y}_B)^2}{\hat{y}_S + \hat{y}_B} dy, \quad (\text{B.5})$$

where \hat{y}_S and \hat{y}_B are, respectively, the signal and background probability density functions of y . The 1/2 factor is used to keep $\langle S^2 \rangle$ within the [0, 1] interval. The separation is zero for identical signal and background shapes, and it is one for shapes with no overlap.

B.7 Additional plots and tables

This section expands the information provided in the main body of the thesis about the training of the different BDT models. The models and ROC curves

of the $\text{BDT}^{\text{Lepton Assignment}}$ are presented in the first section. The second section shows, for the XGBoost-based BDT, the correlation between the variables and the evolution of the training.

B.7.1 $\text{BDT}^{\text{Lepton Assignment}}$

Figure B.9 compare the train and test distributions for the five different models composing the final $\text{BDT}^{\text{Lepton Assignment}}$. A good agreement between the distributions of the train and test subsets is observed through all the folds. This means that no overtraining in our data. See Section 6.4.2.5 for the description of the training of these models.

In Figure B.10 are presented the ROC curves for the five models trained in Section 6.4.2.5

B.7.2 BDTs for region definition

The three BDTs used to define the analysis regions are presented in Section 6.6.2. Two of these are trained on $2\ell \text{OS} + 1\tau_{\text{had}}$ events to target the signal and the $t\bar{t}$ process. These are the $\text{BDT}(tHq|_{\text{OS}})$ and $\text{BDT}(t\bar{t}|_{\text{OS}})$, respectively. For the $2\ell \text{SS} + 1\tau_{\text{had}}$ channel, only one BDT is used for the analysis, the $\text{BDT}(tHq|_{\text{SS}})$, which targets the signal. In this section, the correlations between the variables of each model are presented first. Then, the training evolution is studied with the AUC and LogLoss progression over the training epochs.

B.7.3 Correlation between input features

When choosing the input variables of an ML model, it is important to check for correlations between these variables because highly correlated variables can lead to redundancy, reducing the efficiency and interpretability of the model. This helps to avoid overfitting, ensuring the model generalises well to new, unseen data. For the feature selection described in Section 6.6.2.2, the linear correlations among all variables within a model have been explored and the results are presented in Figures B.11, B.12 and B.13. In these figures, the more intense the colour, the higher the correlation. For the $\text{BDT}(tHq|_{\text{OS}})$ is of particular relevance the correlation between the $\Delta\eta$ (forward non- b -tagged jet, subleading b -tagged jet) and the multiplicity of forward jets. These two variables have a correlation of 87% and 85% for the signal and the background, respectively, so it could be beneficial to remove one of these two. For the $\text{BDT}(t\bar{t}|_{\text{OS}})$ model there is not any highly correlated pair

of variables, being the highest correlation a 52%. Similarly, for the $\text{BDT}(tHq|\text{ss})$, the only correlation appears for the p_T and the energy of the top quark. Since the energy depends on the momentum, it is not surprising that these two present a 66% ($t\bar{t}$) and 72% (other processes) correlation.

B.7.4 Evolution of training

The loss values of the training and test offer a deeper understanding of how learning performance evolves across different epochs. These metrics provide useful insight into how the learning performance changes over the number of epochs. This can help diagnose potential learning issues that might result in an underfitted or overfitted model.

Figures B.14, B.15 and B.16 present the evolution of the AUC and LogLoss functions at each iteration of the training procedure. These are plotted, separately for train and test, so that overtraining can be detected. The values for the AUC and LogLoss of the last epoch correspond to those presented in Figure 6.23 (although small discrepancies can appear). The train and test curves in these plots are not diverging (as in the overtraining example in Figure B.2), indicating that no overfitting is present.

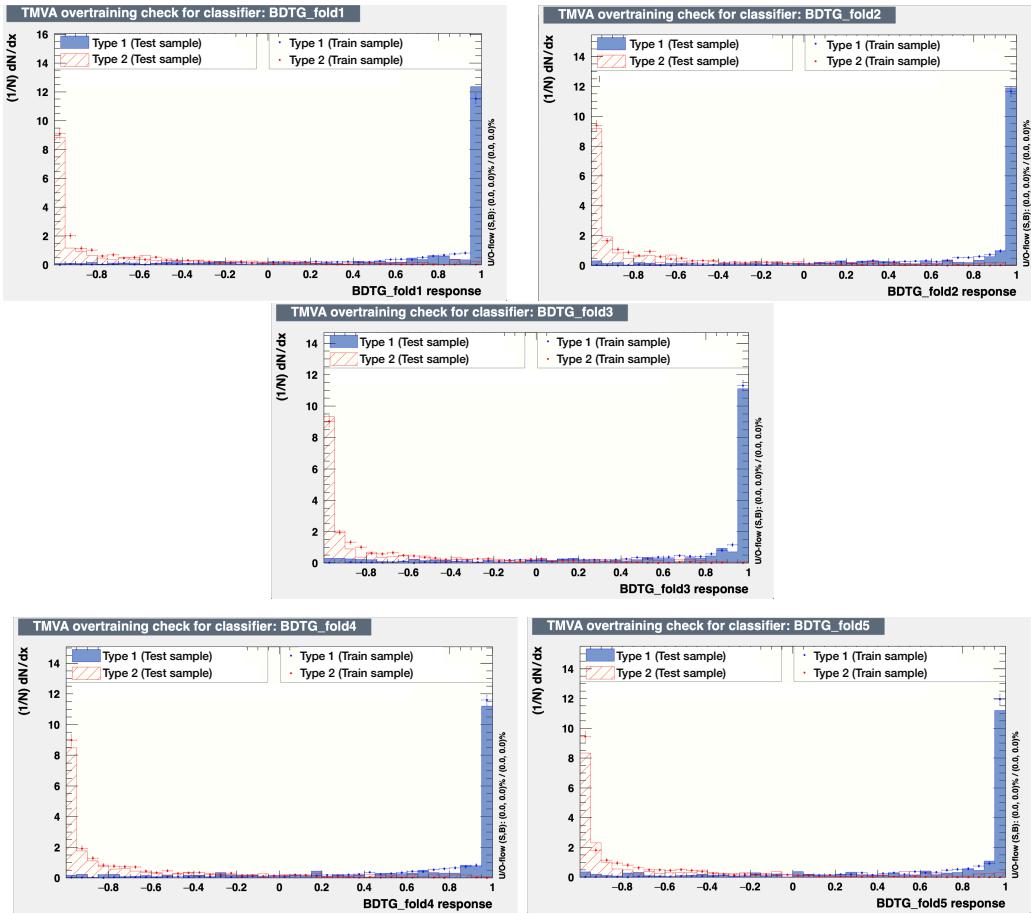


Figure B.9: BDT^{Lepton Assignment} distributions for the different folds. Type 1 and Type 2 categories are presented in blue and red, respectively. The doted distribution corresponds to the scores when the model is evaluated on the training sample. The areas on the histogram refer to the score of the same model when applied to the test sample. On one side, these plots allow to compare the model of each fold. As can be seen, both categories have the same behaviour through the five models. This indicates that the subset of the MC sample used in each split is not affecting the model. On the other side, it can also be checked that the distribution for each fold is the same for the train and the test subsamples.

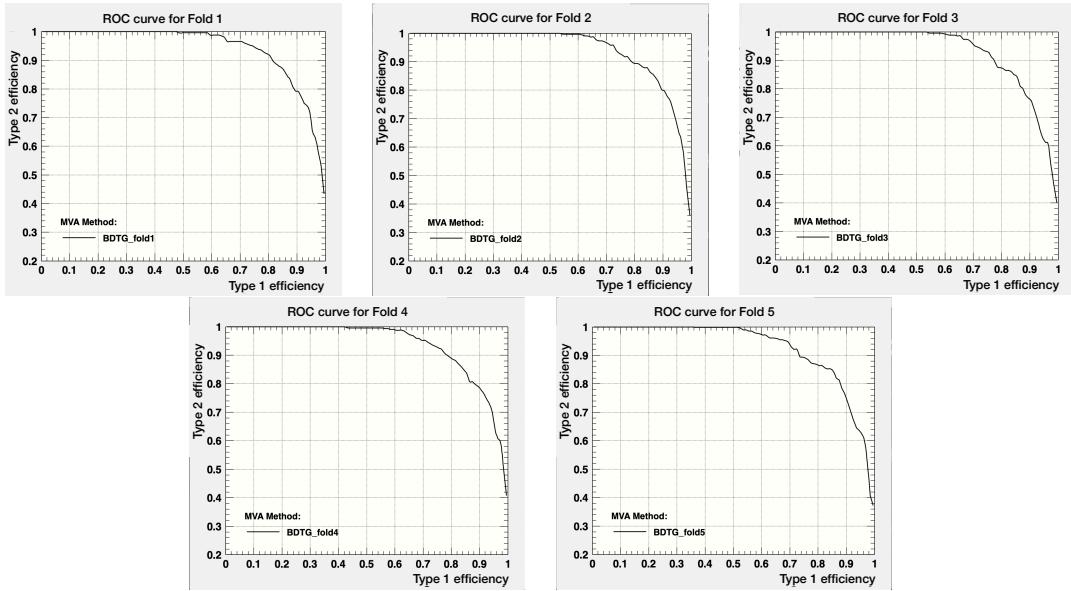


Figure B.10: ROC for the five different BDTs trained with k-folding for the BDT^{Lepton Assignment} in Section 6.4.2. All the curves show a similar behaviour.

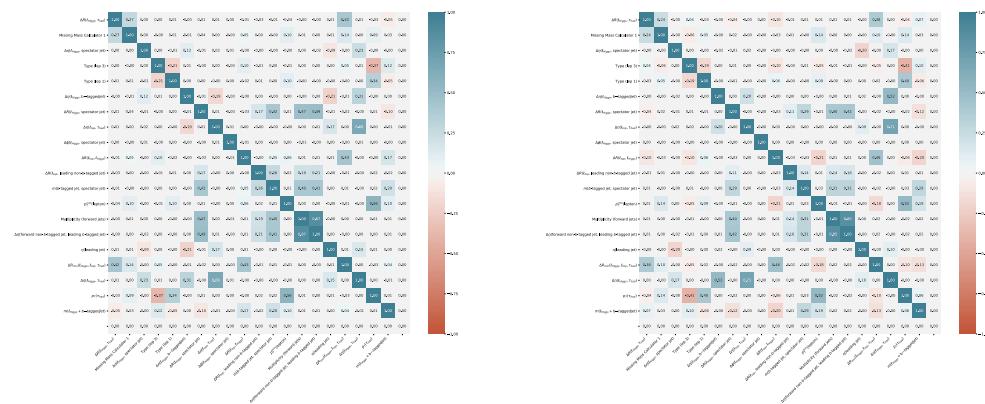


Figure B.11: Correlation matrix of the variables used as input in the BDT($tHq|_{\text{OS}}$). For the left matrix, the correlations among variables are explored for the tHq process only while, for the right plot, these are studied for all the backgrounds. Correlations are presented in blue while anticorrelations are in orange.

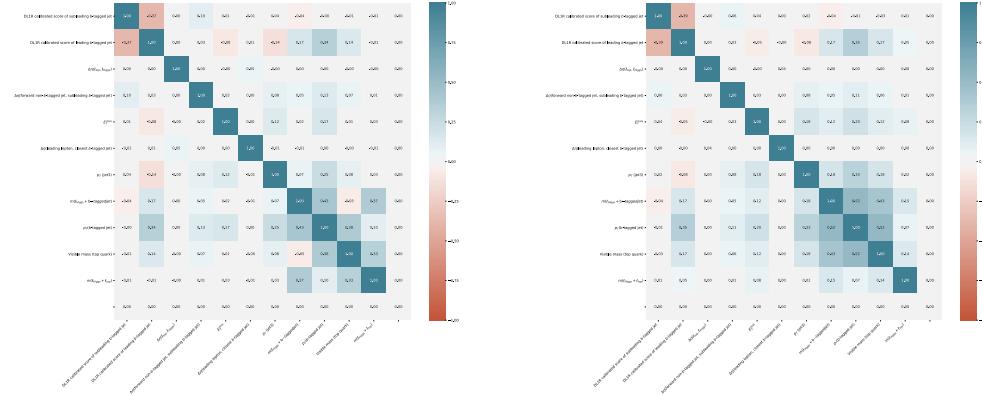
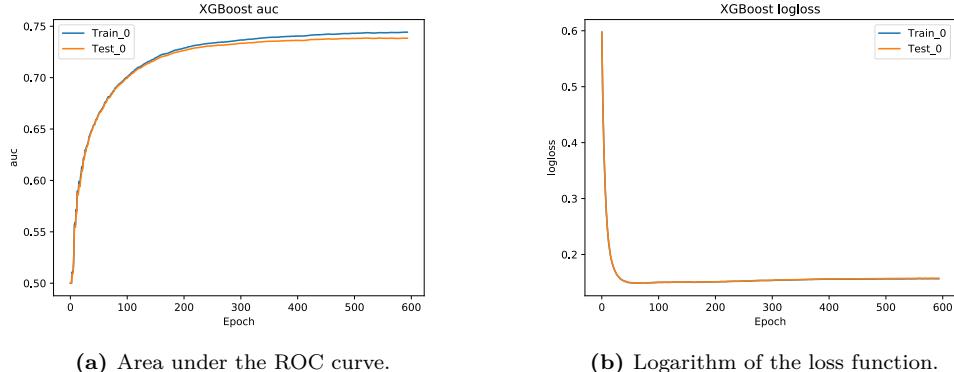


Figure B.12: Correlation matrix of the variables used as input in the BDT($t\bar{t}|_{\text{OS}}$). For the left matrix, the correlations among variables are explored for the $t\bar{t}$ process only while, for the right plot, the correlations are calculated for the other processes. Only linear correlations are shown and higher-order relationships may not be reflected in this matrix.



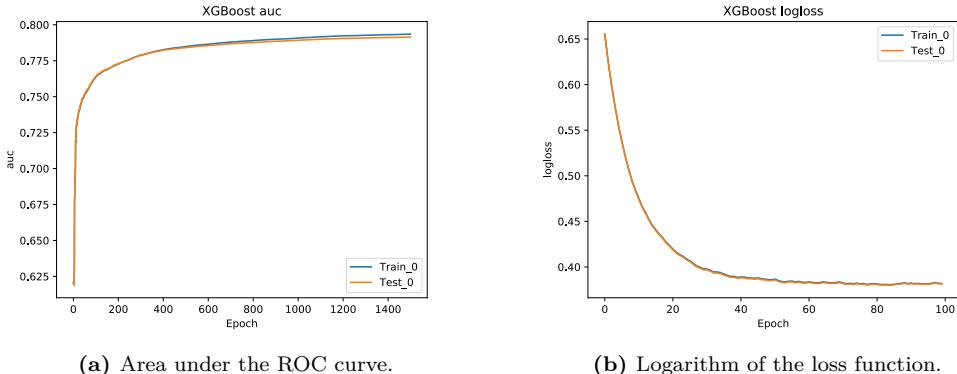
Figure B.13: Correlation matrix of the variables used as input in the BDT($tHq|_{SS}$). For the left matrix, the correlations among variables are explored for the tHq process only while, for the right plot, these are studied for the background processes.



(a) Area under the ROC curve.

(b) Logarithm of the loss function.

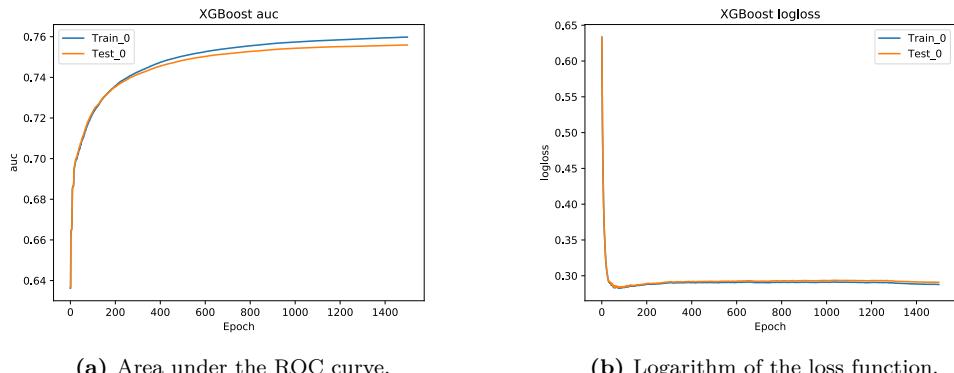
Figure B.14: Evolution of the AUC and LogLoss functions for the training of the BDT($tHq|_{\text{os}}$). The x-axis shows the training iteration.



(a) Area under the ROC curve.

(b) Logarithm of the loss function.

Figure B.15: Evolution of the AUC and LogLoss functions for the training of the BDT($t\bar{t}|_{\text{os}}$). The x-axis shows the training iteration.



(a) Area under the ROC curve.

(b) Logarithm of the loss function.

Figure B.16: Evolution of the AUC and LogLoss functions for the training of the BDT($tHq|_{\text{ss}}$). The x-axis shows the training iteration.

Appendix C

Distribution of kinematic variables

In this appendix are presented all the distributions of the variables used to train the BDT model, with the exception of those already presented in Chapter 6.

C.1 BDT^{Lepton Assignment}

Normalised distributions of the BDT^{Lepton Assignment} input variables for the $2\ell \text{ SS} + 1\tau_{\text{had}}$ samples. The distributions using all events (black) and using only the positively-weighted ones (green) are superimposed. The error bands correspond to the statistical error. Note that the “positive vs negative” distributions have the same profile because here we are not separating according to the Type 1 and Type 2 categories.

In the left subfigure, the distribution using all events (black) is superimposed over the one using only the positively-weighted ones (green) for comparison. In the right plot, the events in which the ℓ_1 comes from the top quark are shown in blue while those for which ℓ_1 is produced from the Higgs boson are plotted in red. Note that only events that contain the truth-reco matching described in Section 6.4.2.1 are used to produce these plots.

Then, uncertainty bands correspond exclusively to the statistical error. In the case of the distributions containing negatively-weighted events, the calculation of the statistical uncertainty is described in Appendix D.2

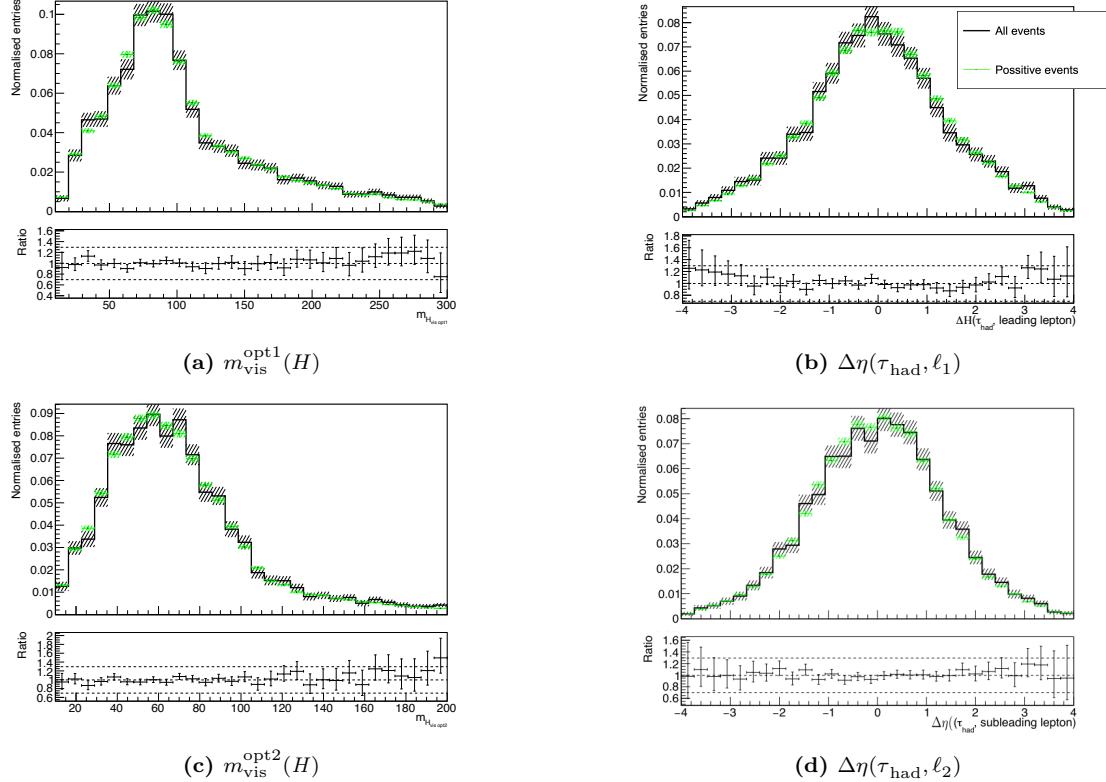


Figure C.1: Normalised distributions of the highest ranked variables for the light-lepton assignment. The black distribution is obtained using all events and the green one corresponds to the scenario in which only the positively weighted events are used. The comparison between the Type 1 and Type 2 categories for these variables is presented on Figure 6.5.

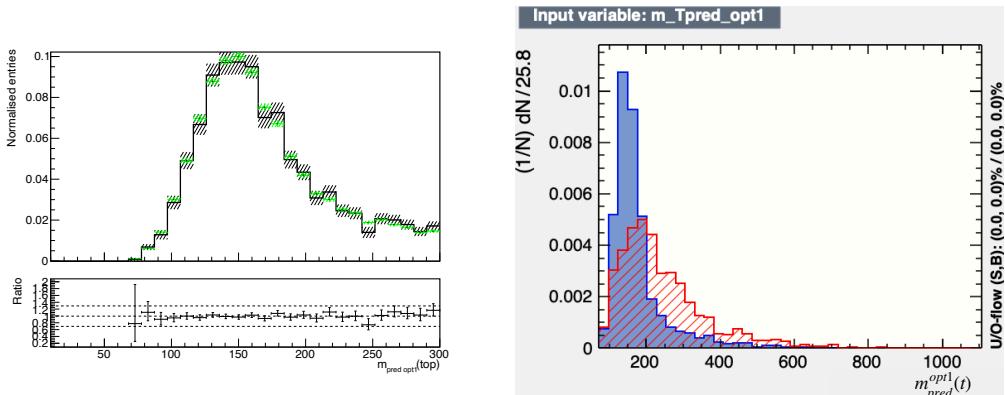


Figure C.2: Normalised distributions for $m_{\text{pred}}^{\text{opt1}}(t)$.

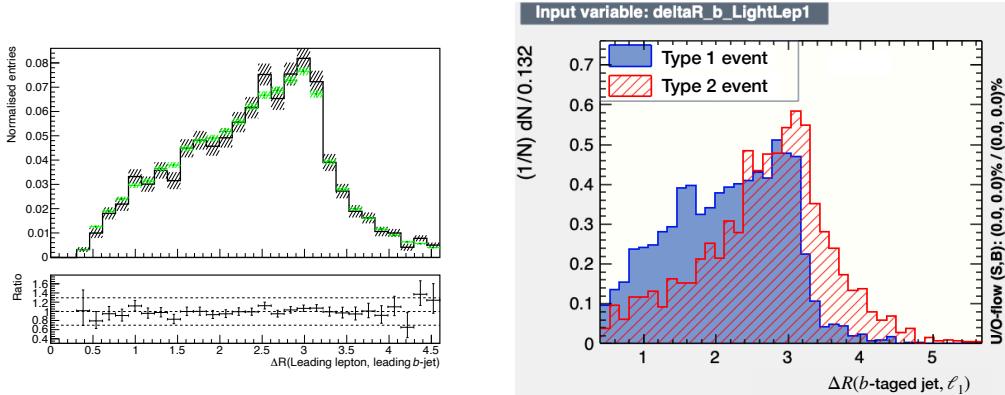


Figure C.3: Normalised distributions for the geometric proximity, ΔR , between the b -tagged jet and the leading-light lepton.

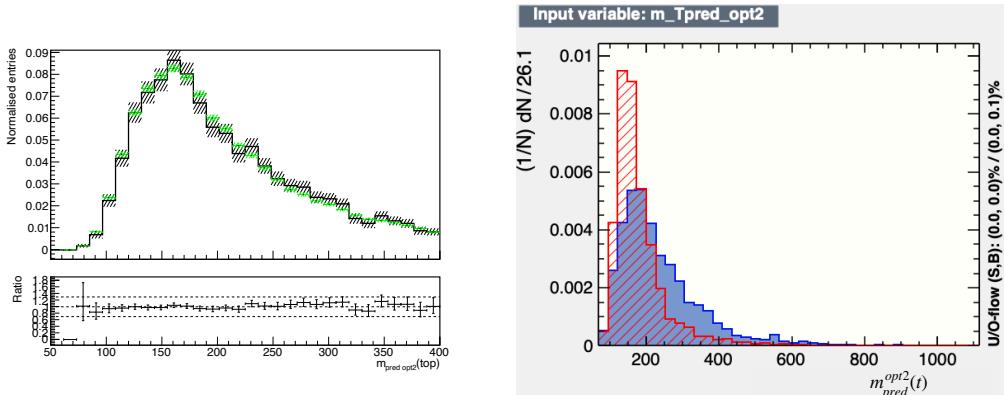


Figure C.4: Normalised distributions for $m_{\text{pred}}^{\text{opt2}}(t)$.

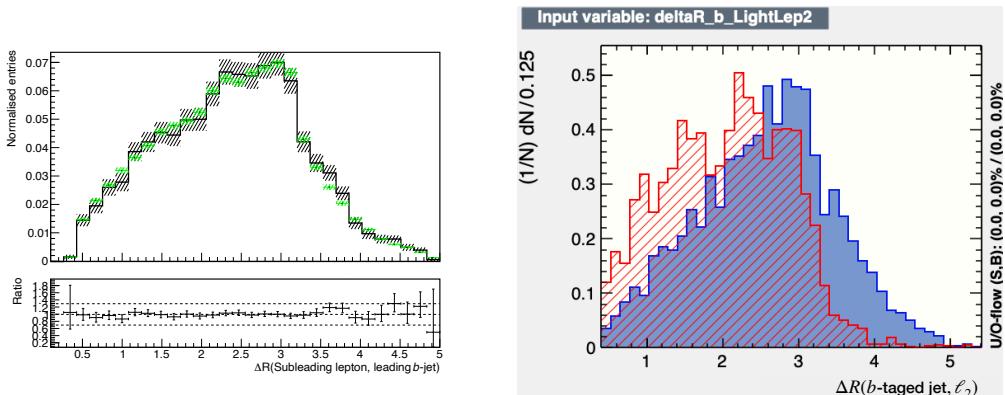


Figure C.5: Normalised distributions for $\Delta R(b\text{-tagged jet}, \ell_2)$.

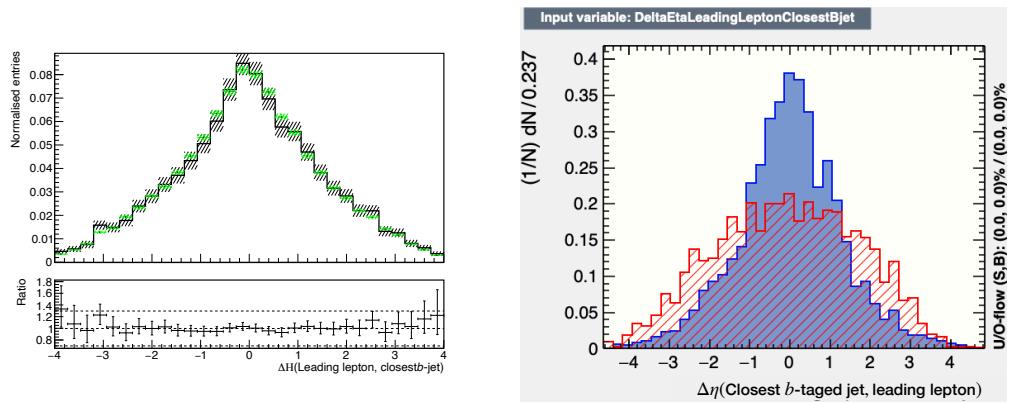


Figure C.6: Normalised distributions for $\Delta\eta(\text{closest } b\text{-tagged jet, leading lepton})$.

C.2 BDT($tHq|_{\text{OS}}$)

For all distributions presented here, the uncertainty bands include the statistical and systematic uncertainties and the lower panel presents the ratio between the collected data and the MC simulation. Additionally, the χ^2 measures the agreement between real and simulated data events.

When training an ML model, it is important to use variables that provide good modelling, i.e., good data/MC agreement. For most of the distributions presented in the rest of the appendix, while the data/MC agreement is far from being exact, it is compatible with the uncertainty. The variables are presented in their order of importance within the BDT (see Figure 6.17a)

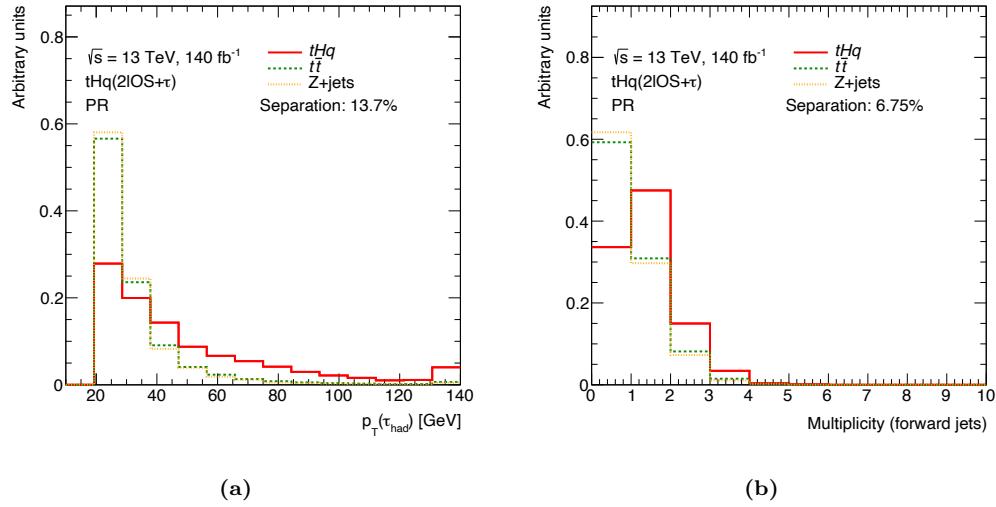


Figure C.7: Separation plot for the two most discriminant variables of the BDT($tHq|_{\text{OS}}$). The original distributions for these features are presented in Figure 6.18. In (a), the transverse momentum of the hadronically-decaying τ -lepton is showed. The reconstructed τ_{had} is more boosted when it is produced from the Higgs-boson decay than in the main backgrounds. In (b), the number of forward jets present in the final state is shown. A forward jet is defined as a jet reconstructed within $2.5 < |\eta| < 4.5$. The jets more forward than $|\eta| < 4.5$ are not detected. The forward jets are selected with the *forward Jet Vertex Tagger* [304]. The spectator-quark-initiated jet in the tHq production is quite forward, so the multiplicity of forward jets peaks at one. In contrast, $t\bar{t}$ and $Z + \text{jets}$ do not have forward jets in hard-scattering process.

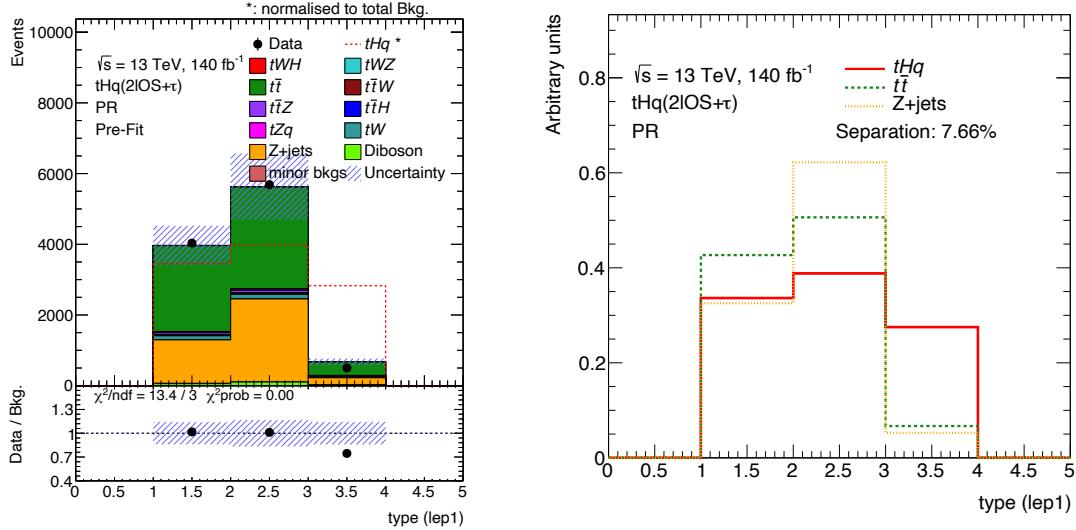


Figure C.8: Distribution and separation plot for the flavour of the leading lepton. The types are e^-/e^+ (1), μ^-/μ^+ (2), and τ^-/τ^+ (3). Note that, for the backgrounds, the reconstructed τ_{had} rarely is the leading lepton. In contrast, for the tHq production, the flavour of the leading lepton is more distributed. There is a disagreement between data and MC prediction for the events in which the τ_{had} is the leading lepton (third bin), this can be linked with the phenomena observed in Figure 6.18a, in which the modelling of the p_T (τ_{had}) misbehave for large p_T .

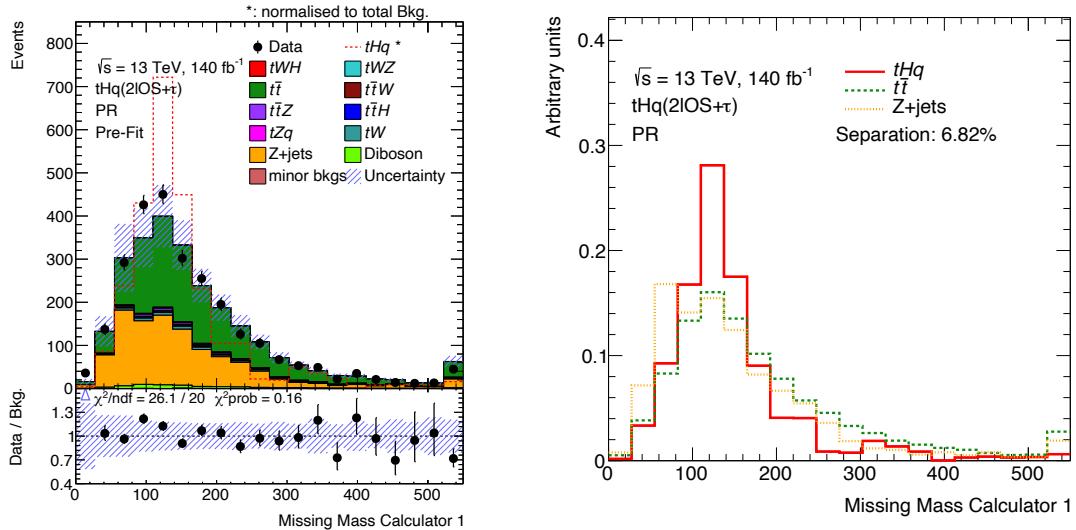


Figure C.9: Distribution and separation plot for the reconstructed mass of the Higgs boson using the MMC method [290]. It peaks around 125.25 ± 0.17 GeV, which corresponds to m_H [10]. This variable allows to separate events with a real Higgs boson (e.g. tWH , tHq , $t\bar{t}H$) from the rest.

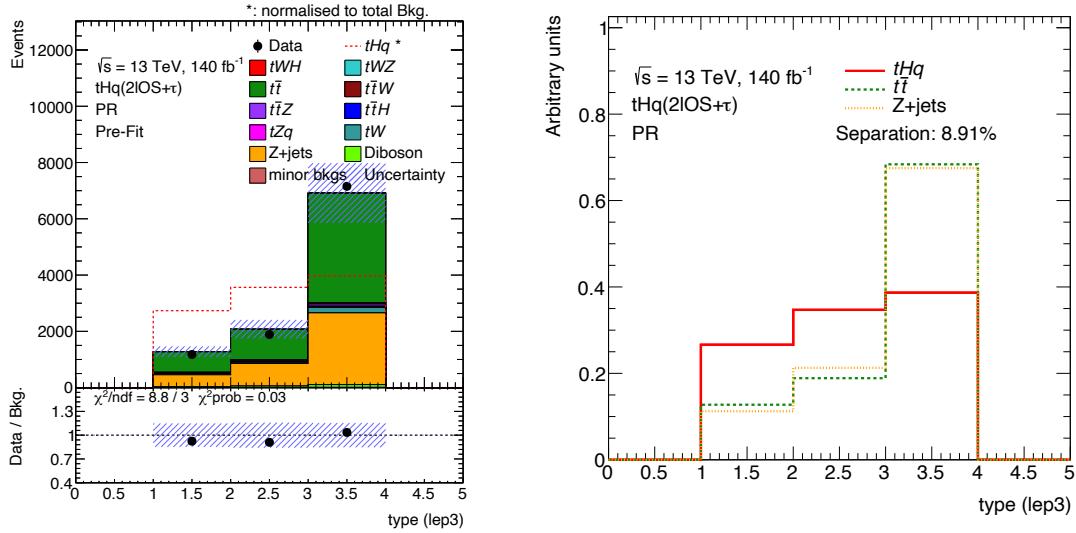


Figure C.10: Distribution and separation plot for the flavour type of the sub-sub-leading lepton. In Figure C.8 is observed that, for the backgrounds, the τ_{had} does not tend to be the leading lepton. In agreement, this Figure shows that the τ_{had} typically is the lepton with lower p_T for the backgrounds. However, this is not the case for the tHq production. Observe that Type(ℓ_1) and Type(ℓ_2) are the only features that use the p_T -based definition to refer to the light leptons.

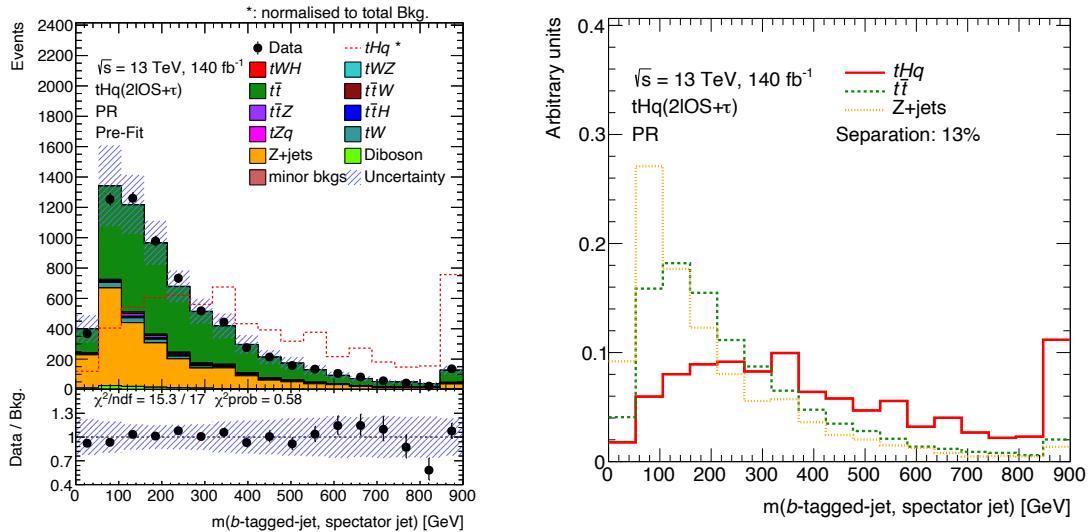


Figure C.11: Distribution and separation plot for the invariant mass of the leading b -tagged jet and the spectator jet. The spectator jet is the one produced by the quark that is scattered in the Feynman diagrams of Figures 2.12a and 2.12b.

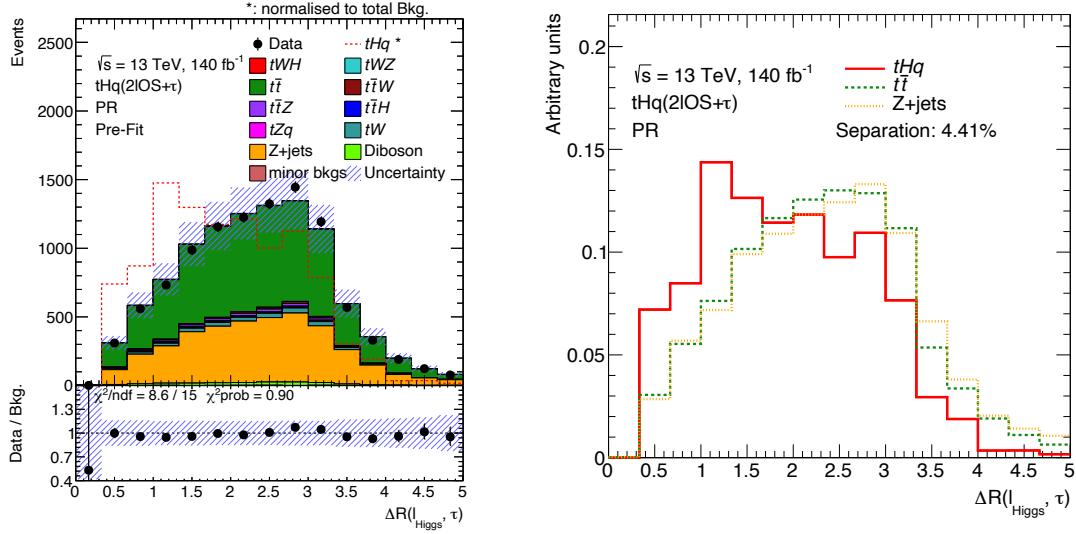


Figure C.12: Distribution and separation plot for the ΔR distance between the ℓ_{Higgs} and the τ_{had} . Since in the tHq process the ℓ_{Higgs} and the τ_{had} are direct-decay products of the Higgs boson (83.71% of times as calculated in Table 6.9) they tend to be geometrically close.

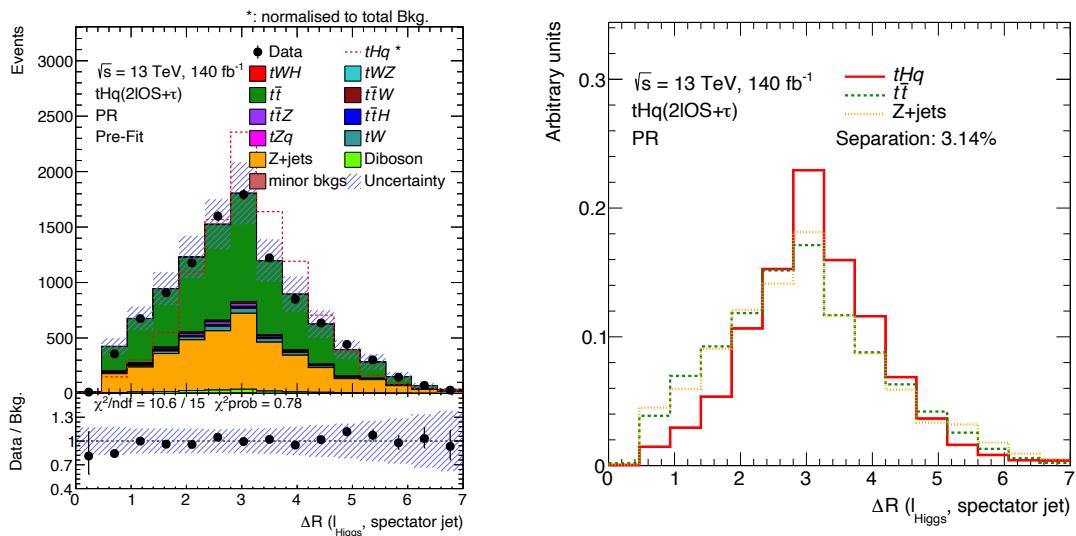


Figure C.13: Distribution and separation plot for the ΔR distance between the ℓ_{Higgs} and spectator jet.

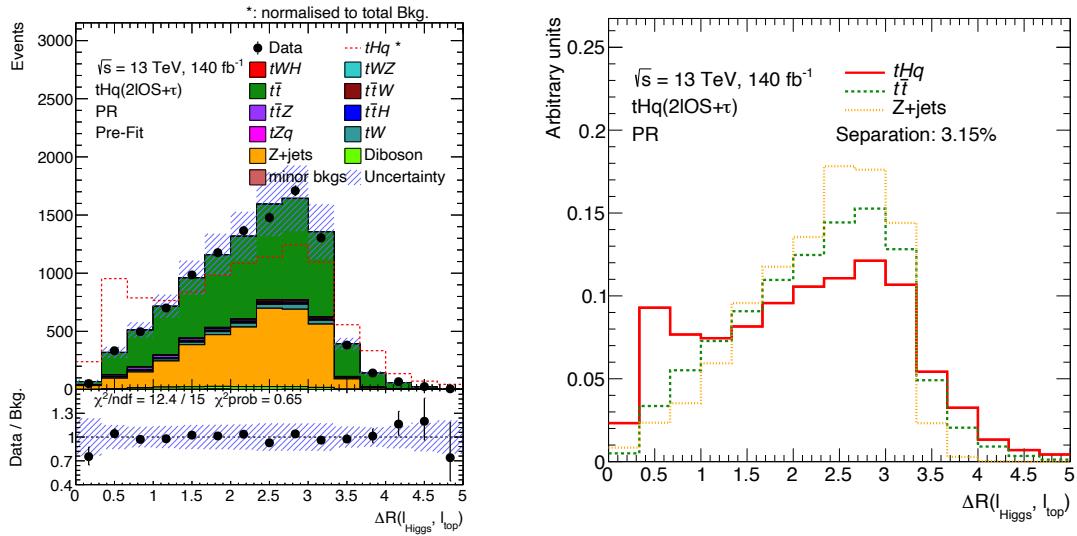


Figure C.14: Distribution and separation plot for the ΔR distance between the two charged light-flavoured leptons.

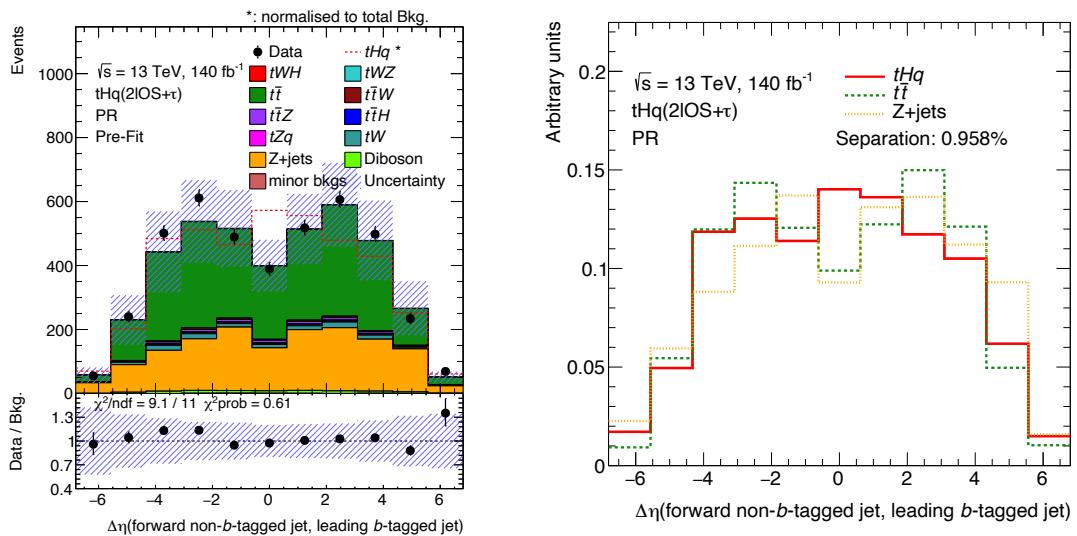
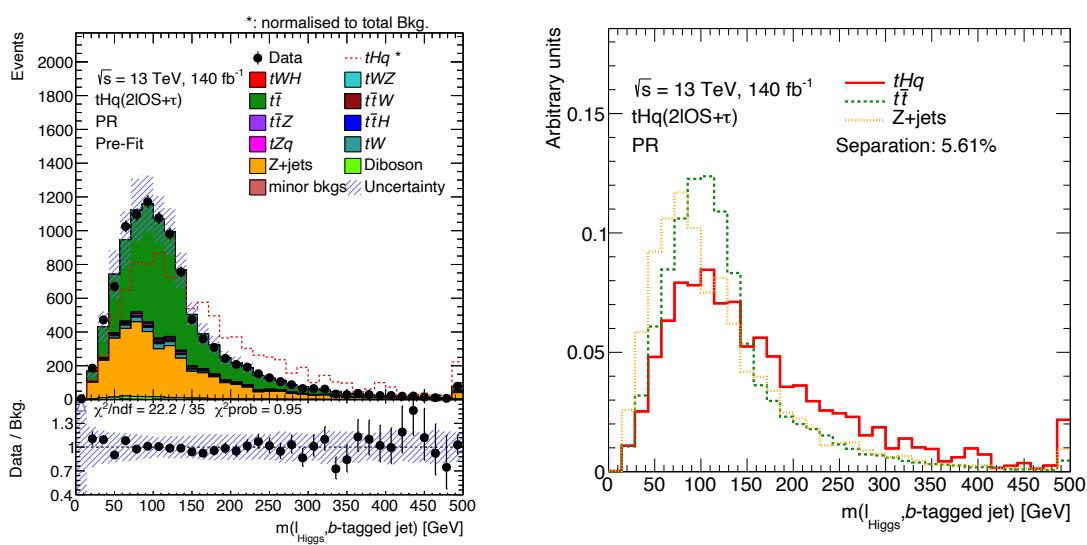
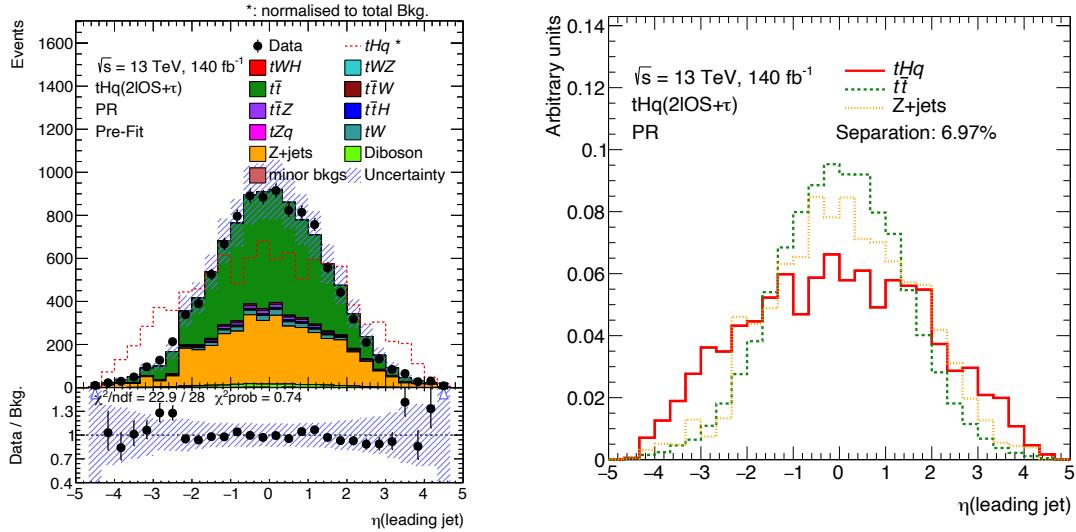


Figure C.15: Distribution and separation plot for the $\Delta\eta$ distance between the forward non- b -tagged jet and the leading b -tagged jet. The non- b -tagged jets are also referred as light jets.



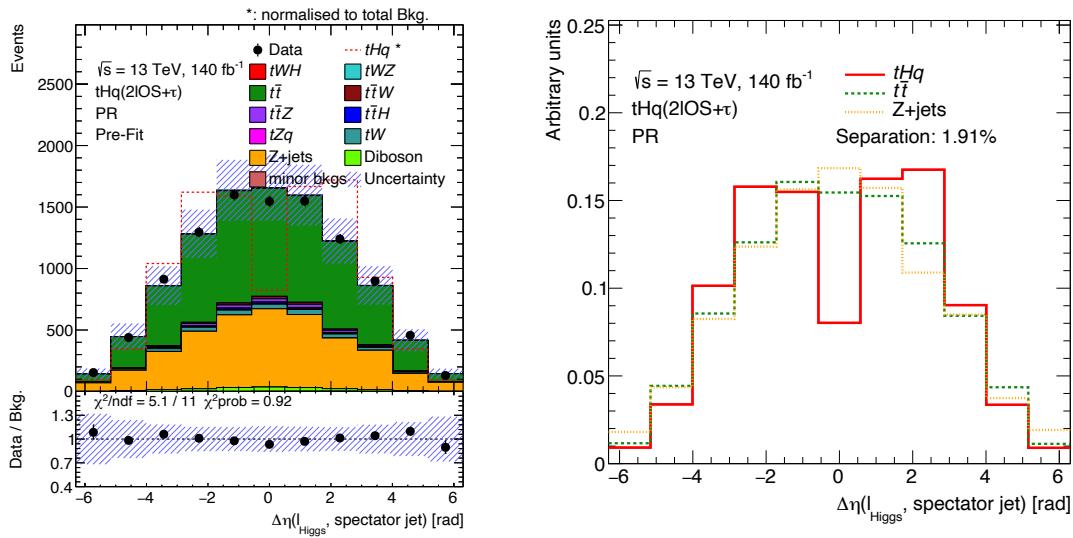


Figure C.18: Distribution and separation plot for the $\Delta\eta$ distance between the ℓ_{Higgs} and the jet initiated by the spectator quark.

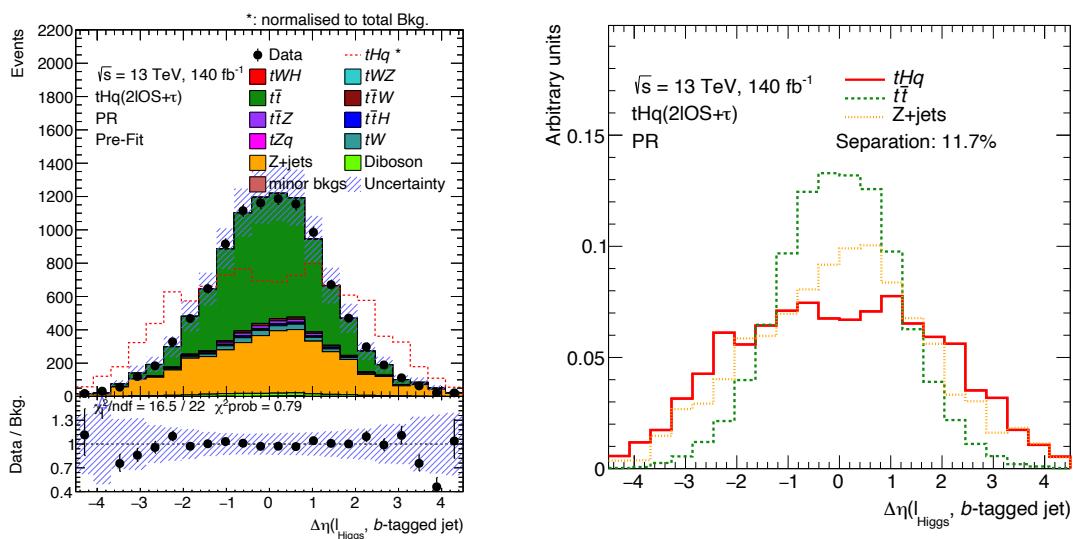


Figure C.19: Distribution and separation plot for the $\Delta\eta$ distance between the ℓ_{Higgs} and the leading b -tagged jet.

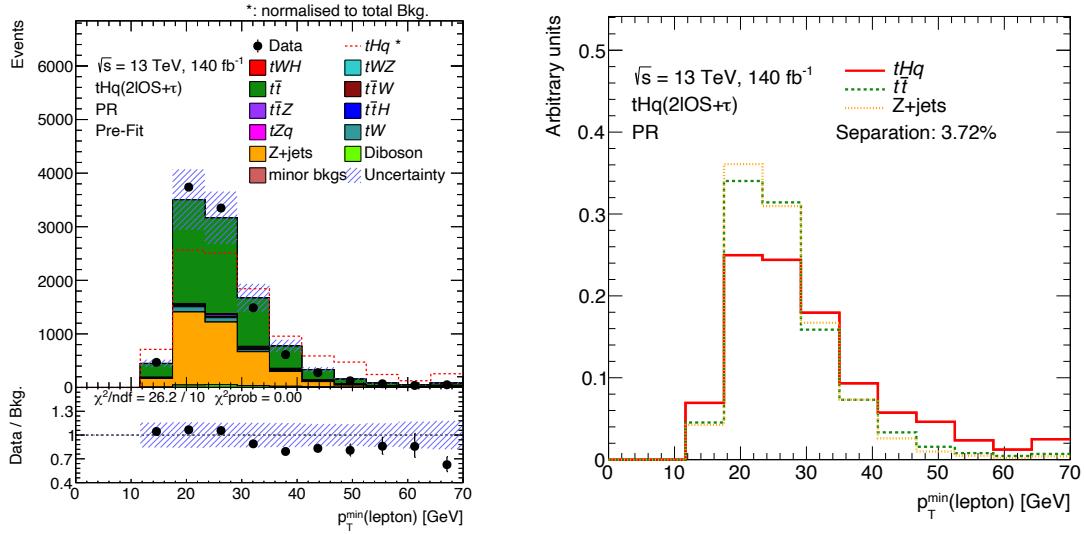


Figure C.20: Distribution and separation plot of the p_T of the softest lepton (including τ -flavoured leptons).

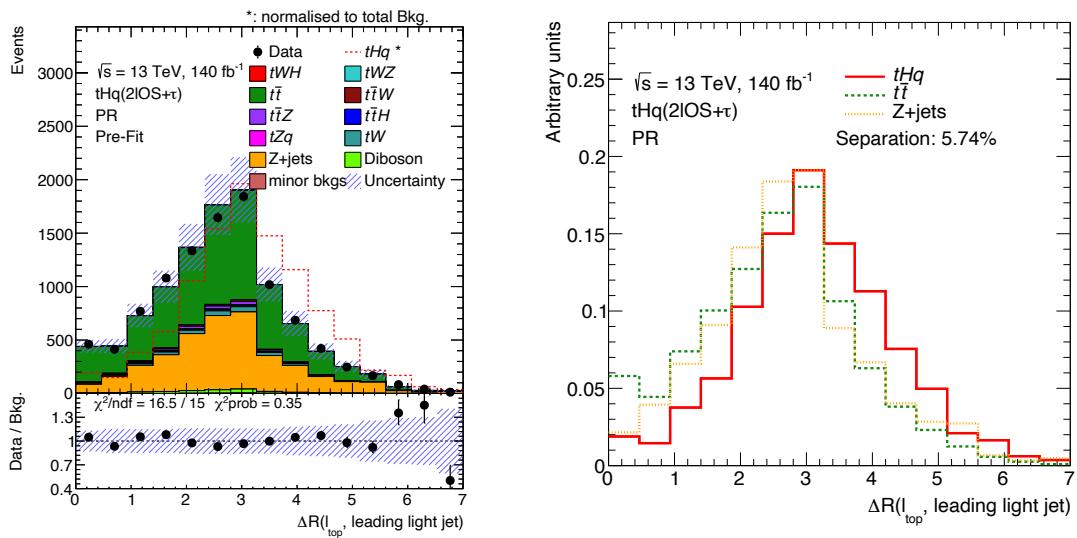


Figure C.21: Distribution and separation plot for the ΔR distance between the ℓ_{top} and the leading non- b -tagged jet.

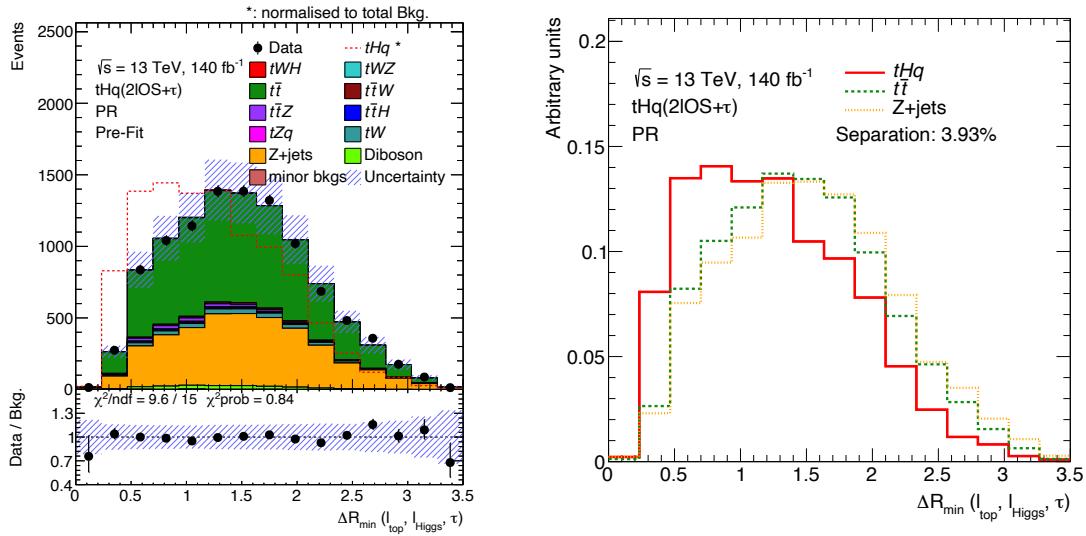


Figure C.22: Distribution and separation plot for the minimum ΔR distance between any two charged leptons form the the three that are present in the final state.

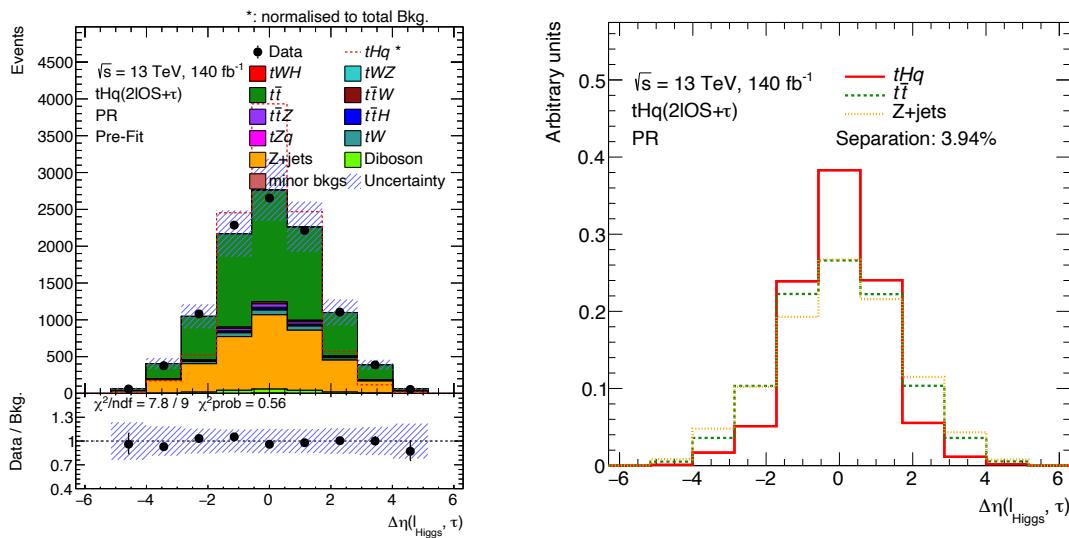


Figure C.23: Distribution and separation plot for the $\Delta\eta$ distance between the ℓ_{Higgs} and the τ_{had} .

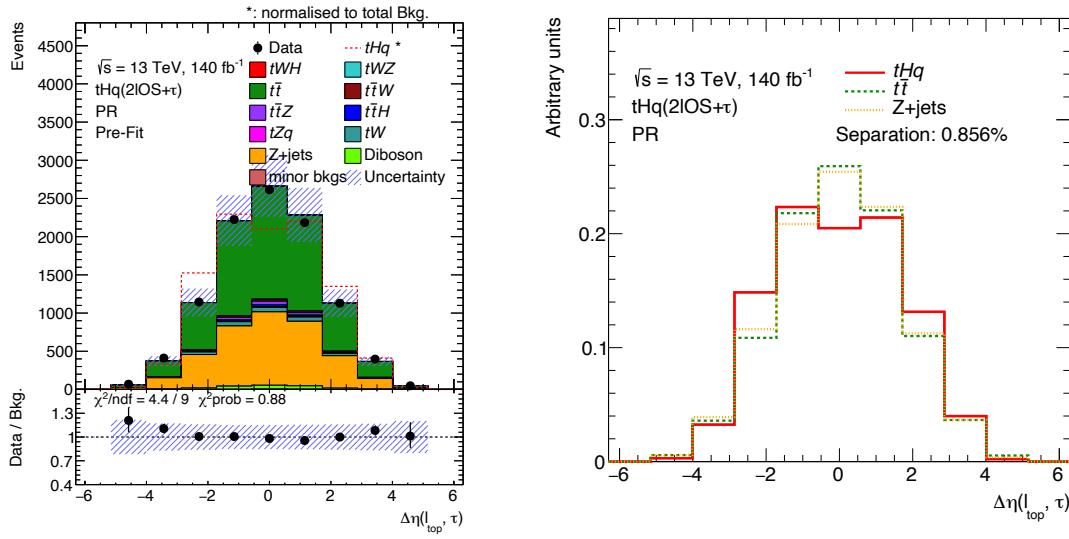


Figure C.24: Distribution and separation plot for the ΔR distance between the ℓ_{top} and τ_{had} .

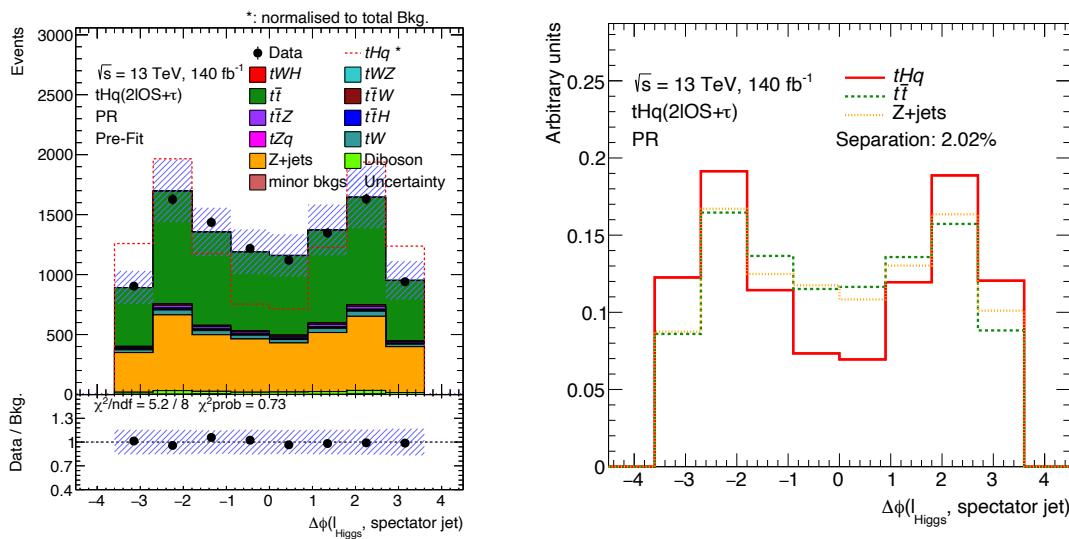


Figure C.25: Distribution and separation plot for the $\Delta\phi$ distance between the ℓ_{Higgs} and the spectator jet.

C.3 BDT($t\bar{t}|_{\text{OS}}$)

This model aims to separate the $t\bar{t}$ and $Z + \text{jets}$ processes to define dedicated regions. Therefore, the variables presented are those that can discriminate between these two backgrounds.

As well as for the distributions above, the uncertainties included here correspond to both the statistical and systematic ones, and the lower panel presents the ratio between the collected data and the MC simulation. The variables are presented in their order of importance for the BDT($t\bar{t}|_{\text{OS}}$), as Figure 6.17b presents.

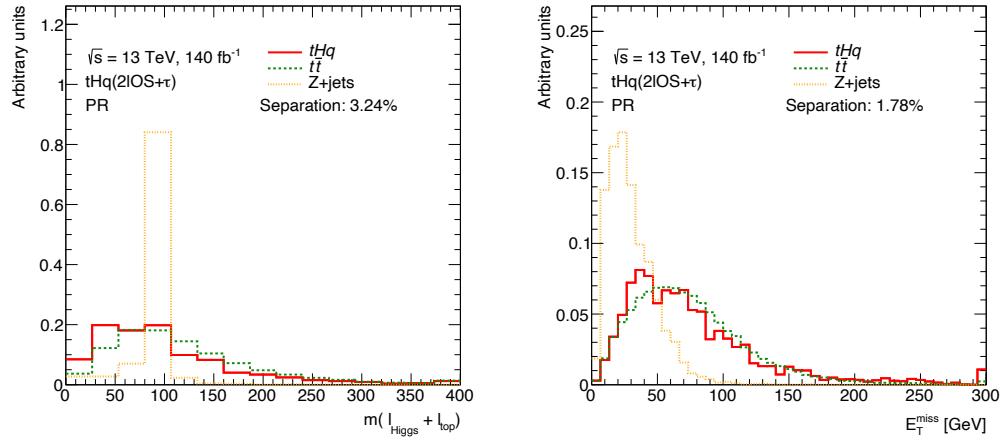


Figure C.26: Separation plot for the two input features with highest discriminant power in the BDT($t\bar{t}|_{\text{OS}}$). The distributions for these two variables showing all processes are presented in Figure 6.19. In (a), the invariant mass of the two charged light-flavoured leptons is presented. The peak around $m(\ell_{\text{Higgs}}, \ell_{\text{top}}) = 90 \text{ GeV}$ corresponds to the events in which a Z boson is produced. The mass of the Z is $m_Z = (91.1876 \pm 0.0021) \text{ GeV}$ [10] and when the mass of the light leptons produced in the Z -boson decay is reconstructed, values close to m_Z are obtained. Therefore, this variable allows us to identify $Z + \text{jets}$ processes. The E_T^{miss} separation plot is shown in (b). Note that while this variable provides good discrimination power against $Z + \text{jets}$, it was not able to separate $t\bar{t}$ and tHq . For this reason, in the studies in which E_T^{miss} is introduced for the training of the BDT($tHq|_{\text{OS}}$), the model becomes specialised in discriminating tHq and $Z + \text{jets}$ but the separation between tHq and $t\bar{t}$ gets worse.

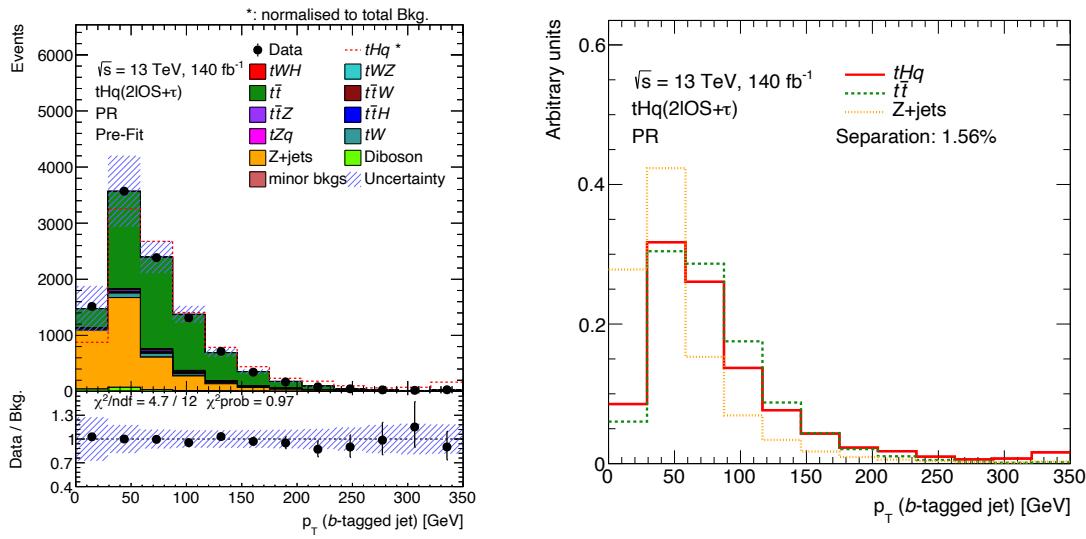


Figure C.27: Distribution and separation plot for the p_T of the leading b -tagged jet.

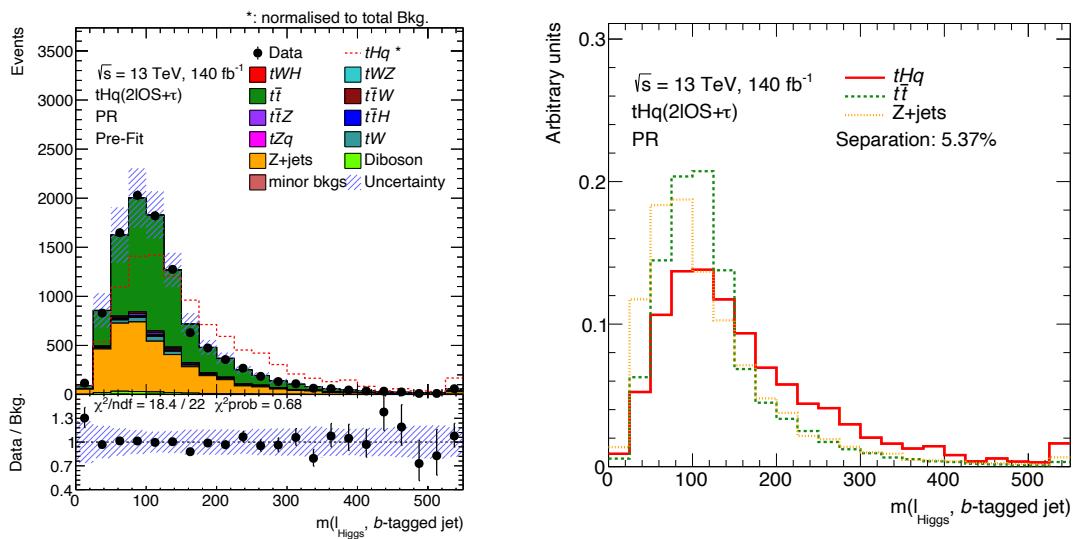


Figure C.28: Distribution and separation plot for the invariant mass of the ℓ_{Higgs} and the leading b -tagged jet. This variable is also present in the training of the BDT($tHq|_{\text{OS}}$).

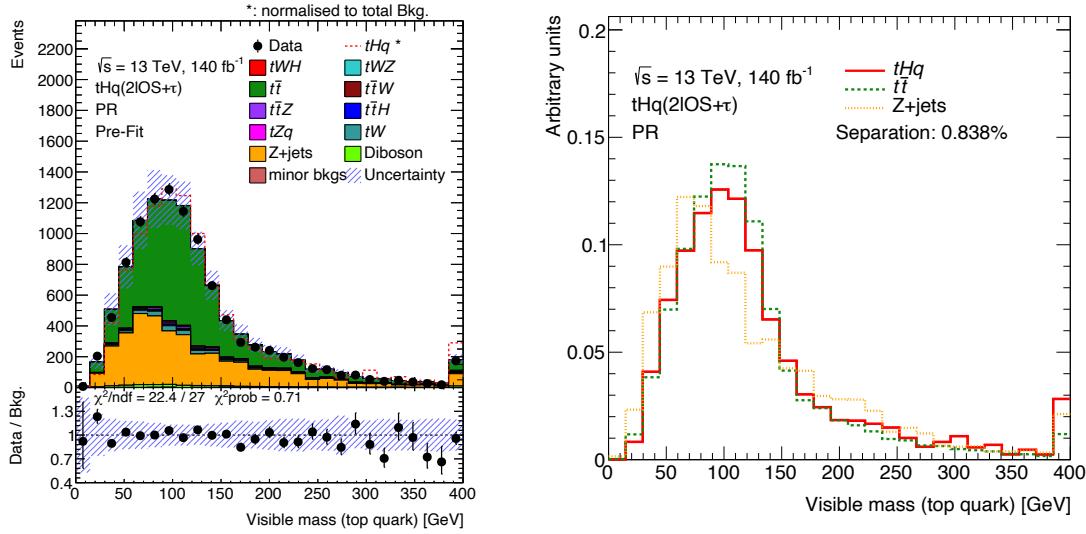


Figure C.29: Distribution and separation plot for the visible mass of the reconstructed top quark. The visible mass refers to the sum of the mass of all particles that are not neutrinos that decayed from the top quark.

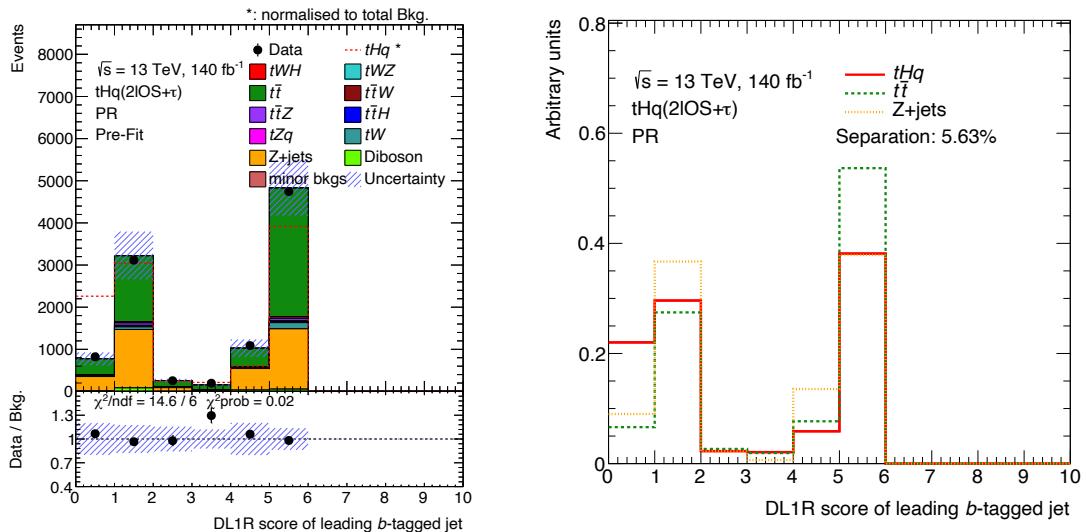


Figure C.30: Distribution and separation plot for binned DL1r score (calibrated) of the leading b -tagged jet [231]. In Equation 5.1 the formulation of the DL1r score is presented.

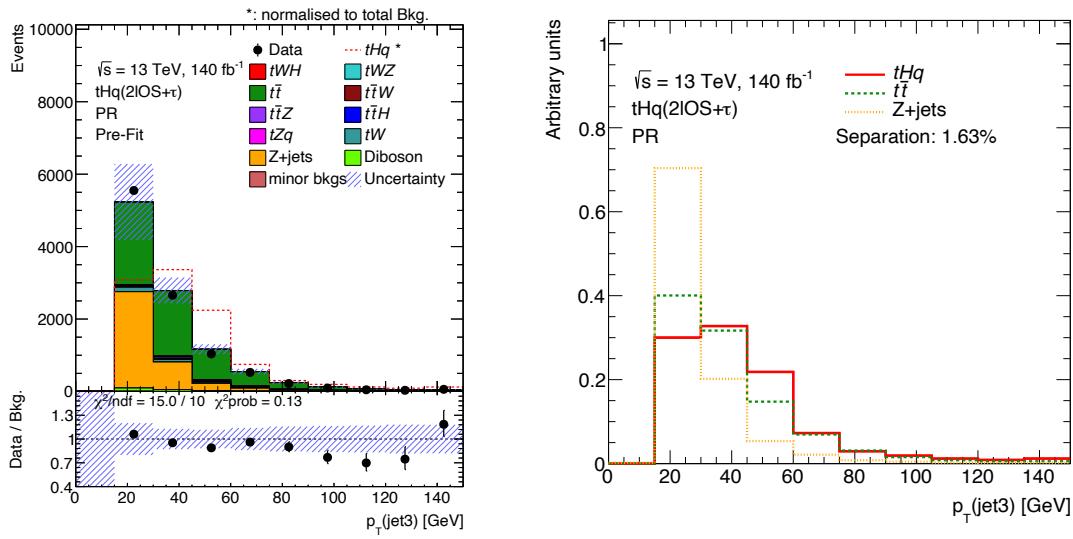


Figure C.31: Distribution and separation plot for the p_T of the softest lepton.

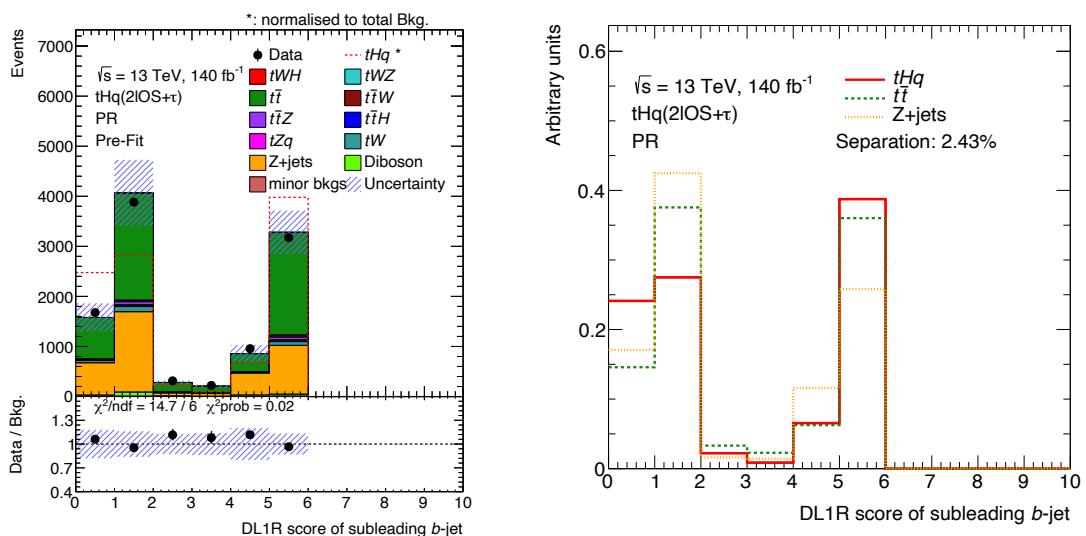


Figure C.32: Distribution and separation plot for binned DL1r score (calibrated) of the sub-leading b -tagged jet.

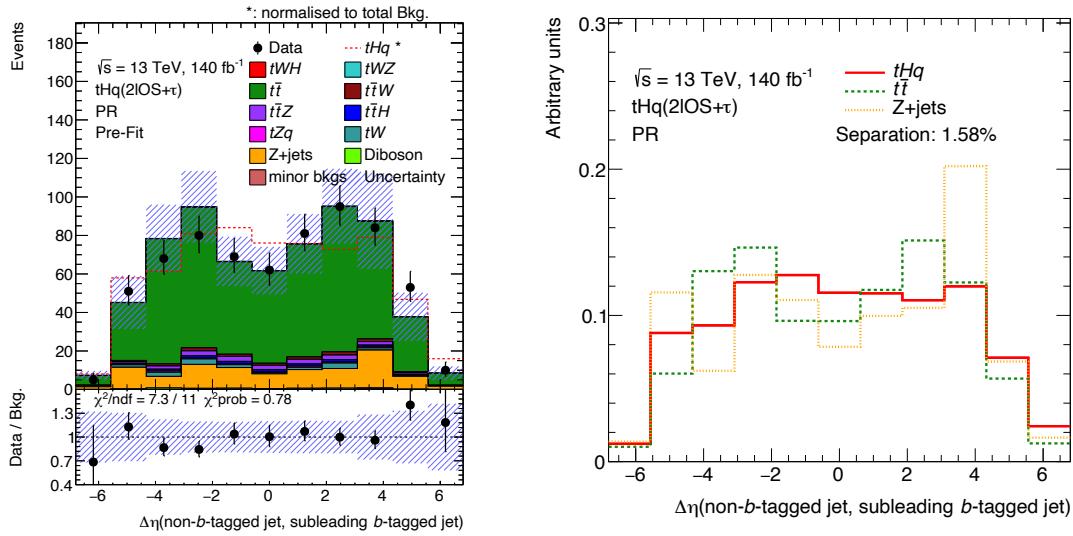


Figure C.33: Distribution and separation plot for the $\Delta\eta$ distance between the leading light jet and the subleading b -tagged jet.

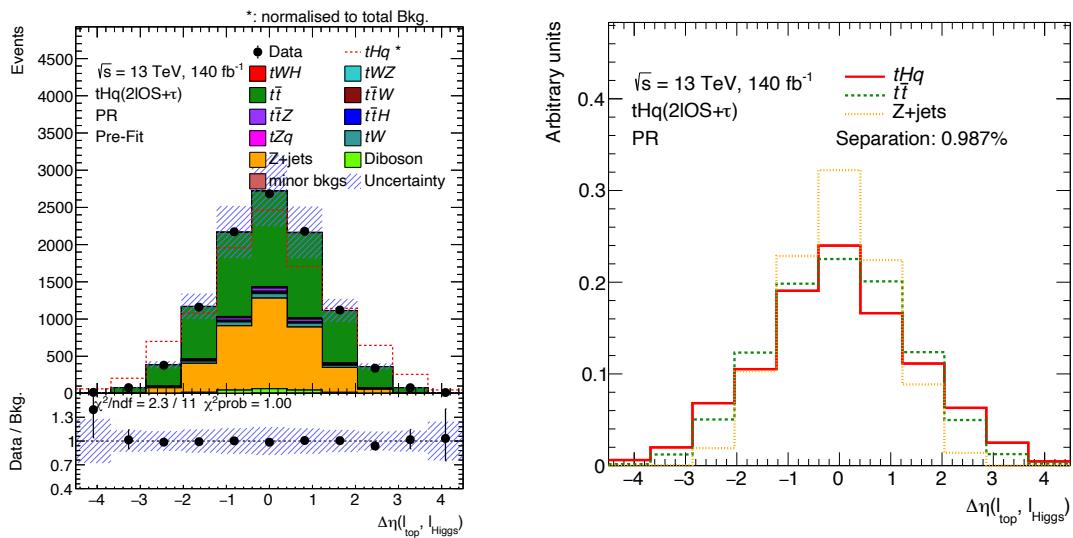


Figure C.34: Distribution and separation plot for the $\Delta\eta$ distance between the two light-flavoured-charged leptons. This variable is also used in the $\text{BDT}(tHq|_{SS})$.

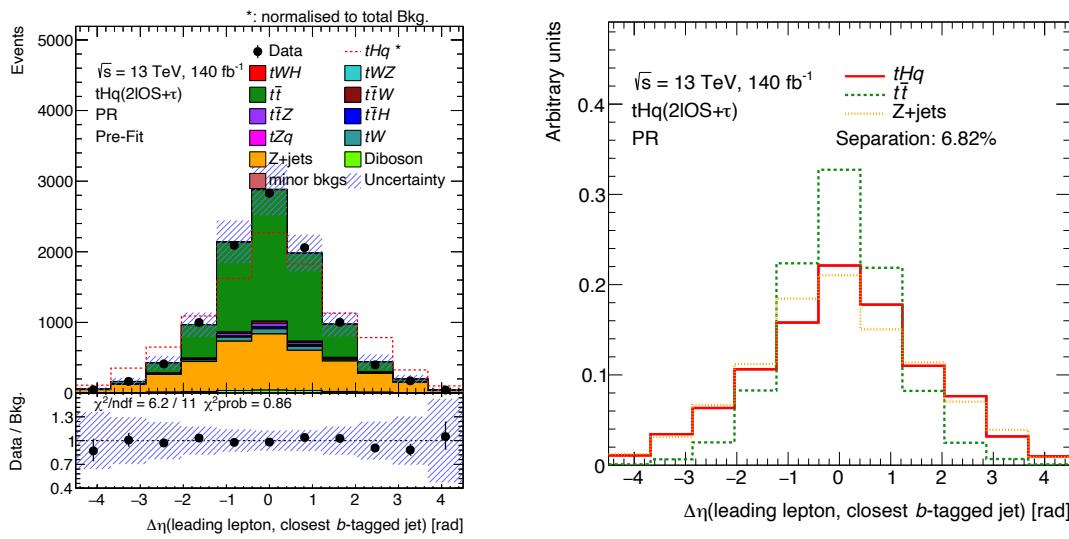


Figure C.35: Distribution and separation plot for the $\Delta\eta$ distance between the leading lepton and the closest b -tagged jet. Note that this variable was also used in the training of the BDT^{Lepton Assignment} (see Figure C.6).

C.4 BDT($tHq|_{\text{SS}}$)

The uncertainty bands depicted here encompass both statistical and systematic uncertainties. The lower panel illustrates the ratio between the actual data and the MC-simulated samples. The variables are presented in their order of importance within the BDT($tHq|_{\text{SS}}$). This ranking is presented in Figure 6.17c)

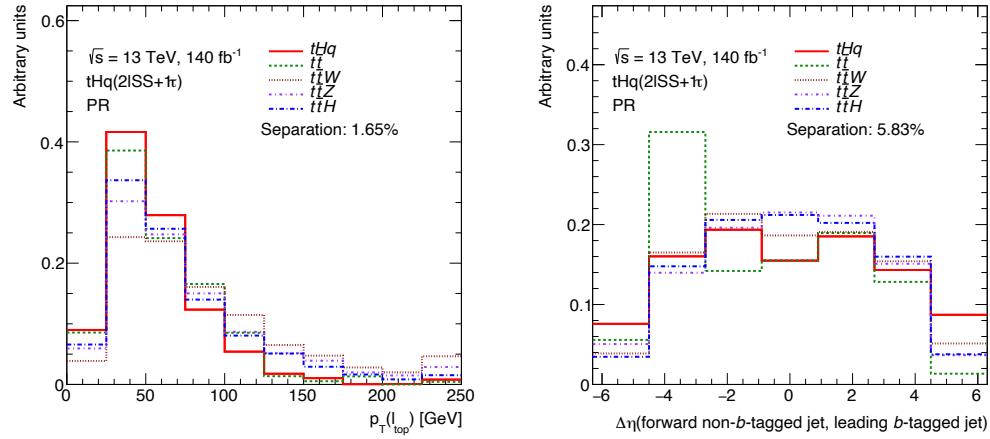


Figure C.36: Separation plot for the two input variables with largest separation power. In (a), the p_T of the light-flavoured lepton originated from the top-quark-decay system is presented. In (b) is shown the $\Delta\eta$ distance between the forward non- b -tagged jet and the leading b -tagged jet. The forward jet is typically the spectator jet and the b -tagged jet is the one that originated from the top-quark-decay system. Therefore For the tHq signal, on average, these two jets are more separated than in the backgrounds. This distribution is the same as for Figure C.15 but for the $2\ell\text{SS} + 1\tau_{\text{had}}$ channel. The distributions for these two variables are presented in Figure 6.20.

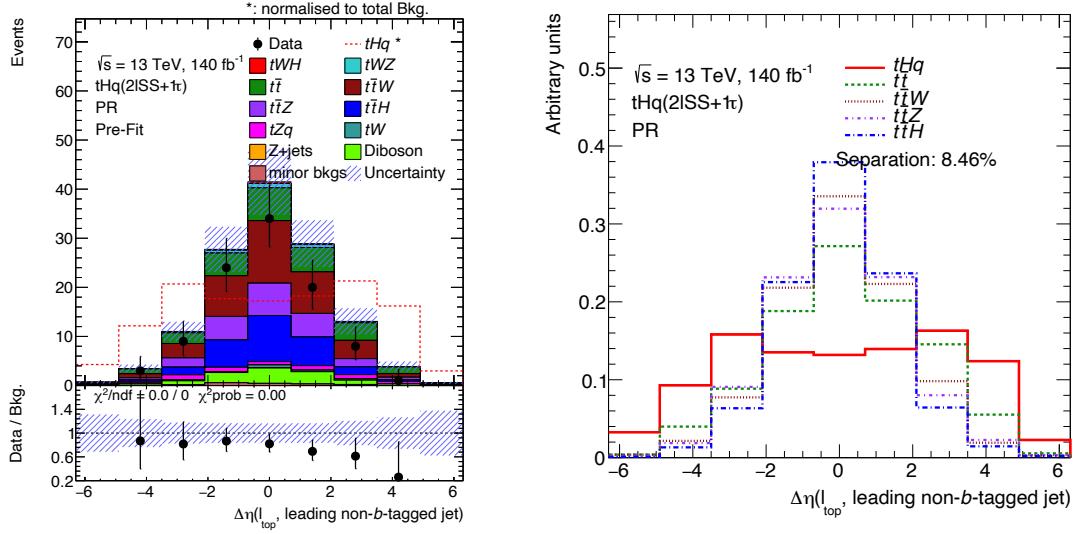


Figure C.37: Distribution and separation plot for the $\Delta\eta$ distance between the ℓ_{top} and the leading-light jet.

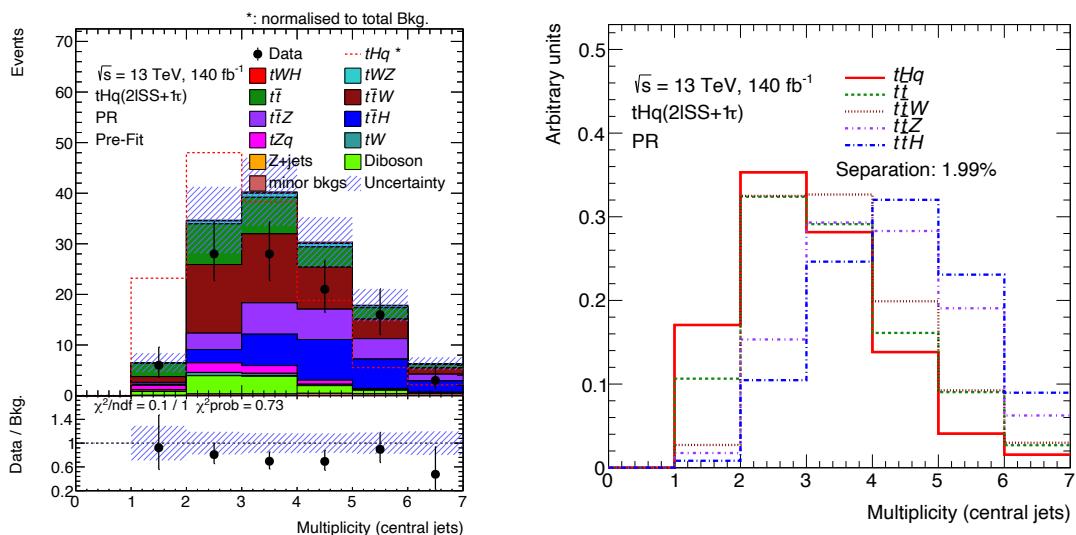


Figure C.38: Distribution and separation plot for the multiplicity of central jets. The central jets are produced in the direction perpendicular to the beam pipe. In this analysis every jet with $|\eta| < 2.5$ is called a central jet.

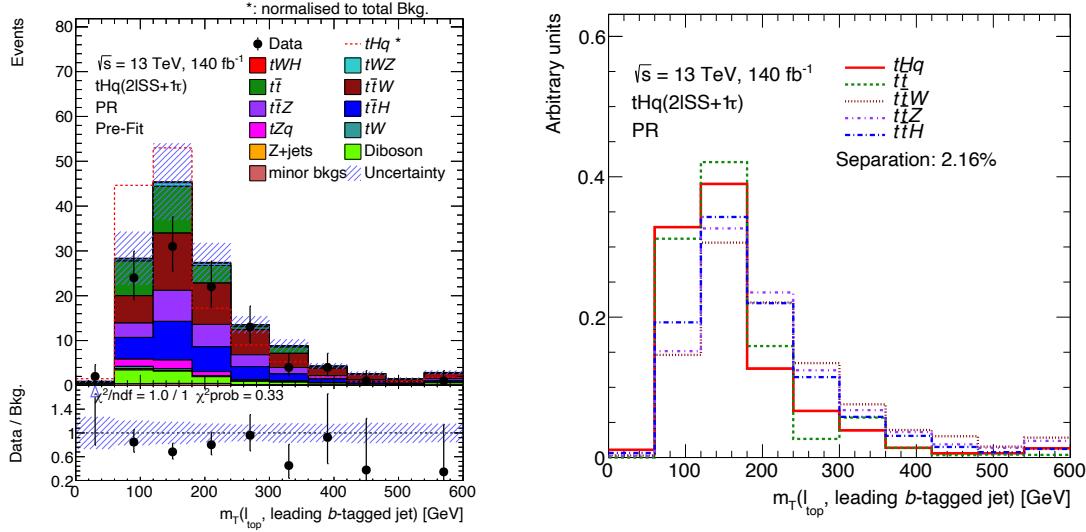


Figure C.39: Distribution and separation plot for the transverse mass of the ℓ_{top} and the leading b -tagged jet. The transverse mass is defined as $m_T = m^2 + p_x^2 + p_y^2 = E^2 - p_z^2$, where the z -direction is along the beam pipe and $p_{x,y}$ are perpendicular to the beam pipe, and m is the invariant mass of the particle. The m_T is a useful quantity because it is invariant under a Lorentz boost along the z -direction.

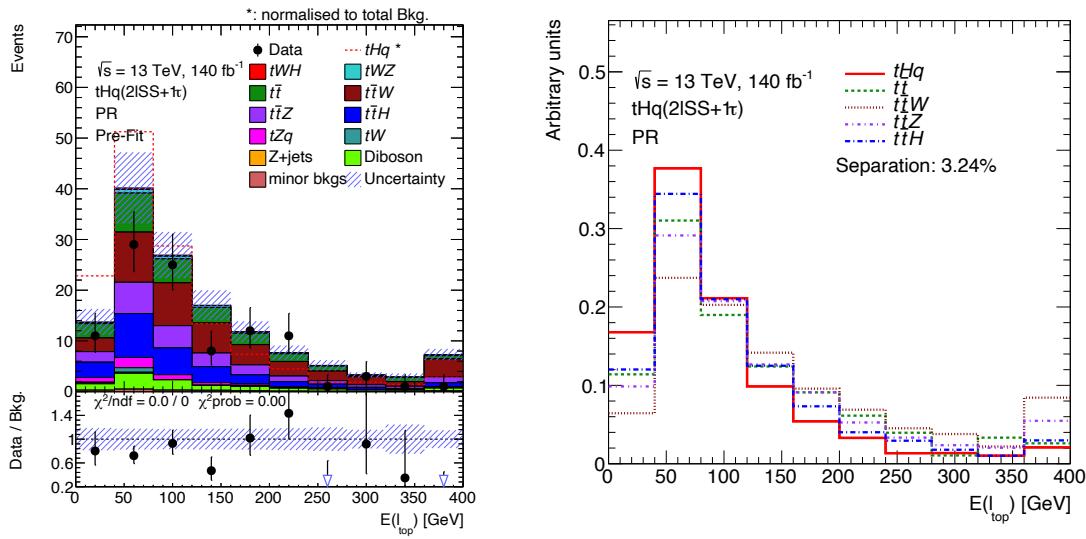


Figure C.40: Distribution and separation plot for the energy of the light-flavoured lepton originated from the top-quark-decay system.

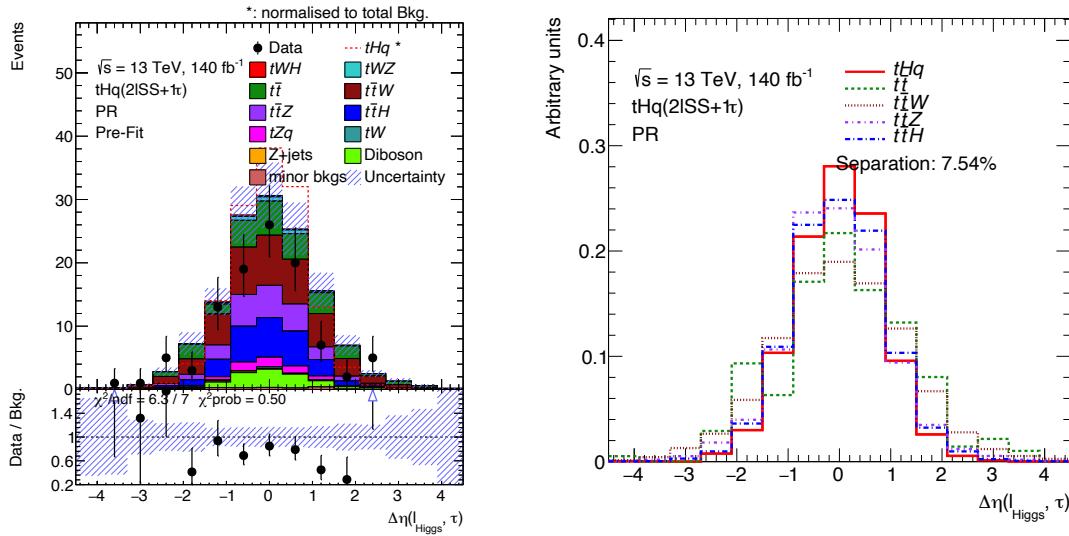


Figure C.41: Distribution and separation plot for the $\Delta\eta$ distance between the ℓ_{Higgs} and the τ_{had} .

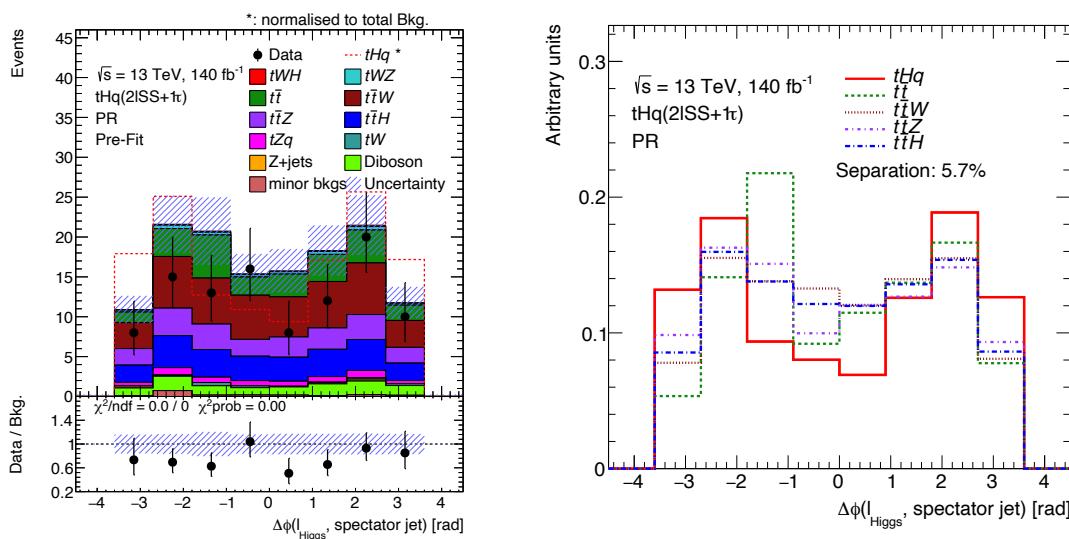


Figure C.42: Distribution and separation plot for the $\Delta\phi$ distance between the ℓ_{Higgs} and the spectator jet.

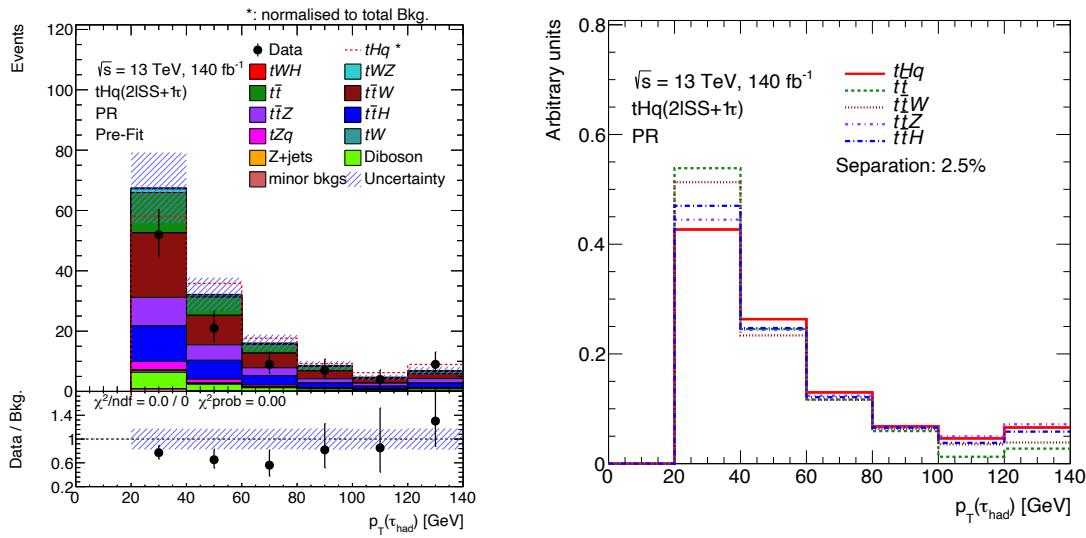


Figure C.43: Distribution and separation plot for the transverse momentum of the hadronically-decaying τ -lepton. This variable is also discriminant for the 2ℓ OS + $1\tau_{\text{had}}$ channel.

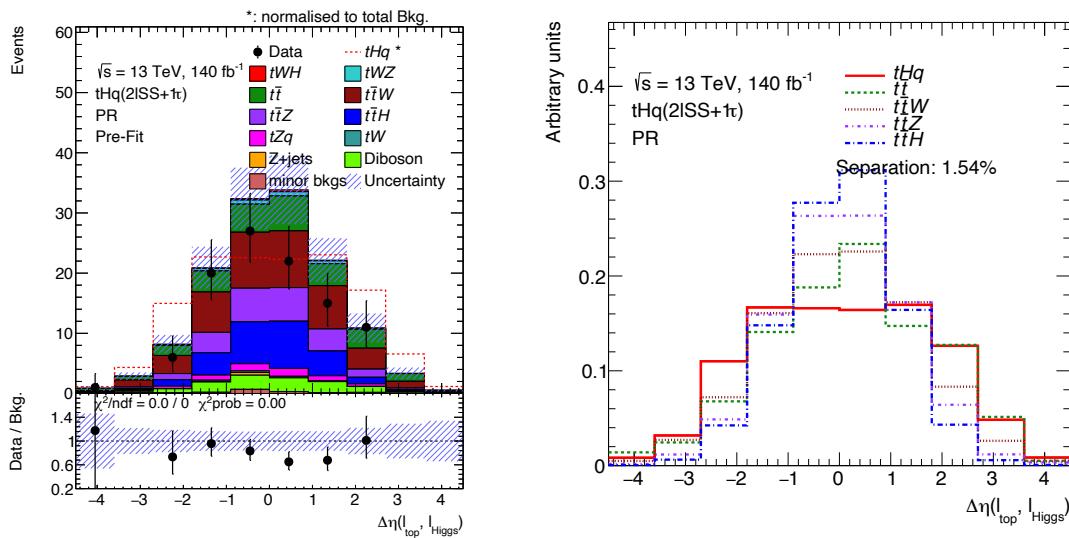


Figure C.44: Distribution and separation plot for the $\Delta\eta$ distance between the two light-flavoured-charged leptons.

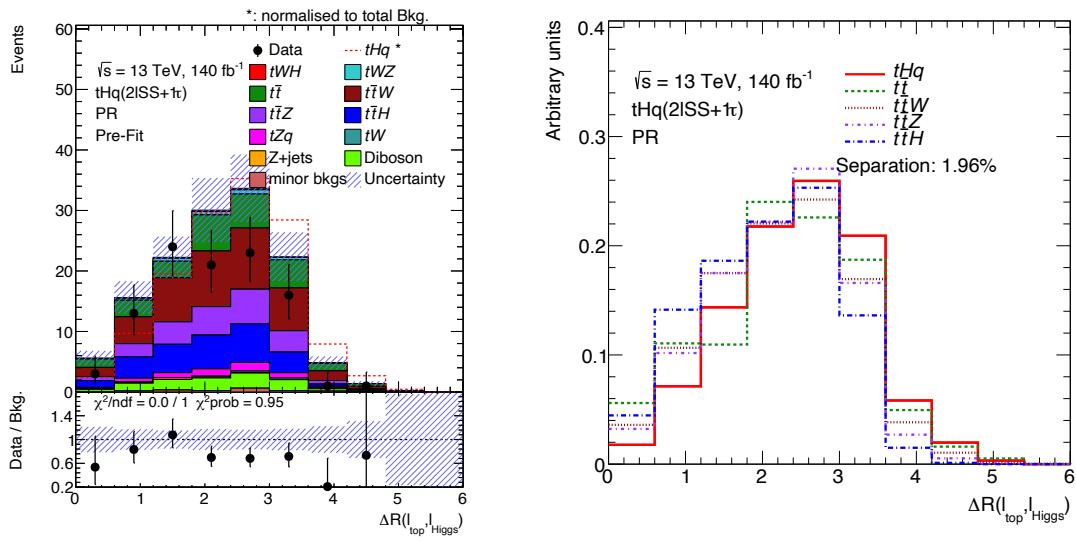


Figure C.45: Distribution and separation plot for the ΔR distance between the two charged light-flavoured leptons.

Appendix D

Effect of negative weights

D.1 Negatively weighted events

The weight in a MC-simulated sample refers to a factor assigned to each event in the simulation to account for various effects such as event generation, detector response, and data-to-simulation discrepancies. These weights are used to scale the simulated events to better match the observed data or to accurately model specific physical processes or background contributions. These negative weights reflect the cancellation of positive and negative contributions to ensure the correct overall probability distribution. The weights are derived based on theoretical calculations, detector simulations, and calibration procedures, and they are crucial for obtaining accurate predictions and comparisons with experimental data in analyses at ATLAS. While negative weights pose challenges for statistical analysis and interpretation, they are necessary to accurately reproduce the expected physics processes and their interference effects.

The negative weights can result in samples with reduced statistical power [317]. Therefore, the presence of negatively-weighted events, as opposed to exclusively positive-weight event samples, implies processing a significantly larger number of events to achieve comparable statistical significance. This issue is particularly pronounced during the final stage of detector simulation, where each event can require hours of CPU time [318].

Another problem with negative weights arises in the training of ML models, as it is further discussed in Appendix B.

D.2 Statistical uncertainty of negative weights

Assume that there is a sample of N MC-simulated events. Of these, a fraction x have negative weights and, therefore, a fraction $(1 - x)$ has a positive weight. The effective number of events is $(N_+ - N_-)$, being $N_+ = (1 - x)N$ the amount of positively weighted events and $N_- = xN$ the same for the negative weights.

The statistical fluctuations are calculated in terms of x and the standard deviation ($\sigma_N = \sqrt{N}$). The number of positive and negative events can fluctuate randomly between $\pm\sigma_-$ for the later and $\pm\sigma_+$ for the former. Here, $\sigma_- = \sqrt{xN} = \sqrt{x}N$ and $\sigma_+ = \sqrt{1-x}N$

The variance ($V = \sigma^2$) of the sample is then

$$V(N_+ - N_-) = xV(N) + (1 - x)V(N) = V(N)$$

and the fractional uncertainty

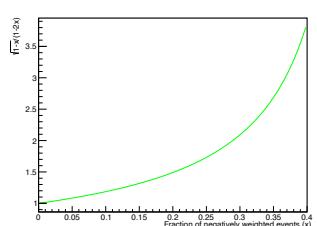
$$\frac{\sigma(N_+ - N_-)}{N_+ - N_-} = \frac{\sigma_N}{(1 - x)N - xN} = \frac{1}{1 - 2x} \frac{\sigma_n}{N}$$

When the fraction of negative events is $x = 0$, $\frac{\sigma(N_+ - N_-)}{N_+ - N_-} = \frac{\sigma_n}{N}$ as expected. In contrast, if $x = 0.5$ the fractional uncertainty is infinite, as expected.

For the signal tHq $2\ell + 1\tau_{\text{had}}$ MC signal sample the fraction of negative weights is between 0.3 and 0.4 depending on the production used.

- $x = 0.3 \rightarrow \frac{\sigma(N_+ - N_-)}{N_+ - N_-} = \frac{1}{0.2} \frac{\sigma_n}{N} = 5.0 \frac{\sigma_n}{N}$
- $x = 0.4 \rightarrow \frac{\sigma(N_+ - N_-)}{N_+ - N_-} = \frac{1}{0.4} \frac{\sigma_n}{N} = 2.5 \frac{\sigma_n}{N}$

The uncertainty of the effective number of events can be compared to that of using only the positively weighted events. If the two fractional uncertainties are divided:



- $x = 0.3 \rightarrow \frac{\frac{\sigma(N_+ - N_-)}{N_+ - N_-}}{\frac{\sigma(N_+)}{N_x}} = 2.09$
- $x = 0.4 \rightarrow \frac{\frac{\sigma(N_+ - N_-)}{N_+ - N_-}}{\frac{\sigma(N_+)}{N_x}} = 3.87$

In Figure D.1, several ΔR distributions are generated using all the events and just the positively weighted ones. As expected, the uncertainty bands are bigger for

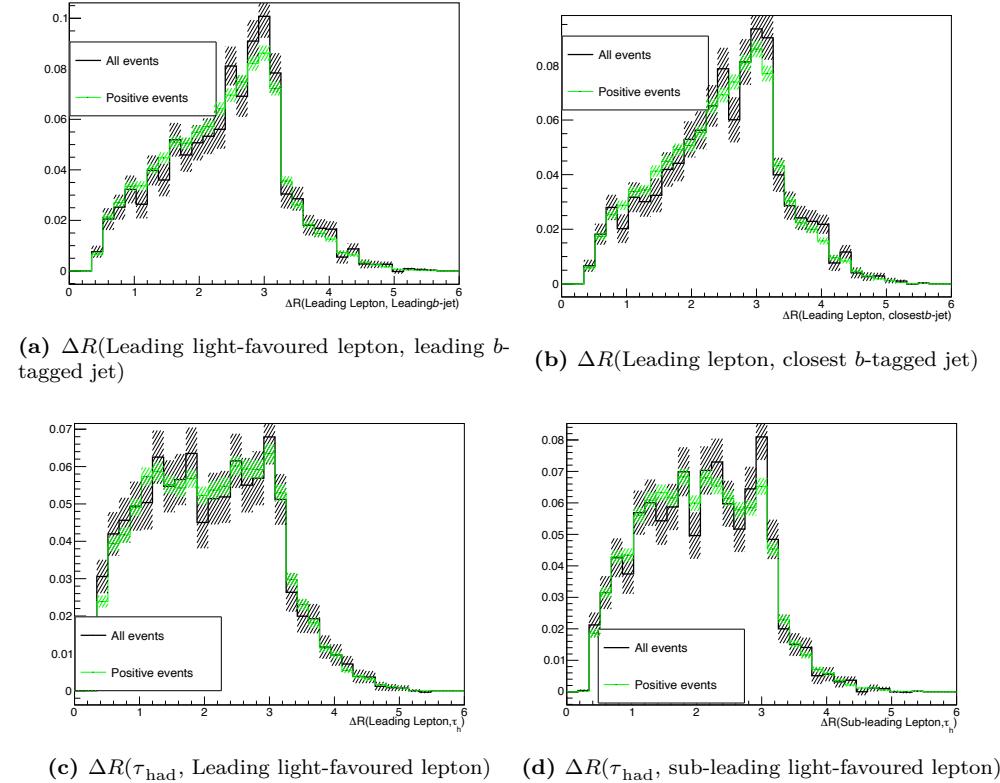


Figure D.1: Some normalised distributions for all the signal events in the $2\ell \text{ SS} + 1\tau_{\text{had}}$ (black) and just the positively weighted events (green). For each bin, the error band is calculated as the square root of the quadratic sum of the weights.

the “All events” than for the “Positive events”. These histograms were produced to verify that using only the events with positive weights in the training of the BDT for the lepton assignment in the SS scenario (Section 6.4.2) was not biasing the result. The size of the error bands is calculated by ROOT as the square root of the quadratic sum of the weights, as explained below.

D.2.1 Errors in binned histograms

If a bin of a histogram has n entries of weighted events w_i with $i = 1, 2, \dots, n$, the size of the bar is $\sum_{i=1}^n w_i$. Therefore, the error of that bar is

$$\sqrt{\sum_{i=1}^n w_i^2} \quad (\text{D.1})$$

This expression for the error of a bin in a histogram is based on error propagation and intrinsic poissonian statistics only. The variance, i.e. the error on the weighted number of events" in that bin, is given by error propagation:

$$V\left(\sum_{i=1}^n w_i\right) = \sqrt{\left(\sum_{i=1}^n w_i^2\right)^2} = \sum_{i=1}^n w_i^2 = \sum_{i=1}^n V(w_i)$$

The variance of the weight w_i , $V(w_i)$, is determined only by the statistical fluctuation of the number of events considered: $V(w_i) = w_i^2$.

D.3 Negative weights in MVA methods

Events coming from the MC generator can be produced with (unphysical) negative weights in some phase-space regions. Such occurrences are frequently inconvenient to deal with, and whether or not they are handled effectively is dependent on the MVA method's actual implementation. Within the ROOT TMVA library, probability and multi-dimensional probability density estimators, as well as BDTs, are among the methods that correctly include occurrences with negative weights. In cases where a method does not properly treat events with negative weights, it is advisable to ignore such events for the training but to include them in the performance evaluation to not bias the results.

Appendix E

Pruning of the non-impactful nuisance parameters

This appendix complements the fit results presented through Section 6.8 presenting the pruning plots for both $2\ell + 1\tau_{\text{had}}$ sub-channels.

E.1 Asimov fit in the $2\ell \text{ OS} + 1\tau_{\text{had}}$ channel

Complementing Section 6.8.4, in this appendix the pruning of the non-impactful NPs is presented. The process of removing certain NPs that have negligible impact on the profile likelihood fit is shown in this section. For all present NPs is evaluated whether to keep the parameter, to drop its normalisation impact, to drop its shape impact or to completely drop the NP. The shape contribution is evaluated by observing the maximum difference in bin entries between the nominal and the varied distribution, under the same normalisation. Regarding the normalisation contribution, it is evaluated by integrating the nominal and varied distribution and measuring the differences.

The pruning is shown first for the instrumental NPs in Figure E.1. Here it can be seen that most of the NPs are completely pruned. Figure E.2 shows that almost all the NPs dedicated to flavour tagging are fully dropped. The JES effective NPs are doped as well but this is not the case for the JER. The pruning of the theory-related NPs is presented in Figures E.3, E.4 and E.5. The PDF uncertainties for $t\bar{t}H$, $t\bar{t}$ and $t\bar{t}W$ are completely dropped. While the PDFs of the diboson and, especially, the tHq production play an important role.

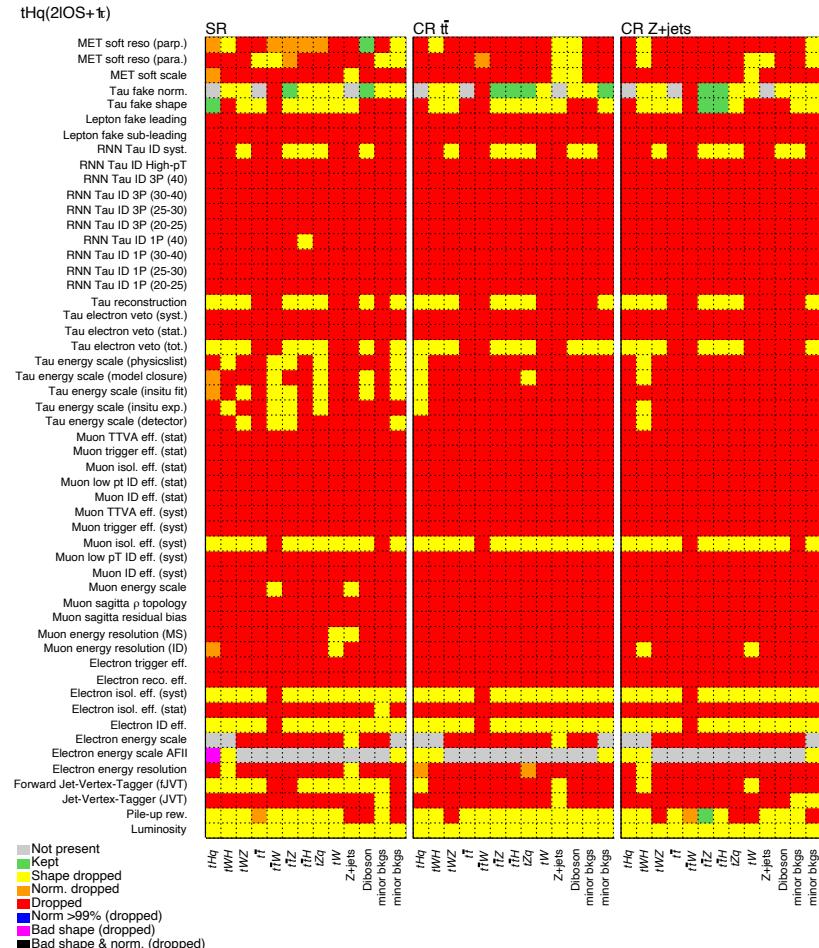


Figure E.1: Pruning of non-impactful instrumental NPs in the Asimov fit of the 2ℓ OS + $1\tau_{\text{had}}$ channel. Grey NPs are not present and green ones are kept. Red combinations are completely dropped. For orange NPs only the shape component is kept, while for yellow ones only the normalisation is kept. Additionally, the list of NPs is split by regions.

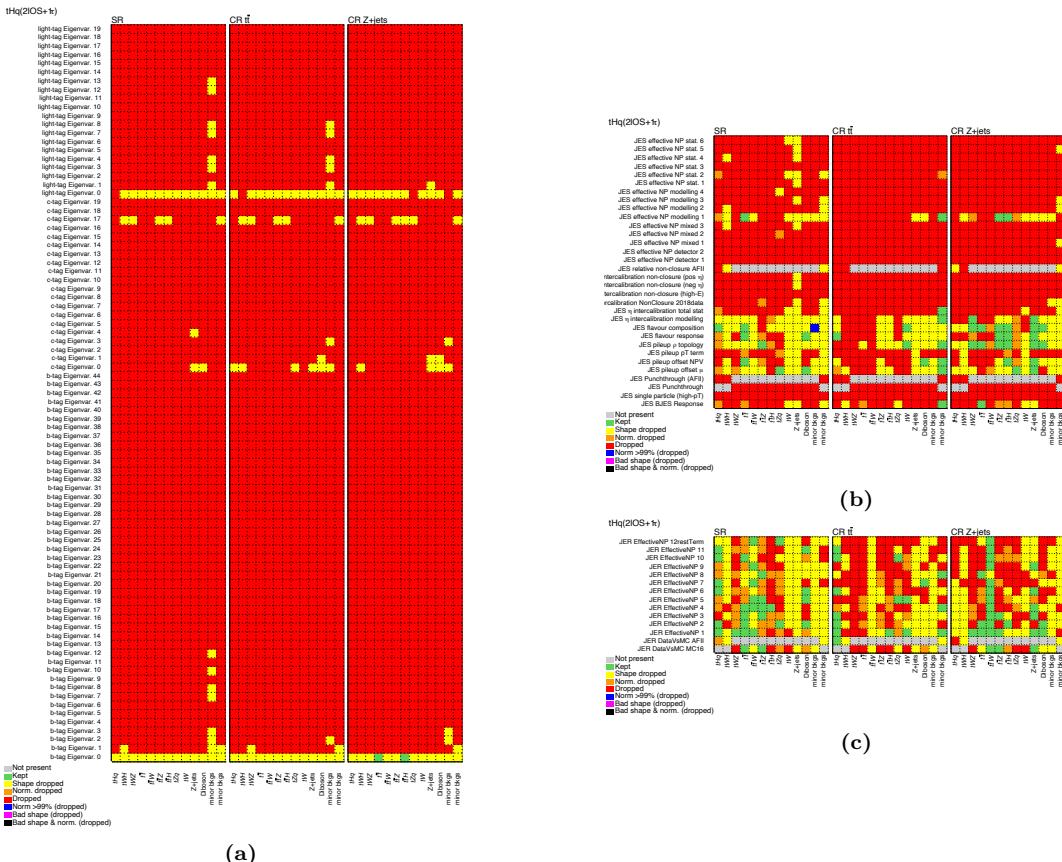


Figure E.2: Pruning of non-impactful (a) flavour-tagging, (b) JES and (c) JER NPs in the Asimov fit of the 2ℓ OS + $1\tau_{\text{had}}$ channel. Grey NPs are not present and green ones are kept. Red combinations are completely dropped. For orange NPs only the shape component is kept, while for yellow ones only the normalisation is kept. Additionally, the list of NPs is split by regions.

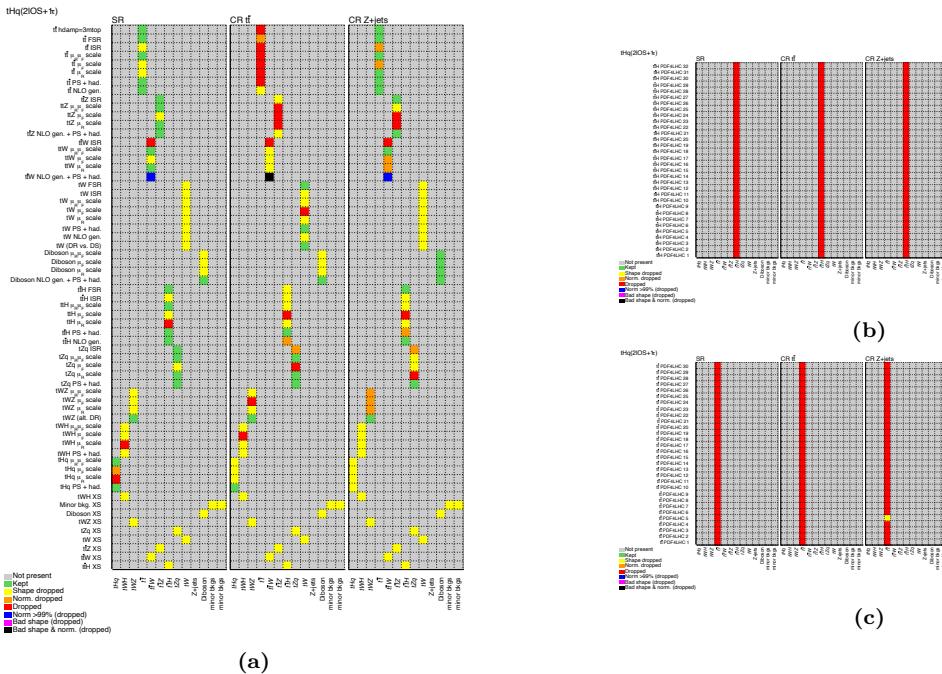


Figure E.3: Pruning of non-impactful (a) modelling NPs as well as the (b) $t\bar{t}H$ and (c) $t\bar{t}$ PDF NPs in the Asimov fit of the 2ℓ OS + $1\tau_{\text{had}}$ channel. The modelling NPs due to PDFs are not included in this figure. Grey NPs are not present and green ones are kept. Red combinations are completely dropped. For orange NPs only the shape component is kept, while for yellow ones only the normalisation is kept. Additionally, the list of NPs is split by regions.

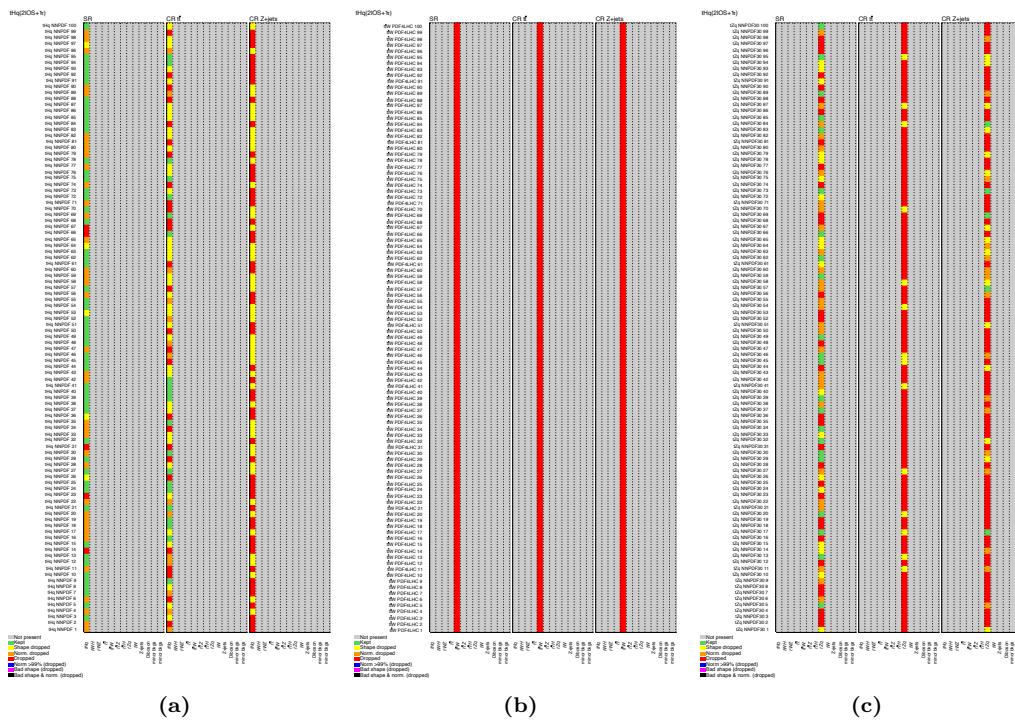


Figure E.4: Pruning of non-impactful (a) tHq , (b) $t\bar{t}W$ and (c) tZq PDF NPs in the Asimov fit of the 2ℓ OS+ $1\tau_{\text{had}}$ channel. Grey NPs are not present and green ones are kept. Red combinations are completely dropped. For orange NPs only the shape component is kept, while for yellow ones only the normalisation is kept. Additionally, the list of NPs is split by regions.

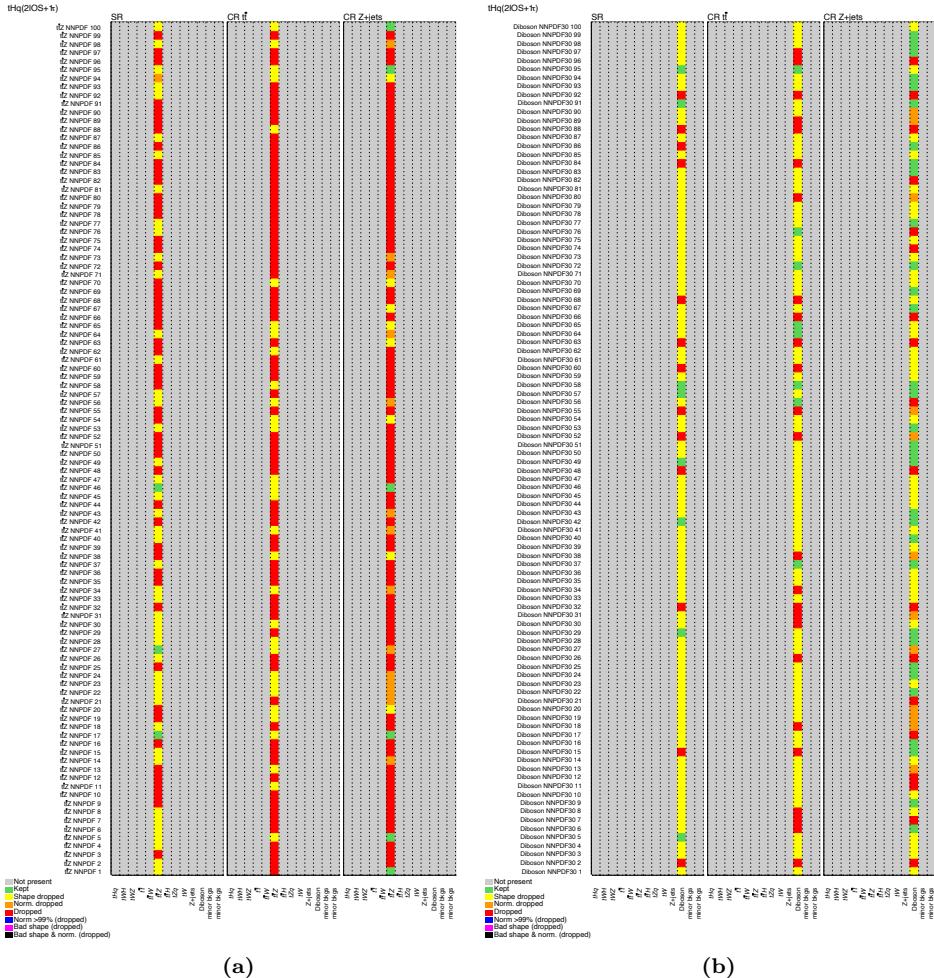


Figure E.5: Pruning of non-impactful (a) $t\bar{t}Z$ and (b) diboson PDF NPs in the Asimov fit of the 2ℓ OS + $1\tau_{\text{had}}$ channel. Grey NPs are not present and green ones are kept. Red combinations are completely dropped. For orange NPs only the shape component is kept, while for yellow ones only the normalisation is kept. Additionally, the list of NPs is split by regions.

E.2 Asimov fit in the 2ℓ SS + $1\tau_{\text{had}}$ channel

Complementing Section 6.8.5, in this appendix the pruning of the non-impactful NPs is presented. The pruning of the experimental NPs in the 2ℓ SS + $1\tau_{\text{had}}$ channel is presented in Figure E.6. For the theory-related NPs is presented these results are shown in Figures E.8, E.9 and E.10. Note that the PDF uncertainties for $t\bar{t}H$, $t\bar{t}$ and $t\bar{t}W$ are completely dropped.



Figure E.6: Pruning of non-impactful instrumental NPs in the Asimov fit of the 2ℓ SS + $1\tau_{\text{had}}$ channel. Grey NPs are not present and green ones are kept. Red combinations are completely dropped. For orange NPs only the shape component is kept, while for yellow ones only the normalisation is kept. Additionally, the list of NPs is split by regions.

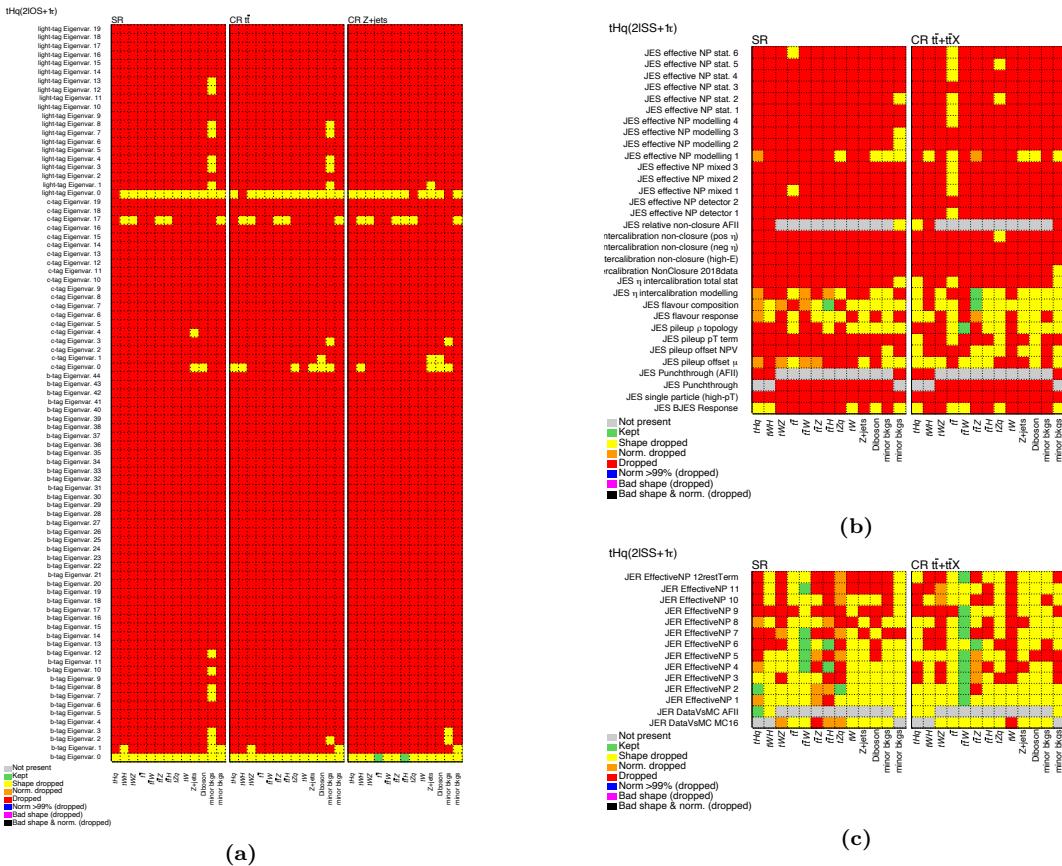


Figure E.7: Pruning of non-impactful (a) flavour-tagging, (b) JES and (c) JER NPs in the Asimov fit of the 2ℓ SS + $1\tau_{\text{had}}$ channel. Grey NPs are not present and green ones are kept. Red combinations are completely dropped. For orange NPs only the shape component is kept, while for yellow ones only the normalisation is kept. Additionally, the list of NPs is split by regions.

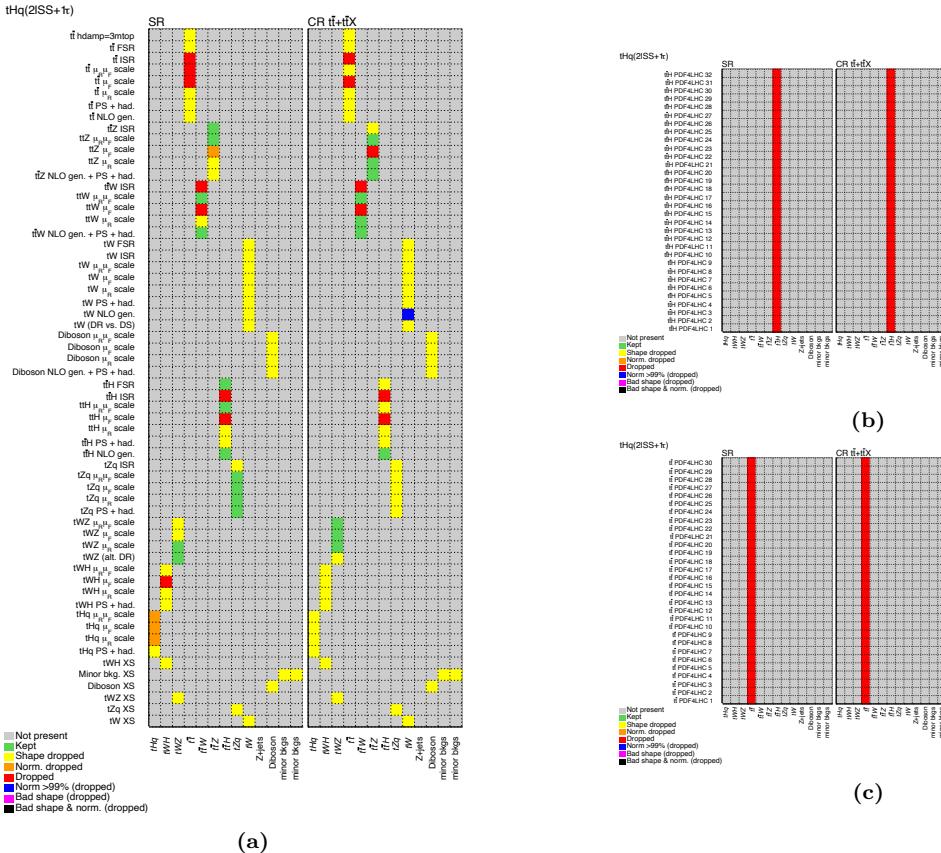


Figure E.8: Pruning of non-impactful (a) modelling NPs as well as the (b) $t\bar{t}H$ and (c) $t\bar{t}$ PDF NPs in the Asimov fit of the 2ℓ SS + $1\tau_{\text{had}}$ channel. The modelling NPs due to PDFs are not included in this figure. Grey NPs are not present and green ones are kept. Red combinations are completely dropped. For orange NPs only the shape component is kept, while for yellow ones only the normalisation is kept. Additionally, the list of NPs is split by regions.

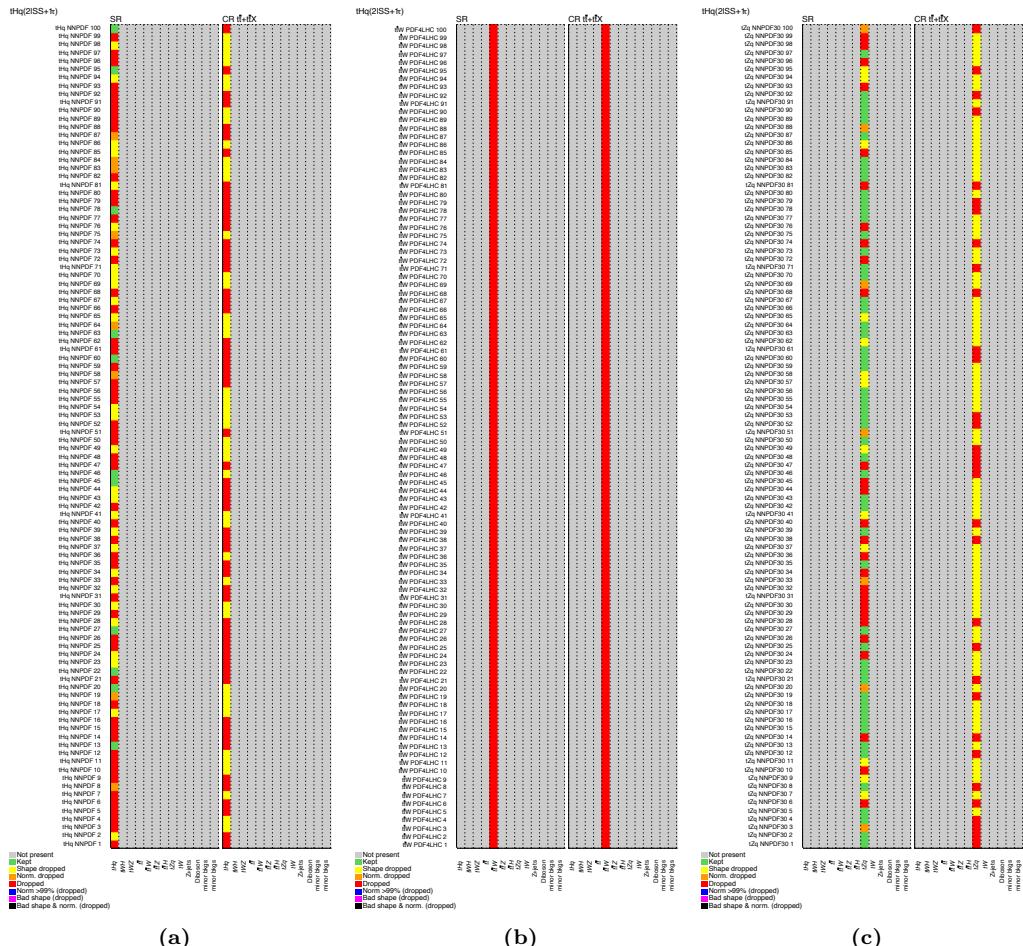


Figure E.9: Pruning of non-impactful (a) tHq , (c) $tTzW$ and (c) tZq PDF NPs in the Asimov fit of the 2ℓ SS+ $1\tau_{\text{had}}$ channel. Grey NPs are not present and green ones are kept. Red combinations are completely dropped. For orange NPs only the shape component is kept, while for yellow ones only the normalisation is kept. Additionally, the list of NPs is split by regions.

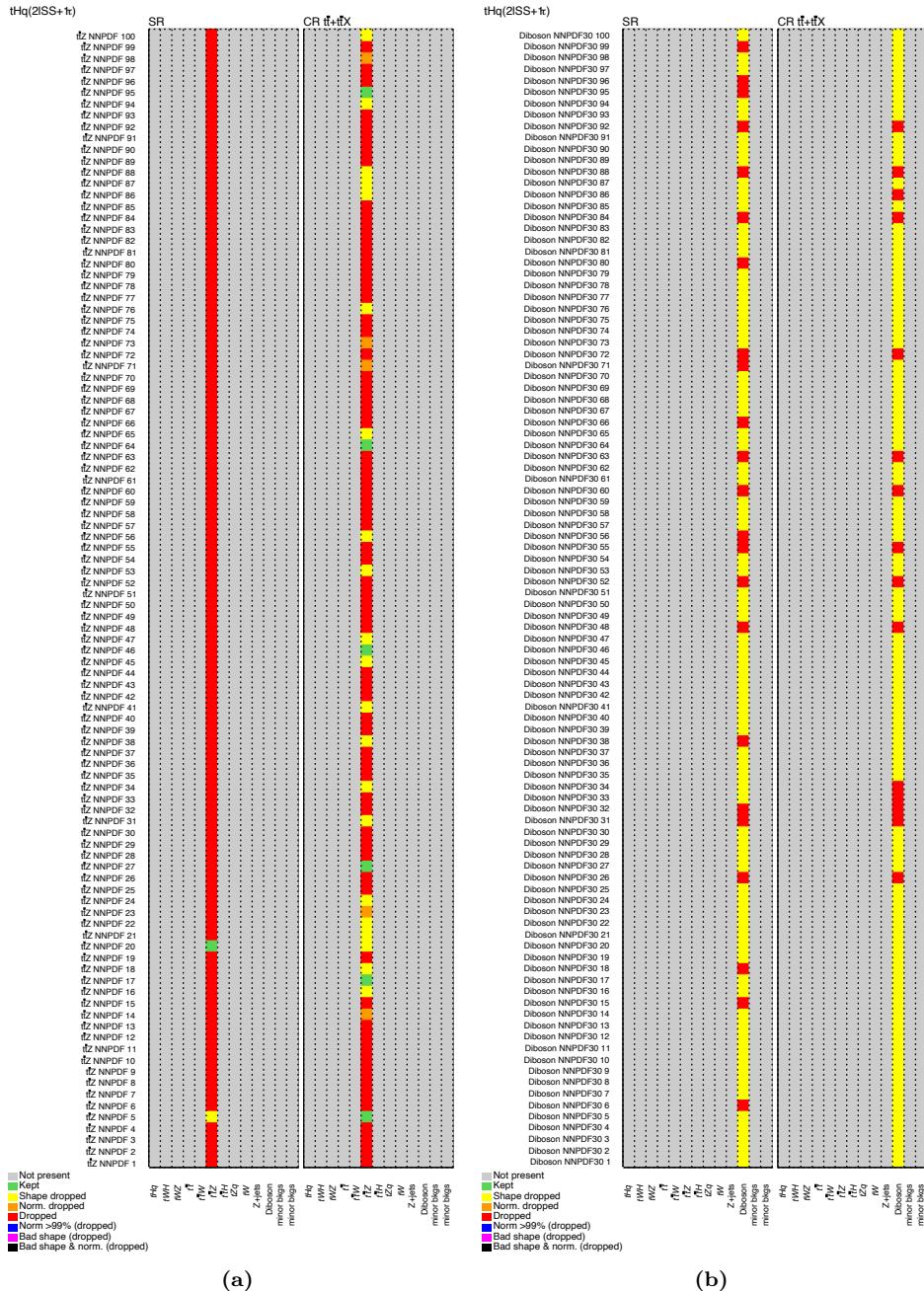


Figure E.10: Pruning of non-impactful (a) $t\bar{t}Z$ and (b) Diboson PDFs NPs in the Asimov fit of the 2ℓ SS + $1\tau_{\text{had}}$ channel. Grey NPs are not present and green ones are kept. Red combinations are completely dropped. For orange NPs only the shape component is kept, while for yellow ones only the normalisation is kept. Additionally, the list of NPs is split by regions.

Bibliography

- [1] Online Etymology Dictionary, "*Physics*", 2022, URL: <https://www.etymonline.com/word/physics> (visited on 15/02/2024).
- [2] Perseus Digital Library, "*fusiko*", 2022, URL: <https://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.04.0057:entry=fusiko/s> (visited on 15/02/2024).
- [3] C. Singer, *A Short History of Science to the Nineteenth Century*, Streeter Press, 1941, ISBN: 0486298876, URL: <https://archive.org/details/inernet.dli.2015.84563>.
- [4] O. Leaman, *Key concepts in Eastern philosophy*, New York: Routledge, 2002, ISBN: 9780415173636, URL: <https://www.routledge.com/Key-Concepts-in-Eastern-Philosophy/Leaman/p/book/9780415173636>.
- [5] C. C. W. Taylor, *The atomists, Leucippus and Democritus: Fragments*, University of Toronto Press, 1999, ISBN: 9781442612129, URL: <http://www.jstor.org/stable/10.3138/9781442671102>.
- [6] A. Purcell, *Go on a particle quest at the first CERN webfest. Le premier webfest du CERN se lance à la conquête des particules*, BUL-NA-2012-269h for the associated production of the Higgs boson, 2012, URL: <https://cds.cern.ch/record/1473657>.
- [7] A. Einstein, *The Foundation of the General Theory of Relativity*, *Annalen Phys.* **49** (1916) 769.
- [8] ATLAS Collaboration, *Measurement of the W -boson mass in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector*, *Eur. Phys. J. C* **78** (2018) 110, arXiv: [1701.07240 \[hep-ex\]](https://arxiv.org/abs/1701.07240), Erratum: *Eur. Phys. J. C* **78** (2018) 898.
- [9] ALEPH Collaboration et al., *Precision electroweak measurements on the Z resonance*, *Phys. Rept.* **427** (2006) 257, arXiv: [hep-ex/0509008](https://arxiv.org/abs/hep-ex/0509008).
- [10] R.L. Workman *et al.* (Particle Data Group), *Review of Particle Physics*, *Prog. Theor. Phys.* (2022) 083C01.

- [11] F. Englert and R. Brout, *Broken Symmetry and the Mass of Gauge Vector Mesons*, Phys. Rev. Lett. **13** (1964) 321.
- [12] P.W. Higgs, *Broken Symmetries and the Masses of Gauge Bosons*, Phys. Rev. Lett. **13** (1964) 508.
- [13] P.A.M. Dirac, *On the Theory of Quantum Mechanics*, Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character **112** (1926) 661, ISSN: 09501207, URL: <http://www.jstor.org/stable/94692>.
- [14] A. Salam, *Weak and Electromagnetic Interactions*, Conf. Proc. C **680519** (1968) 367.
- [15] E. Noether, *Invariante Variationsprobleme*, Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse **1918** (1918) 235.
- [16] R.F. Streater and A.S. Wightman, *PCT, spin and statistics, and all that*, 1989, ISBN: 978-0-691-07062-9.
- [17] J.S. Bell, *Time reversal in field theory*, Proc. Roy. Soc. Lond. A **231** (1955) 479.
- [18] L. Chau and W. Keung, *Comments on the Parametrization of the Kobayashi-Maskawa Matrix*, Phys. Rev. Lett. **53** (1984) 1802.
- [19] O.W. Greenberg, *Spin and Unitary Spin Independence in a Paraquark Model of Baryons and Mesons*, Phys. Rev. Lett. **13** (1964) 598.
- [20] F.J. Yndurain, *Limits on the mass of the gluon*, Phys. Lett. B **345** (1995) 524.
- [21] G.S. Guralnik, C.R. Hagen and T.W.B. Kibble, *Global Conservation Laws and Massless Particles*, Phys. Rev. Lett. **13** (1964) 585.
- [22] J. Goldstone, A. Salam and S. Weinberg, *Broken Symmetries*, Phys. Rev. **127** (1962) 965.
- [23] S.D. Bass, A. De Roeck and M. Kado, *The Higgs boson implications and prospects for future discoveries. The Higgs boson – its implications and prospects for future discoveries*, Nature Rev. Phys. **3** (2021) 608, arXiv: [2104.06821](https://arxiv.org/abs/2104.06821).
- [24] CMS Collaboration, *Measurement of the weak mixing angle using the forward-backward asymmetry of Drell-Yan events in pp collisions at 8 TeV*, Eur. Phys. J. C **78** (2018) 701, arXiv: [1806.00863 \[hep-ex\]](https://arxiv.org/abs/1806.00863).
- [25] MuLan Collaboration, *Measurement of the Positive Muon Lifetime and Determination of the Fermi Constant to Part-per-Million Precision*, Phys. Rev. Lett. **106** (2011) 041803, arXiv: [1010.0991 \[hep-ex\]](https://arxiv.org/abs/1010.0991).

- [26] Z. Maki, M. Nakagawa and S. Sakata, *Remarks on the unified model of elementary particles*, *Prog. Theor. Phys.* **28** (1962) 870.
- [27] B. Pontecorvo, *Inverse beta processes and nonconservation of lepton charge*, *Zh. Eksp. Teor. Fiz.* **34** (1957) 247.
- [28] A.D. Sakharov, *Violation of CP Invariance, C asymmetry, and baryon asymmetry of the universe*, *Pisma Zh. Eksp. Teor. Fiz.* **5** (1967) 32.
- [29] KATRIN Collaboration, *Direct neutrino-mass measurement with sub-electronvolt sensitivity*, *Nature Phys.* **18** (2022) 160, arXiv: [2105.08533 \[hep-ex\]](#).
- [30] DUNE Collaboration, *Deep Underground Neutrino Experiment (DUNE), Far Detector Technical Design Report, Volume I Introduction to DUNE*, *JINST* **15** (2020) T08008, arXiv: [2002.02967 \[physics.ins-det\]](#).
- [31] Planck Collaboration, *Planck 2018 results. VI. Cosmological parameters*, *Astron. Astrophys.* **641** (2020) A6, arXiv: [1807.06209 \[astro-ph.CO\]](#), Erratum: *Astron. Astrophys.* **652** (2021) C4.
- [32] S. Weinberg, *A New Light Boson?*, *Phys. Rev. Lett.* **40** (1978) 223.
- [33] F. Wilczek, *Problem of Strong P and T Invariance in the Presence of Instantons*, *Phys. Rev. Lett.* **40** (1978) 279.
- [34] V.C. Rubin and W.K. Ford Jr., *Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions*, *Astrophys. J.* **159** (1970) 379.
- [35] A.N. Taylor, S. Dye, T.J. Broadhurst, N. Benitez and E. van Kampen, *Gravitational lens magnification and the mass of abell 1689*, *Astrophys. J.* **501** (1998) 539, arXiv: [astro-ph/9801158](#).
- [36] Planck Collaboration, *Planck 2015 results. XIII. Cosmological parameters*, *Astron. Astrophys.* **594** (2016) A13, arXiv: [1502.01589 \[astro-ph.CO\]](#).
- [37] E. Majorana, *Teoria simmetrica dell'elettrone e del positrone*, *Nuovo Cim.* **14** (1937) 171.
- [38] G. Veneziano, *Construction of a crossing - symmetric, Regge behaved amplitude for linearly rising trajectories*, *Nuovo Cim. A* **57** (1968) 190.
- [39] H.P. Nilles, *Supersymmetry, Supergravity and Particle Physics*, *Phys. Rept.* **110** (1984) 1.
- [40] N. Arkani-Hamed, S. Dimopoulos and G. R. Dvali, *The Hierarchy problem and new dimensions at a millimeter*, *Phys. Lett. B* **429** (1998) 263, arXiv: [hep-ph/9803315](#).
- [41] I. Béjar Alonso et al., *High-Luminosity Large Hadron Collider (HL-LHC): Technical design report*, *CERN Yellow Reports: Monographs* **10/2020** (2020).

- [42] G. Mahlon and S.J. Parke, *Spin Correlation Effects in Top Quark Pair Production at the LHC*, Phys. Rev. D **81** (2010) 074024, arXiv: [1001.3422 \[hep-ph\]](https://arxiv.org/abs/1001.3422).
- [43] M. Kobayashi and T. Maskawa, *CP Violation in the Renormalizable Theory of Weak Interaction*, Prog. Theor. Phys. **49** (1973) 652.
- [44] S.L. Glashow, J. Iliopoulos and L. Maiani, *Weak Interactions with Lepton-Hadron Symmetry*, Phys. Rev. D **2** (1970) 1285.
- [45] SLAC-SP-017 Collaboration, *Discovery of a Narrow Resonance in e^+e^- Annihilation*, Phys. Rev. Lett. **33** (1974) 1406.
- [46] E288 Collaboration, *Observation of a Dimuon Resonance at 9.5-GeV in 400-GeV Proton-Nucleus Collisions*, Phys. Rev. Lett. **39** (1977) 252.
- [47] CDF Collaboration, *Observation of top quark production in $\bar{p}p$ collisions*, Phys. Rev. Lett. **74** (1995) 2626, arXiv: [hep-ex/9503002](https://arxiv.org/abs/hep-ex/9503002).
- [48] DØ Collaboration, *Observation of the top quark*, Phys. Rev. Lett. **74** (1995) 2632, arXiv: [hep-ex/9503003](https://arxiv.org/abs/hep-ex/9503003).
- [49] S. Alioli et al., *A new observable to measure the top-quark mass at hadron colliders*, Eur. Phys. J. C **73** (2013) 2438, arXiv: [1303.6415 \[hep-ph\]](https://arxiv.org/abs/1303.6415).
- [50] ALEPH Collaboration et al., *Precision Electroweak Measurements and Constraints on the Standard Model*, CERN-PH-EP-2010-095, 2010, arXiv: [1012.2367 \[hep-ex\]](https://arxiv.org/abs/1012.2367), URL: <https://cds.cern.ch/record/1313716>.
- [51] Gfitter Group, *The global electroweak fit at NNLO and prospects for the LHC and ILC*, Eur. Phys. J. C **74** (2014) 3046, arXiv: [1407.3792 \[hep-ph\]](https://arxiv.org/abs/1407.3792).
- [52] G. Degrassi et al., *Higgs mass and vacuum stability in the Standard Model at NNLO*, JHEP **08** (2012) 098, arXiv: [1205.6497 \[hep-ph\]](https://arxiv.org/abs/1205.6497).
- [53] F.L. Bezrukov and M. Shaposhnikov, *The Standard Model Higgs boson as the inflaton*, Phys. Lett. B **659** (2008) 703, arXiv: [0710.3755 \[hep-th\]](https://arxiv.org/abs/0710.3755).
- [54] A. De Simone, M.P. Hertzberg and F. Wilczek, *Running Inflation in the Standard Model*, Phys. Lett. B **678** (2009) 1, arXiv: [0812.4946 \[hep-ph\]](https://arxiv.org/abs/0812.4946).
- [55] ATLAS Collaboration, *Measurement of the top quark mass in the $t\bar{t} \rightarrow \text{lepton+jets}$ channel from $\sqrt{s} = 8$ TeV ATLAS data and combination with previous results*, Eur. Phys. J. C **79** (2019) 290, arXiv: [1810.01772 \[hep-ex\]](https://arxiv.org/abs/1810.01772).
- [56] CMS Collaboration, *Measurement of the Jet Mass Distribution and Top Quark Mass in Hadronic Decays of Boosted Top Quarks in pp Collisions at $\sqrt{s} = 13$ TeV*, Phys. Rev. Lett. **124** (2020) 202001, arXiv: [1911.03800 \[hep-ex\]](https://arxiv.org/abs/1911.03800).

- [57] CDF and DØ Collaborations, *Combination of CDF and DØ results on the mass of the top quark using up 9.7 fb^{-1} at the Tevatron*, FERMILAB-CONF-16-298-E, 2016, arXiv: [1608.01881 \[hep-ex\]](https://arxiv.org/abs/1608.01881), URL: <https://inspirehep.net/literature/1479761>.
- [58] M. Czakon and A. Ferroglio, *Top quark pair production at complete NLO accuracy with NNLO +NNLL corrections in QCD*, Chinese Physics C **44** (2020) 083104.
- [59] M. Czakon, P. Fiedler and A. Mitov, *Total Top-Quark Pair-Production Cross Section at Hadron Colliders Through $O(\alpha_S^4)$* , Phys. Rev. Lett. **110** (2013) 252004, arXiv: [1303.6254 \[hep-ph\]](https://arxiv.org/abs/1303.6254).
- [60] M. Czakon, D. Heymes and A. Mitov, *Dynamical scales for multi-TeV top-pair production at the LHC*, JHEP **04** (2017) 071, arXiv: [1606 . 03350 \[hep-ph\]](https://arxiv.org/abs/1606.03350).
- [61] ATLAS Collaboration, *Inclusive and differential cross-sections for dilepton $t\bar{t}$ production measured in $\sqrt{s} = 13$ TeV pp collisions with the ATLAS detector*, JHEP **07** (2023) 141, arXiv: [2303.15340 \[hep-ex\]](https://arxiv.org/abs/2303.15340).
- [62] J. Campbell, T. Neumann and Z. Sullivan, *Single-top-quark production in the t-channel at NNLO*, JHEP **02** (2021) 040, arXiv: [2012.01574 \[hep-ph\]](https://arxiv.org/abs/2012.01574).
- [63] ATLAS Collaboration, *Measurement of t-channel production of single top quarks and antiquarks in pp collisions at 13 TeV using the full ATLAS Run 2 dataset*, ATLAS-CONF-2023-026, 2023, URL: <https://cds.cern.ch/record/2860644>.
- [64] N. Kidonakis and N. Yamanaka, *Higher-order corrections for tW production at high-energy hadron colliders*, JHEP **05** (2021) 278, arXiv: [2102 . 11300 \[hep-ph\]](https://arxiv.org/abs/2102.11300).
- [65] CMS Collaboration, *Measurement of inclusive and differential cross sections for single top quark production in association with a W boson in proton-proton collisions at $\sqrt{s} = 13$ TeV*, JHEP **07** (2023) 046, arXiv: [2208.00924 \[hep-ex\]](https://arxiv.org/abs/2208.00924).
- [66] P. Kant et al., *HatHor for single top-quark production: Updated predictions and uncertainty estimates for single top-quark production in hadronic collisions*, Comput. Phys. Commun. **191** (2015) 74, arXiv: [1406 . 4403 \[hep-ph\]](https://arxiv.org/abs/1406.4403).
- [67] ATLAS Collaboration, *Evidence for $t\bar{t}\bar{t}\bar{t}$ production in the multilepton final state in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Eur. Phys. J. C **80** (2020) 1085, arXiv: [2007 . 14858 \[hep-ex\]](https://arxiv.org/abs/2007.14858).
- [68] ATLAS Collaboration, *Measurement of the charge asymmetry in top-quark pair production in association with a photon with the ATLAS experiment*, Phys. Lett. B **843** (2023) 137848, arXiv: [2212 . 10552 \[hep-ex\]](https://arxiv.org/abs/2212.10552).

- [69] ATLAS Collaboration, *Measurement of the polarisation of W bosons produced in top-quark decays using dilepton events at $s=13$ TeV with the ATLAS experiment*, *Phys. Lett. B* **843** (2023) 137829, arXiv: [2209.14903 \[hep-ex\]](#).
- [70] ATLAS Collaboration, *Measurement of the energy asymmetry in $t\bar{t}j$ production at 13 TeV with the ATLAS experiment and interpretation in the SMEFT framework*, *Eur. Phys. J. C* **82** (2022) 374, arXiv: [2110.05453 \[hep-ex\]](#).
- [71] ATLAS Collaboration, *Measurements of top-quark pair spin correlations in the $e\mu$ channel at $\sqrt{s} = 13$ TeV using pp collisions in the ATLAS detector*, *Eur. Phys. J. C* **80** (2020) 754, arXiv: [1903.07570 \[hep-ex\]](#).
- [72] CMS Collaboration, *Measurement of the top quark polarization and $t\bar{t}$ spin correlations using dilepton final states in proton-proton collisions at $\sqrt{s} = 13$ TeV*, *Phys. Rev. D* **100** (2019) 072002, arXiv: [1907.03729 \[hep-ex\]](#).
- [73] ATLAS Collaboration, *Measurement of the polarisation of single top quarks and antiquarks produced in the t-channel at $\sqrt{s} = 13$ TeV and bounds on the tWb dipole operator from the ATLAS experiment*, *JHEP* **11** (2022) 040, arXiv: [2202.11382 \[hep-ex\]](#).
- [74] P. Martínez-Agulló, *Optimisation of selection criteria of t-channel single-top-quark events at $\sqrt{s}=13$ TeV for studies of anomalous couplings in the Wtb vertex*, Master thesis: Universitat de València, 2017, URL: <https://cds.cern.ch/record/2285874>.
- [75] ATLAS Collaboration, *Measurement of the Higgs boson mass in the $H \rightarrow ZZ^* \rightarrow 4\ell$ and $H \rightarrow \gamma\gamma$ channels with $\sqrt{s} = 13$ TeV pp collisions using the ATLAS detector*, *Phys. Lett. B* **784** (2018) 345, arXiv: [1806.00242 \[hep-ex\]](#).
- [76] CMS Collaboration, *A measurement of the Higgs boson mass in the di-photon decay channel*, *Phys. Lett. B* **805** (2020) 135425, arXiv: [2002.06398 \[hep-ex\]](#).
- [77] 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** (2012) 1, arXiv: [1207.7214 \[hep-ex\]](#).
- [78] CMS Collaboration, *Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC*, *Phys. Lett. B* **716** (2012) 30, arXiv: [1207.7235 \[hep-ex\]](#).
- [79] CDF Collaboration, *Combined search for the standard model Higgs boson decaying to a bb pair using the full CDF data set*, *Phys. Rev. Lett.* **109** (2012) 111802, arXiv: [1207.1707 \[hep-ex\]](#).

- [80] DØ Collaboration, *Combined Search for the Standard Model Higgs Boson Decaying to bb Using the D0 Run II Data Set*, Phys. Rev. Lett. **109** (2012) 121802, arXiv: [1207.6631 \[hep-ex\]](https://arxiv.org/abs/1207.6631).
- [81] SLAC Linear Collider Conceptual Design Report, SLAC-0229, 1980, URL: <https://www.osti.gov/biblio/5074122>.
- [82] LEP design report, CERN-LEP/84-01, 1984, URL: <https://cds.cern.ch/record/102083>.
- [83] D. de Florian et al., *Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector*, CERN Yellow Reports: Monographs **2/2017** (2016), arXiv: [1610.07922 \[hep-ph\]](https://arxiv.org/abs/1610.07922).
- [84] M. Farina, C. Grojean, F. Maltoni, E. Salvioni and A. Thamm, *Lifting degeneracies in Higgs couplings using single top production in association with a Higgs boson*, JHEP **05** (2013) 022, arXiv: [1211.3736 \[hep-ph\]](https://arxiv.org/abs/1211.3736).
- [85] S. Biswas, E. Gabrielli and B. Mele, *Single top and Higgs associated production as a probe of the Htt coupling sign at the LHC*, JHEP **01** (2013) 088, arXiv: [1211.0499 \[hep-ph\]](https://arxiv.org/abs/1211.0499).
- [86] ATLAS Collaboration, *Combined measurements of Higgs boson production and decay using up to 139 fb⁻¹ of proton-proton collision data at $\sqrt{s} = 13$ TeV collected with the ATLAS experiment*, ATLAS-CONF-2021-053, 2021, URL: <https://cds.cern.ch/record/2789544>.
- [87] ATLAS Collaboration, *Measurements of the Higgs boson production and decay rates and coupling strengths using pp collision data at $\sqrt{s} = 7$ and 8 TeV in the ATLAS experiment*, ATLAS-CONF-2015-007, 2015, URL: <https://cds.cern.ch/record/2002212>.
- [88] ATLAS Collaboration, *CP Properties of Higgs Boson Interactions with Top Quarks in the t̄H and tH Processes Using H → γγ with the ATLAS Detector*, Phys. Rev. Lett. **125** (2020) 061802, arXiv: [2004.04545 \[hep-ex\]](https://arxiv.org/abs/2004.04545).
- [89] A. Broggio, A. Ferroglio, B.D. Pecjak and L.L. Yang, *NNLL resummation for the associated production of a top pair and a Higgs boson at the LHC*, JHEP **02** (2017) 126, arXiv: [1611.00049 \[hep-ph\]](https://arxiv.org/abs/1611.00049).
- [90] 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** (2018) 173, arXiv: [1806.00425 \[hep-ex\]](https://arxiv.org/abs/1806.00425).
- [91] CMS Collaboeration, *Observation of t̄H production*, Phys. Rev. Lett. **120** (2018) 231801, arXiv: [1804.02610 \[hep-ex\]](https://arxiv.org/abs/1804.02610).

- [92] ATLAS Collaboration, *Search for $H \rightarrow \gamma\gamma$ produced in association with top quarks and constraints on the Yukawa coupling between the top quark and the Higgs boson using data taken at 7 TeV and 8 TeV with the ATLAS detector*, *Phys. Lett. B* **740** (2015) 222, arXiv: [1409.3122 \[hep-ex\]](https://arxiv.org/abs/1409.3122).
- [93] CMS Collaboration, *Search for the associated production of the Higgs boson with a top-quark pair*, *JHEP* **09** (2014) 087, arXiv: [1408.1682 \[hep-ex\]](https://arxiv.org/abs/1408.1682), Erratum: *JHEP* **10** (2014) 106.
- [94] CMS Collaboration, *Measurements of $t\bar{t}H$ Production and the CP Structure of the Yukawa Interaction between the Higgs Boson and Top Quark in the Diphoton Decay Channel*, *Phys. Rev. Lett.* **125** (2020) 061801, arXiv: [2003.10866 \[hep-ex\]](https://arxiv.org/abs/2003.10866).
- [95] CMS Collaboration, *Higgs boson production in association with top quarks in final states with electrons, muons, and hadronically decaying tau leptons at $\sqrt{s} = 13$ TeV*, CMS-PAS-HIG-19-008, 2020, URL: <https://cds.cern.ch/record/2725523>.
- [96] F. Maltoni, G. Ridolfi and M. Ubiali, *b-initiated processes at the LHC: a reappraisal*, *JHEP* **07** (2012) 022, arXiv: [1203.6393 \[hep-ph\]](https://arxiv.org/abs/1203.6393), Erratum: *JHEP* **04** (2013) 095.
- [97] F. Demartin, F. Maltoni, K. Mawatari and M. Zaro, *Higgs production in association with a single top quark at the LHC*, *Eur. Phys. J. C* **75** (2015) 267, arXiv: [1504.00611 \[hep-ph\]](https://arxiv.org/abs/1504.00611).
- [98] F. Demartin, B. Maier, F. Maltoni, K. Mawatari and M. Zaro, *tWH associated production at the LHC*, *Eur. Phys. J. C* **77** (2017) 34, arXiv: [1607.05862 \[hep-ph\]](https://arxiv.org/abs/1607.05862).
- [99] B. Grzadkowski, M. Iskrzynski, M. Misiak and J. Rosiek, *Dimension-Six Terms in the Standard Model Lagrangian*, *JHEP* **10** (2010) 085, arXiv: [1008.4884 \[hep-ph\]](https://arxiv.org/abs/1008.4884).
- [100] T.M.P. Tait and C.P. Yuan, *Single top quark production as a window to physics beyond the standard model*, *Phys. Rev. D* **63** (2000) 014018, arXiv: [hep-ph/0007298](https://arxiv.org/abs/hep-ph/0007298).
- [101] S. Biswas, E. Gabrielli, F. Margaroli and B. Mele, *Direct constraints on the top-Higgs coupling from the 8 TeV LHC data*, *JHEP* **07** (2013) 073, arXiv: [1304.1822 \[hep-ph\]](https://arxiv.org/abs/1304.1822).
- [102] CMS Collaboration, *Measurement of the Higgs boson production rate in association with top quarks in final states with electrons, muons, and hadronically decaying tau leptons at $\sqrt{s} = 13$ TeV*, *Eur. Phys. J. C* **81** (2021) 378, arXiv: [2011.03652 \[hep-ex\]](https://arxiv.org/abs/2011.03652).

- [103] CMS Collaboration, *Search for associated production of a Higgs boson and a single top quark in proton-proton collisions at $\sqrt{s} = 13$ TeV*, Phys. Rev. D **99** (2019) 092005, arXiv: 1811.09696 [hep-ex].
- [104] ATLAS Collaboration, *Measurement of the properties of Higgs boson production at $\sqrt{s} = 13$ TeV in the $H \rightarrow \gamma\gamma$ channel using 139 fb^{-1} of pp collision data with the ATLAS experiment*, ATLAS-CONF-2020-026, 2020, URL: <https://cds.cern.ch/record/2725727>.
- [105] CMS Collaboration, *Measurement of the $t\bar{t}H$ and tH production rates in the $H \rightarrow b\bar{b}$ decay channel with 138 fb^{-1} of proton-proton collision data at $\sqrt{s} = 13$ TeV*, CMS-PAS-HIG-19-011, 2023, URL: <https://cds.cern.ch/record/2868175>.
- [106] ATLAS Collaboration, *ATLAS: Detector and physics performance technical design report. Volume 2*, ATLAS-TDR-15, 1999, URL: <https://cds.cern.ch/record/391176>.
- [107] *Withdrawal of Spain. Retrait de l'Espagne. 41st Session of Council*, 41st Session of Council, 1969, URL: <https://cds.cern.ch/record/22390>.
- [108] EL PAÍS, *España se reintegra al CERN, después de 14 años de ausencia*, EL PAÍS (1982), URL: https://elpais.com/diario/1982/06/28/sociedad/394063206_850215.html (visited on 15/02/2024).
- [109] CERN Web Page, *Our Member States*, 2022, URL: <https://home.web.cern.ch/about/who-we-are/our-governance/member-states> (visited on 15/02/2024).
- [110] T. Fazzini, G. Fidecaro, A.W. Merrison, H. Paul and A.V. Tollestrup, *Electron Decay of the Pion*, Phys. Rev. Lett. **1** (1958) 247.
- [111] Gargamelle Neutrino Collaboration, *Observation of Neutrino Like Interactions Without Muon Or Electron in the Gargamelle Neutrino Experiment*, Phys. Lett. B **46** (1973) 138.
- [112] P.M. Watkins, *Discovery of the W and Z bosons*, Contemporary Physics **27** (1986) 291.
- [113] ALEPH Collaboration, *Determination of the Number of Light Neutrino Species*, Phys. Lett. B **231** (1989) 519.
- [114] PS210 Collaboration, *Production of anti-hydrogen*, Phys. Lett. B **368** (1996) 251.
- [115] NA31 Collaboration, *A New measurement of direct CP violation in the neutral kaon system*, Phys. Lett. B **317** (1993) 233.

- [116] NA48 Collaboration, *A New measurement of direct CP violation in two pion decays of the neutral kaon*, *Phys. Lett. B* **465** (1999) 335, arXiv: [hep-ex/9909022](#).
- [117] LHCb Collaboration, *Observation of $J/\psi p$ Resonances Consistent with Pentaquark States in $\Lambda_b^0 \rightarrow J/\psi K^- p$ Decays*, *Phys. Rev. Lett.* **115** (2015) 072001, arXiv: [1507.03414 \[hep-ex\]](#).
- [118] AD Collaboration, *The Antiproton Decelerator: AD*, *Hyperfine Interact.* **109** (1997) 43.
- [119] CERN Bulletin, *The antimatter factory is ready for another successful year. L'usine à antimatière est prête pour de nouveaux succès*, BUL-NA-2011-125, 2011 7, URL: <https://cds.cern.ch/record/1352088>.
- [120] ISOLDE Collaboration, *The ISOLDE facility*, *J. Phys. G* **44** (2017) 094002.
- [121] AWAKE Collaboration, *AWAKE, The Advanced Proton Driven Plasma Wakefield Acceleration Experiment at CERN*, *Nucl. Instrum. Meth. A* **829** (2016) 76, arXiv: [1512.05498 \[physics.acc-ph\]](#).
- [122] AMS Collaboration, *The Alpha Magnetic Spectrometer (AMS) on the International Space Station. I: Results from the test flight on the space shuttle*, *Phys. Rept.* **366** (2002) 331, Erratum: *Phys. Rept.* **380** (2003) 97.
- [123] L. Evans, *LHC Machine*, *JINST* **3** (2008) S08001.
- [124] CERN Press Office, *CERN Council holds special session on the Large Hadron Collider project*, CERN-PR-91-10-EN, 1991 3, URL: <https://cds.cern.ch/record/859552>.
- [125] LHC Study Group, *The Large Hadron Collider: conceptual design*, CERN-AC-95-05-LHC, 1995, URL: <http://cds.cern.ch/record/291782>.
- [126] S. Myers, *The LEP Collider, from design to approval and commissioning*, John Adams' memorial lecture, 1991, URL: <http://cds.cern.ch/record/226776>.
- [127] E. Wilson and B. J. Holzer, *Beam Dynamics*, Particle Physics Reference Library: Volume 3: Accelerators and Colliders, 2020 15, URL: <https://cds.cern.ch/record/2743947>.
- [128] O. Brüning et al., *LHC Design Report Vol.1: The LHC Main Ring*, CERN-2004-003, 2004, URL: <https://cds.cern.ch/record/782076>.
- [129] J.T. Boyd, *LHC Run-2 and future prospects*, CERN Yellow Rep. School Proc. **5** (2022) 247, arXiv: [2001.04370 \[hep-ex\]](#).
- [130] R. Steerenberg et al., *Operation and performance of the CERN Large Hadron Collider during proton Run 2*, CERN-ACC-2019-067, 2019, URL: <https://cds.cern.ch/record/2696126>.

- [131] L. Lari, H. Gaillard and V. Mertens, *Scheduling the installation of the LHC injection lines*, CERN-TS-2004-017, 2004, URL: <https://cds.cern.ch/record/1069714>.
- [132] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, *JINST* **3** (2008) S08003.
- [133] CMS Collaboration, *The CMS Experiment at the CERN LHC*, *JINST* **3** (2008) S08004.
- [134] LHCb Collaboration, *The LHCb Detector at the LHC*, *JINST* **3** (2008) S08005.
- [135] ALICE Collaboration, *The ALICE experiment at the CERN LHC*, *JINST* **3** (2008) S08002.
- [136] LHCf Collaboration, *The LHCf detector at the CERN Large Hadron Collider*, *JINST* **3** (2008) S08006.
- [137] MATHUSLA Collaboration, *A Letter of Intent for MATHUSLA: A Dedicated Displaced Vertex Detector above ATLAS or CMS.*, CERN-LHCC-2018-025, 2018, arXiv: [1811.00927 \[physics.ins-det\]](https://arxiv.org/abs/1811.00927).
- [138] A. Ball et al., *A Letter of Intent to Install a milli-charged Particle Detector at LHC P5*, 2016, arXiv: [1607.04669 \[physics.ins-det\]](https://arxiv.org/abs/1607.04669).
- [139] MoEDAL Collaboration, *The MoEDAL experiment at the LHC: status and results*, *J. Phys. Conf. Ser.* **873** (2017) 012010, arXiv: [1703.07141 \[hep-ex\]](https://arxiv.org/abs/1703.07141).
- [140] TOTEM Collaboration, *The TOTEM experiment at the CERN Large Hadron Collider*, *JINST* **3** (2008) S08007.
- [141] FASER Collaboration, *Letter of Intent for FASER: ForwArd Search ExpeRiment at the LHC*, CERN-LHCC-2018-030, 2018, arXiv: [1811.10243 \[physics.ins-det\]](https://arxiv.org/abs/1811.10243).
- [142] ATLAS Collaboration, *ATLAS computing: Technical design report*, ATLAS-TRD-017, 2005, URL: <https://cds.cern.ch/record/837738/>.
- [143] WLCG Web Page, *Tier centres*, 2022, URL: <https://wlcg-public.web.cern.ch/tier-centres> (visited on 31/01/2024).
- [144] J. Pequenao, *Computer generated image of the whole ATLAS detector*, CERN-GE-0803012, 2008, URL: <https://cds.cern.ch/record/1095924>.
- [145] ATLAS Collaboration, *ATLAS inner detector: Technical Design Report, 1*, ATLAS-TDR-04, 1997, URL: <https://cds.cern.ch/record/331063>.
- [146] ATLAS Collaboration, *The ATLAS Inner Detector commissioning and calibration*, *Eur. Phys. J. C* **70** (2010) 787, arXiv: [1004 . 5293 \[physics.ins-det\]](https://arxiv.org/abs/1004.5293).

- [147] ATLAS Collaboration, *Studies of radial distortions of the ATLAS Inner Detector*, ATL-PHYS-PUB-2018-003, 2018, URL: <https://cds.cern.ch/record/2309785>.
- [148] ATLAS Collaboration, *Alignment of the ATLAS Inner Detector in Run-2*, Eur. Phys. J. C **80** (2020) 1194, arXiv: 2007.07624 [hep-ex].
- [149] M. Capeans et al., *ATLAS Insertable B-Layer Technical Design Report*, CERN-LHCC-2010-013, ATLAS-TDR-19, 2010, URL: <https://cds.cern.ch/record/1291633>.
- [150] A. La Rosa, *The ATLAS Insertable B-Layer: from construction to operation*, JINST **11** (2016) C12036, arXiv: 1610.01994 [physics.ins-det].
- [151] ATLAS Collaboration, *Technical Design Report for the ATLAS Inner Tracker Pixel Detector*, CERN-LHCC-2017-021, ATLAS-TDR-030, 2017, URL: <https://cds.cern.ch/record/2285585>.
- [152] F. Cavallari, *Performance of calorimeters at the LHC*, Journal of Physics: Conference Series **293** (2011) 012001.
- [153] ATLAS Experiment Web Page, *Detector and Technology*, 2022, URL: <https://atlas.cern/discover/detector> (visited on 02/02/2024).
- [154] ATLAS Collaboration, *ATLAS liquid-argon calorimeter: Technical Design Report*, ATLAS-TDR-2, 1996, URL: <https://cds.cern.ch/record/331061>.
- [155] ATLAS Collaboration, *ATLAS muon spectrometer: Technical design report*, CERN-LHCC-97-22, ATLAS-TDR-10, 1997, URL: <https://cds.cern.ch/record/331068>.
- [156] ATLAS Collaboration, *The ATLAS muon spectrometer*, PoS **HEP2001** (2001) 253.
- [157] M. Livan, *Monitored drift tubes in ATLAS*, ATL-M-PN-129, 1996, URL: <https://cds.cern.ch/record/319197>.
- [158] G. Cattani, *The Resistive Plate Chambers of the ATLAS experiment: performance studies*, Journal of Physics: Conference Series **280** (2011) 012001.
- [159] K. Nagai, *Thin gap chambers in ATLAS*, Nucl. Instrum. Meth. A **384** (1996) 219.
- [160] ATLAS Collaboration, *ATLAS magnet system: Technical Design Report*, 1, ATL-TDR-6, 1997, URL: <https://cds.cern.ch/record/338080>.
- [161] ATLAS Collaboration, *ATLAS high-level trigger, data acquisition and controls: Technical design report*, CERN-LHCC-2003-022, ATLAS-TRD-016, 2003, URL: <https://cds.cern.ch/record/61608>.

- [162] ATLAS Collaboration, *Performance of the ATLAS Trigger System in 2015*, *Eur. Phys. J. C* **77** (2017) 317, arXiv: [1611.09661 \[hep-ex\]](https://arxiv.org/abs/1611.09661).
- [163] P. Mato, *GAUDI-Architecture design document*, LHCb-98-064, CERN-LHCb-98-064, 1998, URL: <https://cds.cern.ch/record/691746/>.
- [164] ATLAS Collaboration, *Design Guidelines*, <https://atlas.cern/design>, Accessed: 2023-10-23, 2023.
- [165] CherryPy Team, *CherryPy - A Minimalist Python Web Framework*, 2017, URL: <https://cherrypy.dev/>.
- [166] T.J. Berners-Lee and R. Cailliau, *WorldWideWeb: proposal for a HyperText Project*, 1990, URL: <https://cds.cern.ch/record/2639699>.
- [167] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, Monographs on Statistics and Applied Probability 57, Chapman & Hall/CRC, 1993, URL: <https://cds.cern.ch/record/526679>.
- [168] F.S. Cafagna, *Latest results for Proton-proton Cross Section Measurements with the TOTEM experiment at LHC*, PoS **ICRC2019** (2021) 207.
- [169] R.P. Feynman, *Very high-energy collisions of hadrons*, *Phys. Rev. Lett.* **23** (1969) 1415.
- [170] J. Butterworth et al., *PDF4LHC recommendations for LHC Run II*, *J. Phys. G* **43** (2016) 023001, arXiv: [1510.03865 \[hep-ph\]](https://arxiv.org/abs/1510.03865).
- [171] S. Alekhin, J. Blumlein and S. Moch, *The ABM parton distributions tuned to LHC data*, *Phys. Rev. D* **89** (2014) 054028, arXiv: [1310.3059 \[hep-ph\]](https://arxiv.org/abs/1310.3059).
- [172] S. Dulat et al., *New parton distribution functions from a global analysis of quantum chromodynamics*, *Phys. Rev. D* **93** (2016) 033006, arXiv: [1506.07443 \[hep-ph\]](https://arxiv.org/abs/1506.07443).
- [173] L.A. Harland-Lang, A.D. Martin, P. Motylinski and R. S. Thorne, *Parton distributions in the LHC era: MMHT 2014 PDFs*, *Eur. Phys. J. C* **75** (2015) 204, arXiv: [1412.3989 \[hep-ph\]](https://arxiv.org/abs/1412.3989).
- [174] NNPDF Collaboration, *Parton distributions for the LHC Run II*, *JHEP* **04** (2015) 040, arXiv: [1410.8849 \[hep-ph\]](https://arxiv.org/abs/1410.8849).
- [175] G. Watt and R.S. Thorne, *Study of Monte Carlo approach to experimental uncertainty propagation with MSTW 2008 PDFs*, *JHEP* **08** (2012) 052, arXiv: [1205.4024 \[hep-ph\]](https://arxiv.org/abs/1205.4024).
- [176] T.J. Hou et al., *New CTEQ global analysis of quantum chromodynamics with high-precision data from the LHC*, *Phys. Rev. D* **103** (2021) 014013, arXiv: [1912.10053 \[hep-ph\]](https://arxiv.org/abs/1912.10053).

- [177] A.D. Martin, W.J. Stirling, R.S. Thorne and G. Watt, *Parton distributions for the LHC*, Eur. Phys. J. C **63** (2009) 189, arXiv: [0901.0002 \[hep-ph\]](#).
- [178] S. Höche, *Introduction to parton-shower event generators*, SLAC-PUB-16160, 2015, arXiv: [1411.4085 \[hep-ph\]](#).
- [179] ATLAS Collaboration, *Luminosity determination in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector at the LHC*, Eur. Phys. J. C **83** (2023) 982, arXiv: [2212.09379 \[hep-ex\]](#).
- [180] ATLAS Collaboration, *Preliminary analysis of the luminosity calibration of the ATLAS 13.6 TeV data recorded in 2022*, ATL-DAPR-PUB-2023-001, 2023, URL: <http://cds.cern.ch/record/2853525>.
- [181] W. Herr and B. Muratori, *Concept of luminosity*, CAS-CERN Accelerator School, 2006, URL: <http://cds.cern.ch/record/941318>.
- [182] ATLAS Collaboration, *Luminosity determination in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector at the LHC*, Eur. Phys. J. C **76** (2016) 653, arXiv: [1608.03953 \[hep-ex\]](#).
- [183] ATLAS Collaboration, *ATLAS data quality operations and performance for 2015–2018 data-taking*, JINST **15** (2020) P04003, arXiv: [1911.04632 \[physics.ins-det\]](#).
- [184] ATLAS Collaboration, *The ATLAS Simulation Infrastructure*, Eur. Phys. J. C **70** (2010) 823, arXiv: [1005.4568 \[physics.ins-det\]](#).
- [185] S. Agostinelli et al., *GEANT4—a simulation toolkit*, Nucl. Instrum. Meth. A **506** (2003) 250.
- [186] S. Frixione, P. Nason and C. Oleari, *Matching NLO QCD computations with parton shower simulations: the POWHEG method*, JHEP **11** (2007) 070, arXiv: [0709.2092 \[hep-ph\]](#).
- [187] T. Gleisberg et al., *Event generation with SHERPA 1.1*, JHEP **02** (2009) 007, arXiv: [0811.4622 \[hep-ph\]](#).
- [188] S. Platzer and S. Gieseke, *Dipole Showers and Automated NLO Matching in Herwig++*, Eur. Phys. J. C **72** (2012) 2187, arXiv: [1109.6256 \[hep-ph\]](#).
- [189] S. Gieseke, P. Stephens and B. Webber, *New formalism for QCD parton showers*, JHEP **12** (2003) 045, arXiv: [hep-ph/0310083](#).
- [190] Y.I. Azimov, Y.L. Dokshitzer, C.A. Khoze and S.I. Troyan, *Similarity of Parton and Hadron Spectra in QCD Jets*, Z. Phys. C **27** (1985) 65.
- [191] B. Andersson, G. Gustafson, G. Ingelman and T. Sjöstrand, *Parton fragmentation and string dynamics*, Phys. Rept. **97** (1983) 31.

- [192] T. Sjöstrand and M. Bengtsson, *The Lund Monte Carlo for Jet Fragmentation and $e^+ e^-$ Physics. Jetset Version 6.3: An Update*, *Comput. Phys. Commun.* **43** (1987) 367.
- [193] I. Borozan and M.H. Seymour, *An Eikonal model for multiparticle production in hadron hadron interactions*, *JHEP* **09** (2002) 015, arXiv: [hep-ph/0207283](#).
- [194] M. Bähr et al., *Herwig++ physics and manual*, *Eur. Phys. J. C* **58** (2008) 639, arXiv: [0803.0883 \[hep-ph\]](#).
- [195] T. Sjöstrand et al., *An introduction to PYTHIA 8.2*, *Comput. Phys. Commun.* **191** (2015) 159, arXiv: [1410.3012 \[hep-ph\]](#).
- [196] ATLAS Collaboration, *The simulation principle and performance of the ATLAS fast calorimeter simulation FastCaloSim*, ATL-PHYS-PUB-2010-013, 2010, URL: <https://cds.cern.ch/record/1300517>.
- [197] T. Yamanaka, *The ATLAS calorimeter simulation FastCaloSim*, *J. Phys. Conf. Ser.* **331** (2011) 032053.
- [198] S. Alioli, P. Nason, C. Oleari and E. Re, *Vector boson plus one jet production in POWHEG*, *JHEP* **01** (2011) 095, arXiv: [1009.5594 \[hep-ph\]](#).
- [199] J. Bellm et al., *Herwig 7.0/Herwig++ 3.0 release note*, *Eur. Phys. J. C* **76** (2016) 196, arXiv: [1512.01178 \[hep-ph\]](#).
- [200] T. Sjöstrand, S. Mrenna and P. Skands, *A brief introduction to PYTHIA 8.1*, *Comput. Phys. Commun.* **178** (2008) 852, arXiv: [0710.3820 \[hep-ph\]](#).
- [201] T. Sjöstrand and M. van Zijl, *A Multiple Interaction Model for the Event Structure in Hadron Collisions*, *Phys. Rev. D* **36** (1987) 2019.
- [202] E. Bothmann et al., *Event Generation with Sherpa 2.2*, *SciPost Phys.* **7** (2019) 034, arXiv: [1905.09127 \[hep-ph\]](#).
- [203] S. Frixione, E. Laenen, P. Motylinski and B.R. Webber, *Angular correlations of lepton pairs from vector boson and top quark decays in Monte Carlo simulations*, *JHEP* **04** (2007) 081, arXiv: [hep-ph/0702198](#).
- [204] P. Artoisenet, R. Frederix, O. Mattelaer R. and R. Rietkerk, *Automatic spin-entangled decays of heavy resonances in Monte Carlo simulations*, *JHEP* **03** (2013) 015, arXiv: [1212.3460 \[hep-ph\]](#).
- [205] D.J. Lange, *The EvtGen particle decay simulation package*, *Nucl. Instrum. Meth. A* **462** (2001) 152.
- [206] J. Pequenao and P. Schaffner, *How ATLAS detects particles: diagram of particle paths in the detector*, CERN-EX-1301009, 2013, URL: <https://cds.cern.ch/record/1505342>.

- [207] ATLAS Collaboration, *Performance of the ATLAS Track Reconstruction Algorithms in Dense Environments in LHC Run 2*, *Eur. Phys. J. C* **77** (2017) 673, arXiv: [1704.07983 \[hep-ex\]](#).
- [208] R. Fruhwirth, *Application of Kalman filtering to track and vertex fitting*, *Nucl. Instrum. Meth. A* **262** (1987) 444.
- [209] ATLAS Collaboration, *Electron and photon performance measurements with the ATLAS detector using the 2015–2017 LHC proton-proton collision data*, *JINST* **14** (2019) P12006, arXiv: [1908.00005 \[hep-ex\]](#).
- [210] ATLAS Collaboration, *Electron reconstruction and identification in the ATLAS experiment using the 2015 and 2016 LHC proton-proton collision data at $\sqrt{s} = 13$ TeV*, *Eur. Phys. J. C* **79** (2019) 639, arXiv: [1902.04655 \[physics.ins-det\]](#).
- [211] ATLAS Collaboration, *Electron reconstruction and identification in the ATLAS experiment using the 2015 and 2016 LHC proton–proton collision data at $\sqrt{s} = 13$ TeV*, *Eur. Phys. J. C* **79** (2019) 639, arXiv: [1902.04655 \[hep-ex\]](#).
- [212] ATLAS Collaboration, *Electron reconstruction and identification efficiency measurements with the ATLAS detector using the 2011 LHC proton-proton collision data*, *Eur. Phys. J. C* **74** (2014) 2941, arXiv: [1404.2240 \[hep-ex\]](#).
- [213] ATLAS Collaboration, *Electron performance measurements with the ATLAS detector using the 2010 LHC proton-proton collision data*, *Eur. Phys. J. C* **72** (2012) 1909, arXiv: [1110.3174 \[hep-ex\]](#).
- [214] A. Straessner and M. Schott, *A new tool for measuring detector performance in ATLAS*, *J. Phys. Conf. Ser.* **219** (2010) 032023, URL: <https://cds.cern.ch/record/1354502>.
- [215] ATLAS Collaboration, *Muon reconstruction performance of the ATLAS detector in proton–proton collision data at $\sqrt{s} = 13$ TeV*, *Eur. Phys. J. C* **76** (2016) 292, arXiv: [1603.05598 \[hep-ex\]](#).
- [216] ATLAS Collaboration, *Muon reconstruction and identification efficiency in ATLAS using the full Run 2 pp collision data set at $\sqrt{s} = 13$ TeV*, *Eur. Phys. J. C* **81** (2021) 578, arXiv: [2012.00578 \[hep-ex\]](#).
- [217] J. Illingworth and J. Kittler, *A survey of the hough transform*, *Computer Vision, Graphics, and Image Processing* **44** (1988) 87.
- [218] Belle Collaboration, *Measurement of the τ -lepton lifetime at Belle*, *Phys. Rev. Lett.* **112** (2014) 031801, arXiv: [1310.8503 \[hep-ex\]](#).
- [219] S. Catani, Y. L. Dokshitzer, M.H. Seymour and B.R. Webber, *Longitudinally invariant K_t clustering algorithms for hadron hadron collisions*, *Nucl. Phys. B* **406** (1993) 187.

- [220] CMS Collaboration, *A Cambridge-Aachen (C-A) based Jet Algorithm for boosted top-jet tagging*, CMS-PAS-JME-09-001, 2009, URL: <https://cds.cern.ch/record/1194489>.
- [221] M. Cacciari, G.P. Salam and G. Soyez, *The anti- k_t jet clustering algorithm*, JHEP **04** (2008) 063, arXiv: [0802.1189 \[hep-ph\]](https://arxiv.org/abs/0802.1189).
- [222] ATLAS Collaboration, *Reconstruction, Identification, and Calibration of hadronically decaying tau leptons with the ATLAS detector for the LHC Run 3 and reprocessed Run 2 data*, ATL-PHYS-PUB-2022-044, 2022, URL: <https://cds.cern.ch/record/2827111>.
- [223] ATLAS Collaboration, *Identification and energy calibration of hadronically decaying tau leptons with the ATLAS experiment in pp collisions at $\sqrt{s}=8$ TeV*, Eur. Phys. J. C **75** (2015) 303, arXiv: [1412.7086 \[hep-ex\]](https://arxiv.org/abs/1412.7086).
- [224] ATLAS Collaboration, *Performance of the Reconstruction and Identification of Hadronic Tau Decays in ATLAS with 2011 Data*, ATLAS-CONF-2012-142, 2012, URL: <https://cds.cern.ch/record/1485531>.
- [225] ATLAS Collaboration, *Identification of hadronic tau lepton decays using neural networks in the ATLAS experiment*, ATL-PHYS-PUB-2019-033, 2019, URL: <https://cds.cern.ch/record/2688062>.
- [226] UA1 Collaboration, *Hadronic Jet Production at the CERN Proton-anti-Proton Collider*, Phys. Lett. B **132** (1983) 214.
- [227] ATLAS Collaboration, *Boosted hadronic vector boson and top quark tagging with ATLAS using Run 2 data*, ATL-PHYS-PUB-2020-017, 2020, URL: <https://cds.cern.ch/record/2724149>.
- [228] ATLAS Collaboration, *Performance of top-quark and W-boson tagging with ATLAS in Run 2 of the LHC*, Eur. Phys. J. C **79** (2019) 375, arXiv: [1808.07858 \[hep-ex\]](https://arxiv.org/abs/1808.07858).
- [229] ATLAS Collaboration, *ATLAS b-jet identification performance and efficiency measurement with $t\bar{t}$ events in pp collisions at $\sqrt{s} = 13$ TeV*, Eur. Phys. J. C **79** (2019) 970, arXiv: [1907.05120 \[hep-ex\]](https://arxiv.org/abs/1907.05120).
- [230] ATLAS Collaboration, *Optimisation and performance studies of the ATLAS b-tagging algorithms for the 2017-18 LHC run*, ATL-PHYS-PUB-2017-013, 2017, URL: <https://cds.cern.ch/record/2273281>.
- [231] ATLAS Collaboration, *Identification of Jets Containing b-Hadrons with Recurrent Neural Networks at the ATLAS Experiment*, ATL-PHYS-PUB-2017-003, 2017, URL: <https://cds.cern.ch/record/2255226>.
- [232] ATLAS Collaboration, *Secondary vertex finding for jet flavour identification with the ATLAS detector*, ATL-PHYS-PUB-2017-011, 2017, URL: <https://cds.cern.ch/record/2270366>.

- [233] ATLAS Collaboration, *Topological b-hadron decay reconstruction and identification of b-jets with the JetFitter package in the ATLAS experiment at the LHC*, ATL-PHYS-PUB-2018-025, 2018, URL: <https://cds.cern.ch/record/2645405>.
- [234] ATLAS Collaboration, *Optimisation and performance studies of the ATLAS b-tagging algorithms for the 2017-18 LHC run*, ATL-PHYS-PUB-2017-013, 2017, URL: <https://cds.cern.ch/record/2273281>.
- [235] ATLAS Collaboration, *ATLAS flavour-tagging algorithms for the LHC Run 2 pp collision dataset*, Eur. Phys. J. C **83** (2023) 681, arXiv: [2211.16345 \[physics.data-an\]](https://arxiv.org/abs/2211.16345).
- [236] ATLAS Collaboration, *ATLAS b-jet identification performance and efficiency measurement with $t\bar{t}$ events in pp collisions at $\sqrt{s} = 13$ TeV*, Eur. Phys. J. C **79** (2019) 970, arXiv: [1907.05120 \[hep-ex\]](https://arxiv.org/abs/1907.05120).
- [237] ATLAS Collaboration, *Measurement of b-tagging efficiency of c-jets in $t\bar{t}$ events using a likelihood approach with the ATLAS detector*, ATLAS-CONF-2018-001, 2018, URL: <https://cds.cern.ch/record/2306649>.
- [238] ATLAS Collaboration, *Calibration of light-flavour b-jet mistagging rates using ATLAS proton–proton collision data at $\sqrt{s} = 13$ TeV*, ATLAS-CONF-2018-006, 2018, URL: <https://cds.cern.ch/record/2314418>.
- [239] ATLAS Collaboration, *Performance of missing transverse momentum reconstruction with the ATLAS detector using proton-proton collisions at $\sqrt{s} = 13$ TeV*, Eur. Phys. J. C **78** (2018) 903, arXiv: [1802.08168 \[hep-ex\]](https://arxiv.org/abs/1802.08168).
- [240] ATLAS Collaboration, *E_T^{miss} performance in the ATLAS detector using 2015-2016 LHC pp collisions*, ATLAS-CONF-2018-023, 2018, URL: <https://cds.cern.ch/record/2625233>.
- [241] T. Holm, *Towards a measurement of the tHq process in channels with hadronic tau lepton decays*, PhD: Rheinische Friedrich-Wilhelms-Universität Bonn, 2023, URL: <https://www.pi.uni-bonn.de/brock/en/results/data/t00000060.pdf>.
- [242] K. Tariq, *Search for Higgs Boson production associated with single top quark in same sign di-lepton final state at ATLAS*, PhD: Shandong University, 2022, URL: <https://cds.cern.ch/record/2883582>.
- [243] J. Guerrero Rojas, *Search for associated production of a Higgs boson and a single top quark in the multi-lepton final state with ATLAS*, PhD: Universitat de València, 2023, URL: <https://cds.cern.ch/record/2875564>.
- [244] LUCID Collaboration, *The new LUCID-2 detector for luminosity measurement and monitoring in ATLAS*, JINST **13** (2018) P07017.

- [245] ATLAS Collaboration, *Luminosity determination in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector at the LHC*, ATLAS-CONF-2019-021, 2019, URL: <http://cds.cern.ch/record/2677054>.
- [246] ATLAS Collaboration, *The Pythia 8 A3 tune description of ATLAS minimum bias and inelastic measurements incorporating the Donnachie-Landshoff diffractive model*, ATL-PHYS-PUB-2016-017, 2016, URL: <https://cds.cern.ch/record/2206965>.
- [247] R.D. Ball et al., *Parton distributions with LHC data*, Nucl. Phys. B **867** (2013) 244, arXiv: [1207.1303 \[hep-ph\]](https://arxiv.org/abs/1207.1303).
- [248] Richard D. Ball et al., *Parton distributions for the LHC run II*, JHEP **04** (2015) 040, arXiv: [1410.8849 \[hep-ph\]](https://arxiv.org/abs/1410.8849).
- [249] ATLAS Collaboration, *ATLAS Pythia 8 tunes to 7 TeV data*, ATL-PHYS-PUB-2014-021, 2014, URL: <https://cds.cern.ch/record/1966419>.
- [250] ATLAS Collaboration, *Measurement of the Inelastic Proton-Proton Cross Section at $\sqrt{s} = 13$ TeV with the ATLAS Detector at the LHC*, Phys. Rev. Lett. **117** (2016) 182002, arXiv: [1606.02625 \[hep-ex\]](https://arxiv.org/abs/1606.02625).
- [251] J. Alwall et al., *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, JHEP **07** (2014) 079, arXiv: [1405.0301 \[hep-ph\]](https://arxiv.org/abs/1405.0301).
- [252] S. Frixione, P. Nason and G. Ridolfi, *A positive-weight next-to-leading-order Monte Carlo for heavy flavour hadroproduction*, JHEP **09** (2007) 126, arXiv: [0707.3088 \[hep-ph\]](https://arxiv.org/abs/0707.3088).
- [253] P. Nason, *A new method for combining NLO QCD with shower Monte Carlo algorithms*, JHEP **11** (2004) 040, arXiv: [hep-ph/0409146](https://arxiv.org/abs/hep-ph/0409146).
- [254] S. Alioli, P. Nason, C. Oleari and E. Re, *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*, JHEP **06** (2010) 043, arXiv: [1002.2581 \[hep-ph\]](https://arxiv.org/abs/1002.2581).
- [255] ATLAS Collaboration, *Studies on top-quark Monte Carlo modelling for Top2016*, ATL-PHYS-PUB-2016-020, 2016, URL: <https://cds.cern.ch/record/2216168>.
- [256] M. Beneke, P. Falgari, S. Klein and C. Schwinn, *Hadronic top-quark pair production with NNLL threshold resummation*, Nucl. Phys. B **855** (2012) 695, arXiv: [1109.1536 \[hep-ph\]](https://arxiv.org/abs/1109.1536).
- [257] M. Cacciari, M. Czakon, M. Mangano, A. Mitov and P. Nason, *Top-pair production at hadron colliders with next-to-next-to-leading logarithmic soft-gluon resummation*, Phys. Lett. B **710** (2012) 612, arXiv: [1111.5869 \[hep-ph\]](https://arxiv.org/abs/1111.5869).

- [258] P. Bärnreuther, M. Czakon and A. Mitov, *Percent-Level-Precision Physics at the Tevatron: Next-to-Next-to-Leading Order QCD Corrections to $q\bar{q} \rightarrow t\bar{t} + X$* , Phys. Rev. Lett. **109** (2012) 132001, arXiv: [1204.5201 \[hep-ph\]](#).
- [259] M. Czakon and A. Mitov, *NNLO corrections to top-pair production at hadron colliders: the all-fermionic scattering channels*, JHEP **12** (2012) 054, arXiv: [1207.0236 \[hep-ph\]](#).
- [260] M. Czakon and A. Mitov, *NNLO corrections to top pair production at hadron colliders: the quark-gluon reaction*, JHEP **01** (2013) 080, arXiv: [1210.6832 \[hep-ph\]](#).
- [261] M. Czakon and A. Mitov, *Top++: A program for the calculation of the top-pair cross-section at hadron colliders*, Comput. Phys. Commun. **185** (2014) 2930, arXiv: [1112.5675 \[hep-ph\]](#).
- [262] E. Bothmann et al., *Event generation with Sherpa 2.2*, SciPost Phys. **7** (2019) 034, arXiv: [1905.09127 \[hep-ph\]](#).
- [263] J.C. Winter, F. Krauss and G. Soff, *A modified cluster-hadronization model*, Eur. Phys. J. C **36** (2004) 381, arXiv: [hep-ph/0311085](#).
- [264] S. Höche, F. Krauss, M. Schönherr and F. Siegert, *A critical appraisal of NLO+PS matching methods*, JHEP **09** (2012) 049, arXiv: [1111.1220 \[hep-ph\]](#).
- [265] C. Anastasiou, Lance J. Dixon, K. Melnikov and F. Petriello, *High precision QCD at hadron colliders: Electroweak gauge boson rapidity distributions at next-to-next-to leading order*, Phys. Rev. D **69** (2004) 094008, arXiv: [hep-ph/0312266](#).
- [266] R. Frederix, E. Re and P. Torrielli, *Single-top t-channel hadroproduction in the four-flavour scheme with POWHEG and aMC@NLO*, JHEP **09** (2012) 130, arXiv: [1207.5391 \[hep-ph\]](#).
- [267] ATLAS Collaboration, *Performance of the ATLAS trigger system in 2015*, Eur. Phys. J. C **77** (2017) 317, arXiv: [1611.09661 \[hep-ex\]](#).
- [268] ATLAS Collaboration, *Tools for estimating fake/non-prompt lepton backgrounds with the ATLAS detector at the LHC*, JINST **18** (2023) T11004, arXiv: [2211.16178 \[hep-ex\]](#).
- [269] ATLAS Collaboration, *Reconstruction of hadronic decay products of tau leptons with the ATLAS experiment*, Eur. Phys. J. C **76** (2016) 295, arXiv: [1512.05955 \[hep-ex\]](#).
- [270] ATLAS Collaboration, *Jet reconstruction and performance using particle flow with the ATLAS Detector*, Eur. Phys. J. C **77** (2017) 466, arXiv: [1703.10485 \[hep-ex\]](#).

- [271] M. Cacciari, G. P. Salam and G. Soyez, *FastJet user manual*, *Eur. Phys. J. C* **72** (2012) 1896, arXiv: [1111.6097 \[hep-ph\]](https://arxiv.org/abs/1111.6097).
- [272] ATLAS Collaboration, *Tagging and suppression of pileup jets with the ATLAS detector*, ATLAS-CONF-2014-018, 2014, URL: <https://cds.cern.ch/record/1700870>.
- [273] ATLAS Collaboration, *Performance of pile-up mitigation techniques for jets in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector*, *Eur. Phys. J. C* **76** (2016) 581, arXiv: [1510.03823 \[hep-ex\]](https://arxiv.org/abs/1510.03823).
- [274] ATLAS Collaboration, *Jet energy scale and resolution measured in proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, *Eur. Phys. J. C* **81** (2020) 689, arXiv: [2007.02645 \[hep-ex\]](https://arxiv.org/abs/2007.02645).
- [275] ATLAS Collaboration, *Commissioning of the ATLAS b-tagging algorithms using $t\bar{t}$ events in early Run 2 data*, ATL-PHYS-PUB-2015-039, 2015, URL: <https://cds.cern.ch/record/2047871>.
- [276] ATLAS Collaboration, *Identification of Jets Containing b-Hadrons with Recurrent Neural Networks at the ATLAS Experiment*, ATL-PHYS-PUB-2017-003, 2017, URL: <https://cds.cern.ch/record/2255226>.
- [277] ATLAS Collaboration, *Performance of missing transverse momentum reconstruction with the ATLAS detector using proton–proton collisions at $\sqrt{s} = 13$ TeV*, *Eur. Phys. J. C* **78** (2018) 903, arXiv: [1802.08168 \[hep-ex\]](https://arxiv.org/abs/1802.08168).
- [278] ATLAS Collaboration, *E_T^{miss} performance in the ATLAS detector using 2015–2016 LHC pp collisions*, ATLAS-CONF-2018-023, 2018, URL: <https://cds.cern.ch/record/2625233>.
- [279] ATLAS Collaboration, *Performance of missing transverse momentum reconstruction in proton–proton collisions at $\sqrt{s} = 7$ TeV with ATLAS*, *Eur. Phys. J. C* **72** (2012) 1844, arXiv: [1108.5602 \[hep-ex\]](https://arxiv.org/abs/1108.5602).
- [280] ATLAS Collaboration, *Performance of algorithms that reconstruct missing transverse momentum in $\sqrt{s} = 8$ TeV proton–proton collisions in the ATLAS detector*, *Eur. Phys. J. C* **77** (2017) 241, arXiv: [1609.09324 \[hep-ex\]](https://arxiv.org/abs/1609.09324).
- [281] ATLAS Collaboration, *Performance of missing transverse momentum reconstruction with the ATLAS detector in the first proton–proton collisions at $\sqrt{s} = 13$ TeV*, ATL-PHYS-PUB-2015-027, 2015, URL: <https://cds.cern.ch/record/2037904>.
- [282] C.P. Walther, *Higgs Boson Mass Reconstruction Using Neural Network Approaches for ttH and tH Analyses*, Master thesis: Technische Universitaet Dortmund, 2020, URL: <https://cds.cern.ch/record/2743809>.

- [283] R. Brun and F. Rademakers, *ROOT: An object oriented data analysis framework*, Nucl. Instrum. Meth. A **389** (1997) 81.
- [284] K. Albertsson et al., *TMVA-toolkit for multivariate data analysis*, CERN-OPEN-2007-007, 2007, arXiv:physics/0703039 [physics.data-an].
- [285] L. Breiman, J. Friedman, R.A. Olshen and C.J. Stone, *Classification and Regression Trees*, 1st, Routledge, 1984, ISBN: 0412048418, URL: <https://doi.org/10.1201/9781315139470>.
- [286] C. Rubbia, *Physics results of the UA1 Collaboration at the CERN proton-antiproton collider; rev. version*, CERN-EP-84-135, 1984, URL: <https://cds.cern.ch/record/155129>.
- [287] R.K. Ellis, I. Hinchliffe, M. Soldate and J.J. van der Bij, *Higgs Decay to $\tau^- \tau^+$: A Possible Signature of Intermediate Mass Higgs Bosons at high energy hadron colliders*, Nucl. Phys. B **297** (1988) 221.
- [288] P. Jackson and C. Rogan, *Recursive Jigsaw Reconstruction: HEP event analysis in the presence of kinematic and combinatoric ambiguities*, Phys. Rev. D **96** (2017) 112007, arXiv: 1705.10733 [hep-ph].
- [289] J. Erdmann et al., *A likelihood-based reconstruction algorithm for top-quark pairs and the KLFitter framework*, Nucl. Instrum. Meth. A **748** (2014) 18, arXiv: 1312.5595 [hep-ex].
- [290] A. Elagin, P. Murat, A. Pranko and A. Safonov, *A new mass reconstruction technique for resonances decaying to di-tau*, Nucl. Instrum. Meth. A **654** (2011) 481, ISSN: 0168-9002.
- [291] ATLAS Collaboration, *Search for the Standard Model Higgs boson in the $H \rightarrow \tau^+ \tau^-$ decay mode in $\sqrt{s} = 7$ TeV pp collisions with ATLAS*, JHEP **09** (2012) 070, arXiv: 1206.5971 [hep-ex].
- [292] G. Cowan, H. Cranmer, E. Gross and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, Eur. Phys. J. C **71** (2011) 1554, arXiv: 1007.1727 [physics.data-an].
- [293] T. Chen and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016), arXiv: 1603.02754 [cs.LG].
- [294] M. Mitchell, *An Introduction to Genetic Algorithms*, 2nd, Cambridge: The MIT Press, 1998, ISBN: 9780262631853, URL: <https://mitpress.mit.edu/9780262631853/an-introduction-to-genetic-algorithms/>.
- [295] ATLAS Collaboration, *Studies on top-quark Monte Carlo modelling with Sherpa and MG5_aMC@NLO*, ATL-PHYS-PUB-2017-007, 2017, URL: <https://cds.cern.ch/record/2261938>.

- [296] A.D. Martin, W.J. Stirling, R.S. Thorne and G. Watt, *Uncertainties on α_S in global PDF analyses and implications for predicted hadronic cross sections*, *Eur. Phys. J. C* **64** (2009) 653, arXiv: [0905.3531 \[hep-ph\]](#).
- [297] H.L. Lai et al., *New parton distributions for collider physics*, *Phys. Rev. D* **82** (2010) 074024, arXiv: [1007.2241 \[hep-ph\]](#).
- [298] J. Gao et al., *CT10 next-to-next-to-leading order global analysis of QCD*, *Phys. Rev. D* **89** (2014) 033009, arXiv: [1302.6246 \[hep-ph\]](#).
- [299] F.A. Berends, H. Kuijf, B. Tausk and W.T. Giele, *On the production of a W and jets at hadron colliders*, *Nuclear Physics B* **357** (1991) 32.
- [300] Z. Marshall, *Simulation of Pile-up in the ATLAS Experiment*, *J. Phys. Conf. Ser.* **513** (2014) 022024.
- [301] W. Buttinger, *Using Event Weights to account for differences in Instantaneous Luminosity and Trigger Prescale in Monte Carlo and Data*, ATL-COM-SOFT-2015-119, 2015, URL: <https://cds.cern.ch/record/2014726>.
- [302] ATLAS Collaboration, *Jet energy scale and resolution measured in proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, *Eur. Phys. J. C* **81** (2021) 689, arXiv: [2007.02645 \[hep-ex\]](#).
- [303] ATLAS Collaboration, *Jet energy scale measurements and their systematic uncertainties in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, *Phys. Rev. D* **96** (2017) 072002, arXiv: [1703.09665 \[hep-ex\]](#).
- [304] ATLAS Collaboration, *Identification and rejection of pile-up jets at high pseudorapidity with the ATLAS detector*, *Eur. Phys. J. C* **77** (2017) 580, arXiv: [1705.02211 \[hep-ex\]](#), Erratum: *Eur. Phys. J. C* **77** (2017) 712.
- [305] ATLAS Collaboration, *Electron and photon performance measurements with the ATLAS detector using the 2015–2017 LHC proton–proton collision data*, *JINST* **14** (2019) P12006, arXiv: [1908.00005 \[hep-ex\]](#).
- [306] ATLAS Collaboration, *Muon reconstruction and identification efficiency in ATLAS using the full Run 2 pp collision data set at $\sqrt{s} = 13$ TeV*, *Eur. Phys. J. C* **81** (2020) 578, arXiv: [2012.00578 \[hep-ex\]](#).
- [307] ATLAS Collaboration, *Muon reconstruction performance of the ATLAS detector in proton–proton collision data at $\sqrt{s} = 13$ TeV*, *Eur. Phys. J. C* **76** (2016) 292, arXiv: [1603.05598 \[hep-ex\]](#).
- [308] A.L. Read, *Presentation of search results: the CL_S technique*, *J. Phys. G* **28** (2002) 2693.

- [309] ATLAS Collaboration, *Measurement of the total and differential cross-sections of $t\bar{t}W$ production in pp collisions at 13 TeV with the ATLAS detector*, ATLAS-CONF-2023-019, 2023, URL: <https://cds.cern.ch/record/2855337>.
- [310] CMS Collaboration, *Measurement of the cross section of top quark-antiquark pair production in association with a W boson in proton-proton collisions at $\sqrt{s} = 13$ TeV*, JHEP **07** (2023) 219, arXiv: [2208.06485 \[hep-ex\]](https://arxiv.org/abs/2208.06485).
- [311] Y. Freund and R. E. Schapire, *A desicion-theoretic generalization of on-line learning and an application to boosting*, J. Comput. Syst. Sci. **55** (1997) 119.
- [312] J. Zimmermann, *Statistical learning methods: Basics, control and performance*, Nucl. Instrum. Meth. A **559** (2006) 106.
- [313] R.A. Fisher, *The Use of Multiple Measurements in Taxonomic Problems*, Annals of Eugenics **7** (1936) 179.
- [314] C.J.C. Burges, R. Ragno and Q. Le, *Learning to Rank with Nonsmooth Cost Functions*, Inf. Retr. **19** (2006).
- [315] Q. Wu, C.J.C. Burges, K.M. Svore et al., *Adapting boosting for information retrieval measures*, Inf. Retr. **13** (2010) 254.
- [316] G. Ke et al., *Lightgbm: A highly efficient gradient boosting decision tree*, Advances in neural information processing systems **30** (2017) 3146, URL: <https://dl.acm.org/doi/10.5555/3294996.3295074>.
- [317] K. Danziger, S. Höche and F. Siegert, *Reducing negative weights in Monte Carlo event generation with Sherpa*, (2021), arXiv: [2110.15211 \[hep-ph\]](https://arxiv.org/abs/2110.15211).
- [318] J.R. Andersen and A. Maier, *Unbiased elimination of negative weights in Monte Carlo samples*, Eur. Phys. J. C **82** (2022) 433, arXiv: [2109.07851 \[hep-ph\]](https://arxiv.org/abs/2109.07851).

Recerca de la producció associada de bosó de Higgs i un quark top amb dos leptons i tau hadrònic en l'estat final

Les meves contradiccions són les meves esperances.

—JOAN FUSTER,
CONSELLS, PROVERBIS I INSOLÈNCIES (1968)

El Model Estàndard (SM) de la física de partícules és una teoria notablement exitosa, però també manifesta limitacions significatives. Aquest model unifica totes les partícules elementals que constitueixen l'univers coneigut en una teoria única. Dins d'aquest marc, el quark top i el bosó de Higgs desperten un interès especial, ja que poden contribuir a respondre algunes de les qüestions encara pendents. Aquesta tesi es centra en l'estudi d'ambdues partícules i la seua interacció. El marc teòric per a l'estudi de la física d'aquestes partícules es presenta a la Secció 1.

Per dur a terme aquest estudi, s'han utilitzat dades de col·lisions protó-protó (pp) amb una lluminositat integrada de 140fb^{-1} , a una energia de centre de masses de 13 TeV, recopilades pel detector ATLAS durant el Run 2 del Gran Col·lisionador d'Hadrons (LHC) l'Organització Europea per a la Recerca Nuclear (CERN). L'ATLAS és el detector més gran de l'LHC, el més potent accelerador de partícules del món. El marc experimental en el qual s'emmarca aquest treball es descriu a la Secció 2. La recopilació de les dades, la generació de simulacions de Monte Carlo, i la reconstrucció i identificació dels objectes físics són descrits a la Secció 3.

El bosó de Higgs fou descobert pels experiments ATLAS [77] i CMS [78] en 2012 i, en conseqüència, va obrir un nou camp d'exploració en la física de partícules. Per comprendre millor l'SM, és d'un gran interès determinar l'acoblament de Yukawa del bosó de Higgs amb el quark top (y_t). Aquest quark és la partícula fonamental més massiva i, per tant, presenta l'acoblament més fort amb el bosó de Higgs.

La mesura directa de y_t només és possible a l'LHC a través de dues produccions associades del bosó de Higgs: amb un parell de quark–antiquark de top ($t\bar{t}H$) i amb un únic quark top juntament amb un partó addicional (tHq). Mentre que el $t\bar{t}H$ només permet determinar la magnitud de y_t , l'única manera de mesurar simultàniament el seu signe i la seua magnitud és mitjançant la producció tH [97]. L'observació d'un excés d'esdeveniments de senyal en comparació amb la prediccio de l'SM podria ser una evidència de nova física en termes de violació de CP a l'acoblament y_t .

En aquest treball, es presenta una cerca de la producció tHq a l'estat final definit per dos leptons lleugers carregats (electrons o muons) i un leptó τ que es desintegra de manera hadrònica. Aquesta configuració es coneix com a canal $2\ell + 1\tau_{\text{had}}$. Aquesta búqueda presenta un repte a causa de l'extremadament menuda secció eficaç del procés tHq en general, i sobretot pel canal final $2\ell + 1\tau_{\text{had}}$, que representa només el 3.5% de la producció total de tHq .

Per distingir els esdeveniments de senyal tHq dels de fons s'han fet servir tècniques d'aprenentatge automàtic. En concret, s'han utilitzat arbres de decisió potenciats (BDT) per definir tant les regions enriquides de senyal com les regions de control que limiten els processos de fons més importants. Els fons rellevants inclouen la producció de parells de quark–antiquark de top en solitari i amb un bosó addicional ($t\bar{t}$, $t\bar{t}H$, $t\bar{t}W$ i $t\bar{t}Z$) i el bosó Z juntament amb jets¹.

A més, per ajudar a identificar els esdeveniments de senyal dins les dades, la reconstrucció de l'esdeveniment juga un paper crucial. En situacions en què els leptons lleugers tenen la mateixa càrrega elèctrica, no és possible determinar a priori quin leptó prové del bosó de Higgs i quin del quark top. Atés aquest fet s'ha desenvolupat una eina basada en un BDT per assignar exitosament l'origen.

S'aconsegueix una supressió significativa dels esdeveniments de fons, imposant requisits estrictes d'identificació i aïllament per als electrons i muons. Al mateix temps, es demana als taus hadrònics que superen un discriminador basat en xarxes neuronals recurrents per reduir les identificacions errònies provinents dels jets.

¹Un jet es refereix a un feix concentrat de partícules produïdes quan un quark o gluó es desintegra i s'hadrònitzca, donant lloc un grup de partícules detectables que emergeixen en una direcció similar formant un con.

Totes les ferramentes mencionades per a la recerca de processos tH i el procediment per a l'anàlisi estan descrits a la Secció 4, on també es presenten i discuteixen els resultats.

1 Marc teòric

1.1 El Model Estàndard

L'SM és un marc teòric que descriu els constituents bàsics de la matèria i les seues interaccions. És el model més àmpliament acceptat i confirmat experimentalment en la física de partícules. La Figura 1 recull totes les partícules fonamentals descrites per l'SM.

L'SM inclou dos tipus de partícules elementals: els fermions i els bosons. Els fermions són partícules subatòmiques que segueixen les regles de l'estadística de Fermi-Dirac. Aquest tipus de partícula es caracteritza per tenir un espíñ semienter i per seguir el principi d'exclusió de Pauli, el qual estableix que dos fermions no poden ocupar el mateix estat quàntic simultaneament. Els fermions es divideixen en quarks i leptons. Ambdós tipus de fermions són els constituents bàsics de la matèria, però són diferents entre si.

D'una banda, els quarks són partícules que tenen càrrega elèctrica fraccionària. A més són la unitat fonamental dels protons i els neutrons. Aquestes partícules es combinen en grups per formar hadrons (mesons i barions). Els barions inclouen els protons i els neutrons, que són les partícules subatòmiques més abundants en la matèria. Els mesons tenen un nombre parell de quarks, la qual cosa fa que tinguen espíñ enter i siguen bosons. Els quarks es divideixen en sis “sabors” diferents: amunt, avall, encant, estrany, superior i inferior. La forma més habitual de referir-se a ells és pel seu nom en anglès: *up*, *down*, *charm*, *strange*, *top* i *bottom*.

D'una altra banda, els leptons tenen càrrega elèctrica sencera i es classifiquen en leptons carregats (electró, muó i tauó, amb), que tenen càrrega -1 i una massa relativament petita comparada amb altres partícules subatòmiques com els quarks, i els neutrins (neutrí electrònic, neutrí muònic i neutrí tauònic), amb càrrega neutra i massa quasi nul · la.

Els altres elements que componen l'SM són els bosons, partícules amb espíñ enter que medien les interaccions fonamentals de la física. Els bosons de calibre (espíñ 1) són els responsables de descriure tres de les quatre forces fonamentals de la naturalesa²:

²La gravetat queda fora de l'SM.

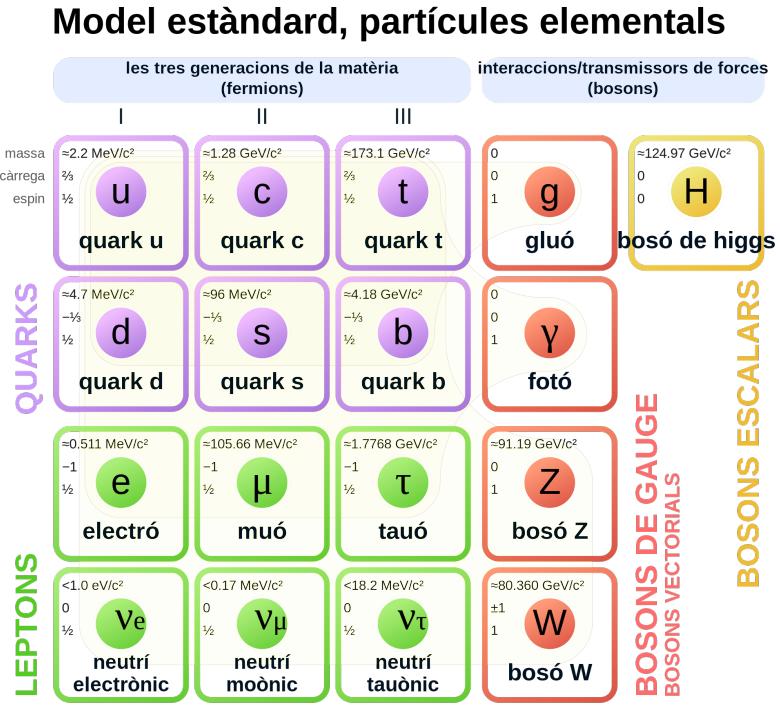


Figura 1: Model estàndard de les partícules elementals, amb les tres generacions de partícules de matèria, els bosons de gauge i el bosó de Higgs. Les superfícies en marró clar indiquen quins bosons (roig) s'acoblen a quins fermions (morat i verd)..

- Interacció electromagnètica: Mediada pel fotó (γ), és la teoria que estudia els fenòmens elèctrics i magnètics. Totes les partícules carregades interactuen entre si a través d'aquesta força. Les principals característiques de la interacció electromagnètica són el seu abast infinit i l'absència de massa dels seus portadors. En aquest sentit, és responsable de l'estabilitat dels àtoms, ja que manté units els electrons en òrbita al voltant del nucli, i de la transmissió de la llum i altres formes de radiació electromagnètica. La teoria que descriu aquesta interacció es denomina electrodinàmica quàntica.
- Interacció nuclear feble: Mediada per dos bosons W (W^+ i W^-) i el bosó Z . Aquesta és responsable de la radioactivitat β , en la qual un neutró es descompon en un protó, un electron i un antineutrí. També és la força a través de la qual té lloc la desintegració del quark top en un quark b i un bosó W . A més, la interacció nuclear feble és crucial en el procés de fusió en les estrelles, on es combinen protons per formar elements més pesants. Les forces nuclear feble i electromagnètica es descriuen simultàniament per la teoria electrofeble (EW).

- Interacció nuclear forta: Mediada pel gluó, és responsable de mantenir units els protons i neutrons al nucli atòmic. És la interacció més forta de la naturalesa, però el seu abast d'acció està limitat a distàncies subatòmiques. A causa del confinament per color de la teoria nuclear forta, ni els gluons ni els quarks apareixen aïllats (excepte a altes energies). La teoria que descriu aquesta interacció es diu cromodinàmica quàntica (QCD). Aquesta teoria, igual que l'electrodinàmica quàntica i la teoria electrofeble, està basada en el formalisme de la teoria quàntica de camps.

1.2 La física del quark top

El quark top (t), o simplement top, és el quark de tipus up de la tercera generació de fermions. La seua característica més distintiva és la seua enorme massa, sent aquesta la més gran entre totes les partícules fonamentals. La seua existència va ser postulada al 1973 per Kobayashi i Maskawa [43] i es va observar per primera vegada al Tevatron a l'any 1995 [47, 48]. Sovint, el quark top és considerat com una finestra per a la nova física perquè proporciona un laboratori únic per provar la comprensió de l'SM.

La seua fenomenologia està determinada per la seua gran massa, deguda a la qual, la vida mitjana del quark top és la més curta entre totes les partícules de l'SM. ($\tau_t = 5 \times 10^{-25}$ s [35]). Això representa una oportunitat única per estudiar quarks en estat lliure, fet que és excepcional a causa del confinament de color. De fet, el quark top és l'únic quark que es pot investigar sense estar lligat. La seua vida mitjana també és menor que l'escala de temps de decorrelació d'espín ($m_t/\Lambda_{QCD}^2 \sim 10^{-21}$ s [42]), implicant que els estats del quark top conserven el seu estat d'espín des de la producció fins la desintegració. D'aquest mode, les propietats del quark top, com la informació d'espín, es poden obtenir a través dels seus productes de desintegració i, per tant, mesurar-se.

A més, la gran massa del quark top propicia que consegüentment, aquesta partícula siga l'única amb un acoblament de Yukawa al bosó de Higgs (y_t) de l'ordre de l'unitat. L'objectiu principal d'aquesta tesi és, precisament, l'estudi de la interacció entre el quark top i el bosó de Higgs per mesurar la seua secció eficaç de producció i ajudar a determinar si el y_t és el que prediu l'SM o si hi ha alguna fase que viola la simetria de CP i afecta així al signe de l'acoblament Yukawa entre el Higgs i el top.

La masa del quark top (m_t) és un paràmetre lliure de l'SM; llavors, la teoria no prediu el seu valor i, per tant, aquest ha de ser determinat experimentalment. Els estudis més recents per a la mesura de la massa del quark top aboquen $m_t = 172.76 \pm 0.30$ GeV [10].

Pel que respecta a la seu producció, l'LHC és conegut com una fàbrica de quarks top a causa de la seu capacitat per produir aquestes partícules. En aquest col·lisionador, el quark top es produeix principalment a través de dos mecanismes: mitjançant QCD, en forma de parelles de quark top i anti-top ($t\bar{t}$), i mitjançant el vèrtex tWb de la interacció EW, on es produeix un sol quark top en associació amb altres partícules.

- $t\bar{t}$: És la font més gran de producció de quarks top en col·lisions hadròniques. Aquest procés és un dels més importants a l'LHC, ja que permet estudiar amb precisió les propietats del quark top. A més, a causa de la seu dominància en la producció, $t\bar{t}$ és també un fons rellevant en moltes anàlisis tal i com la d'aquesta tesi.
- Quark top en solitari: Aquest mecanisme té una secció eficaç tres vegades menor a la de $t\bar{t}$ i es produeix gairebé exclusivament a través del vèrtex tWb de la interacció EW. En primer ordre (LO), hi ha tres modes de producció per al quark top en solitari, sent el canal- t el mecanisme dominant a l'LHC amb aproximadament el 70% de la secció eficaç d'aquest mode a $\sqrt{s} = 13$ TeV. Els altres processos són el s -channel i la producció associada tW . Només el canal- t i la producció tW són rellevants per a la producció d'un sol quark top.

Les produccions associades de quark top són processos importants per tal de mesurar l'acoblamet del quark top amb les altres partícules de l'SM. El quark top en solitari es pot produir en associació amb altres partícules (tX) com els bosons Z , W , γ o H . El tema d'aquesta tesi gira precisament al voltant d'una producció associada particular. Més concretament, la producció amb un bosó de Higgs (tHq).

Pel que fa a la seu desintegració, el quark top es desintegra quasi completament ($\sim 99.8\%$) a través del vèrtex tWb en un quark b i un bosó W . Per això és habitual fer la classificació segons la desintegració subsegüent del bosó W . Per al W^+ , les BR (fraccions de desintegració) per als diferents modes de desintegració són [10]:

$W^+ \rightarrow e^+ \nu_e$	$(10.71 \pm 0.16)\%$
$W^+ \rightarrow \mu^+ \nu_\mu$	$(10.63 \pm 0.15)\%$
$W^+ \rightarrow \tau^+ \nu_\tau$	$(11.38 \pm 0.21)\%$
$W^+ \rightarrow q\bar{q}$ (hadrons)	$(67.41 \pm 0.27)\%$
$W^+ \rightarrow \text{invisible}$	$(1.4 \pm 2.9)\%$

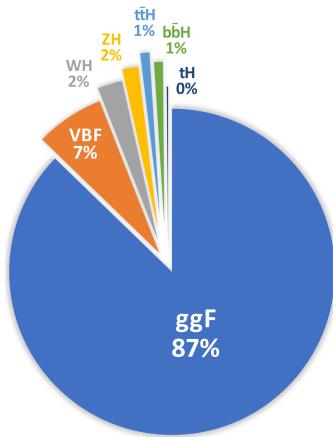
Per als processos conjugats que involucren el W^- , les BR són les mateixes.

1.3 La física del bosó de Higgs

Després del quark top, el bosó de Higgs (H) o simplement Higgs és la partícula més massiva del SM amb una massa de $m_H = 125.25 \pm 0.17$ GeV [10]. La seua existència va ser teoritzada l'any 1964 per tres grups teòrics treballant independentment: Englert-Brout [11], Higgs [12] i Guralnik-Hagen-Kibble [21].

Els experiments ATLAS [77] i CMS [78] de l'LHC van anunciar el descobriment del bosó de Higgs en l'any 2012, la qual cosa va suposar un de les grans fites de l'SM.

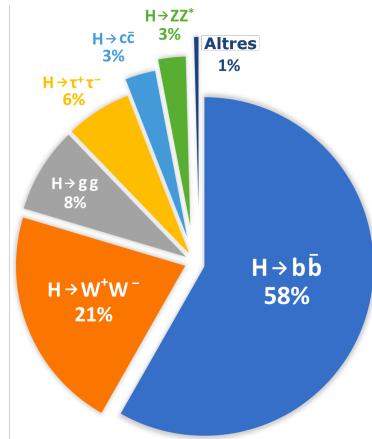
Els quatre processos més dominants per a la producció de bosons de Higgs en l'LHC són: fusió de gluons (ggF), fusió de bosons vectorials (VBF), Higgsstrahlung (VH), producció associada a parell de quarks ($q\bar{q}H$) i producció associada a un quark top i un Higgs ($t\bar{t}H$). Per a una massa $m_H = 125.2$ GeV, les seccions eficaces de producció de Higgs són [83]:



$$\begin{aligned}\sigma_{ggF} &= 48.5^{+2.2}_{-3.3} \text{ pb} \\ \sigma_{VBF} &= 3.78 \pm 0.05 \text{ pb} \\ \sigma_{WH} &= 1.37 \pm 0.03 \text{ pb} \\ \sigma_{ZH} &= 0.89^{+0.04}_{-0.03} \text{ pb} \\ \sigma_{t\bar{t}H} &= 0.5^{+0.03}_{-0.05} \text{ pb} \\ \sigma_{b\bar{b}H} &= 0.49^{+0.10}_{-0.11} \text{ pb} \\ \sigma_{tH} &= 0.09 \pm 0.01 \text{ pb}\end{aligned}$$

El bosó de Higgs té una vida mitjana relativament curta³ ($\tau_H = 1.6 \times 10^{-22}$ s [83]) i, per tant, sempre es detecta a través dels seus productes de desintegració. Les seves fraccions de desintegració segons la seu importància i assumint una $m_H = 125.2$ GeV són [83]:

³La més baixa després del quak top i els bosons EW.



$H \rightarrow b\bar{b}$	$(57.92 \pm 0.29)\%$
$H \rightarrow W^+ W^-$	$(21.70 \pm 0.11)\%$
$H \rightarrow gg$	$(8.17 \pm 0.26)\%$
$H \rightarrow \tau^+ \tau^-$	$(6.24 \pm 0.03)\%$
$H \rightarrow c\bar{c}$	$(2.888 \pm 0.014)\%$
$H \rightarrow ZZ$	$(2.667 \pm 0.013)\%$
$H \rightarrow \gamma\gamma$	$(2.270 \pm 0.023)\%$
$H \rightarrow \mu^+ \mu^-$	$(2.165 \pm 0.011)\%$
$H \rightarrow Z\gamma$	$(0.155 \pm 0.008)\%$
$H \rightarrow \text{Altres}$	$< 0.2\%$

1.4 Interacció entre el quark top i el bosó de Higgs

Com que el quark top és la partícula més massiva, s'espera que y_t siga el més fort entre tots els acoblaments del bosón de Higgs amb fermions i, per tant, el seu estudi té una importància crucial, tal com es discuteix a les referències [84] i [85]. S'espera que l'acoblament Yukawa estiga de l'ordre de la unitat:

$$y_t = \frac{\sqrt{2} \cdot m_{\text{top}}}{v} = 2^{3/4} \sqrt{G_F \cdot m_{\text{top}}} \approx 0.995 \approx 1,$$

on G_F és la constant. Aquest valor és molt més gran que els acoblaments dels altres quarks. Per comparació, $y_b \approx 0.025$ i $y_c \approx 0.007 \gg y_{s,d,u}$.

La producció d'una parella de quarks top juntament amb un bosó de Higgs ($t\bar{t}H$) és el principal mecanisme de producció associada d'aquestes dues partícules. El process $t\bar{t}H$ permet mesurar el valor absolut de y_t . D'altra banda, amb una secció eficaç molt més baixa que $t\bar{t}H$, la producció del bosó de Higgs accompanyant un sol quark top (tH) aporta informació valuosa, especialment pel que fa al signe del coupling de Yukawa. La producció de processos tH en general i tHq en particular constitueix el tema central de la recerca descrita a aquesta tesi.

L'associada producció tH té lloc a través de tres tipus diferents de processos. En primer lloc, el canal- t , on el bosó de Higgs es radia desde el quark top o el bosó W (Figures 2a i 2b, respectivament). En aquest canal, el top i el Higgs es creen juntament amb un quark addicional, donant lloc a la denominada producció tHq . L'altre procés d'interès és la producció tWH , que consisteix en un procés tW on el Higgs s'emiteix desde el quark top (Figura 2c). El tercer mode de producció és el canal- s . Aquest últim és poc rellevant degut a la seua petita secció eficaç.

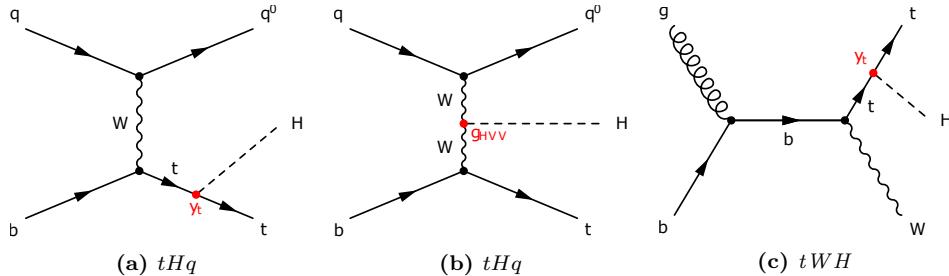


Figura 2: Representació dels diagrames de Feynman a LO per a tH . A les figures (a) i (b) es mostra el procés tHq . Ací el bosó de Higgs s'acopla al quark top (a) o al bosó W (b). A la figura (b), g_{HVV} és el coupling del bosó de Higgs amb els bosons vectorials. A la figura (c), es mostra la producció tWH .

1.4.1 Caracterització del bosó de Higgs en tH i sensitivitat cap a y_t

El Lagrangià efectiu que descriu l'acoblament de Yukawa entre el bosó de Higgs i el quark top per sota de l'escala de ruptura de la simetria EW és el següent [97]:

$$\mathcal{L} = -\frac{y_t}{\sqrt{2}} \bar{\psi}_t [\cos(\alpha) \kappa_{Htt} + i \sin(\alpha) \kappa_{Att} \gamma^5] \psi_t X_0 .$$

Ací, ψ_t i X_0 representen el quark top i el bosó de Higgs respectivament, i α és l'angle de mescla de CP. Els paràmetres d'escala κ_{Htt} i κ_{Att} són adimensionals i reals.

La Figura 3 mostra la secció eficaç per a la producció tX_0 en el canal- t com a funció de l'angle α . Per a comparar, també es mostra la secció eficaç $t\bar{t}X_0$. Ací tX_0 i $t\bar{t}X_0$ representen respectivament els processos els tH i $t\bar{t}H$.

A l'hora d'observar la Figura 3, la primera apreciació a tenir en compte és que la secció eficaç $t\bar{t}H$ és simètrica al voltant de $\alpha = \pi/2$. Això implica que mesurant $\sigma(t\bar{t}H)$ no seria possible discriminar entre escenaris de CP parells i imparells. Per contra, per al procés tH , no existeix eixa simetria i, llavors $\sigma(tH)$ és sensible a un possible acoblament que viole CP. Això es deu al fet que en l'SM, la producció tHq on el H es coupla amb el W (Figura 2 (b)) interfereix destructivament amb aquells en els quals el H és radiat des del top (Figura 2 (a)).

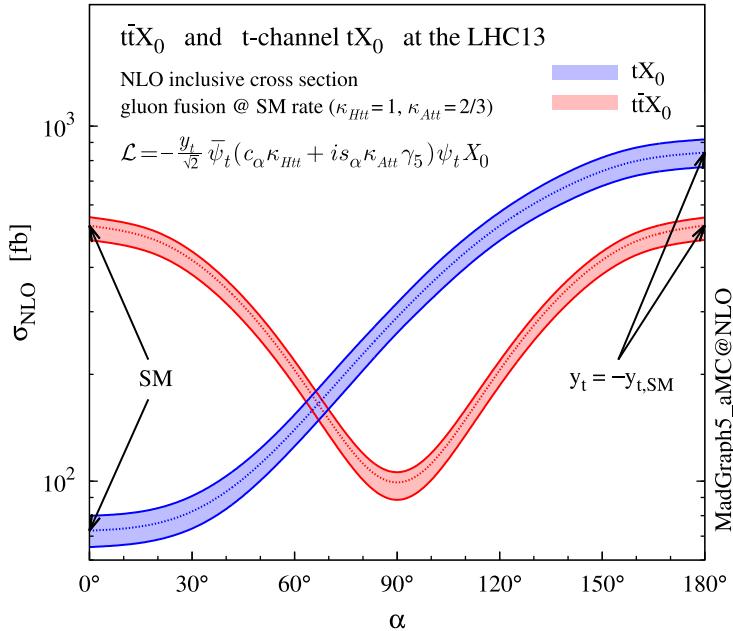


Figura 3: Secció eficaç a NLO com a funció de l'angle de mescla de CP per a la producció tX_0 en el canal- t i $t\bar{t}X_0$ a $\sqrt{s} = 13$ TeV. La partícula X_0 representa un bosó de Higgs amb violació de CP [97].

2 L'experiment ATLAS de l'LHC al CERN

El Gran Col·lisionador d'Hadrons (LHC, per les seues sigles en anglés) és un accelerador de partícules que es troba al CERN (Centre Europeu per a la Recerca Nuclear o Laboratori Europeu de Física de Partícules Elementals), en Ginebra, Suïssa. Va ser dissenyat per a col·lisionar protons i ions pesats amb alta energia, permetint així als científics estudiar l'estructura subatòmica de la matèria i buscar noves partícules.

L'LHC és una màquina d'avantguarda que utilitza tecnologia avançada per accelerar partícules fins a velocitats properes a les de la llum abans de xocar-les entre si. El funcionament de l'LHC es basa en l'ús de camps magnètics i radiofreqüències per accelerar les partícules carregades. A continuació, els feixos de partícules són dirigits per col·lisionar en els punts específics on es troben els detectors com ATLAS i CMS. Aquestes col·lisions generen partícules secundàries que es registren mitjançant una xarxa de detectors situats en el seu interior. Les dades recopilades d'aquestes col·lisions es fan servir per a investigar la física de partícules. Amb energies de $\sqrt{s} = 13$ TeV, l'LHC és l'accelerador de partícules més gran construït, la qual cosa constitueix una eina clau per a l'avanç de la ciència. Els quatre principals detectors que envolten l'LHC són: ATLAS, CMS, LHCb i

ALICE. El primer d'aquests és l'experiment en el qual es desenvolupa el treball que es descriu en aquesta tesi doctoral.

2.1 El detector ATLAS

El detector ATLAS s'utilitza per mesurar les propietats de les partícules resultants de les col·lisions d'hadrons. ATLAS té una estructura cilíndrica i és un dels detectors de partícules més grans del món, amb aproximadament 46 metres de llargària i 25 metres de diàmetre. Està compost per diversos components i subcomponents. Cada un d'aquests sistemes s'encarrega de registrar un tipus d'informació diferent. En ordre de dins cap a fora, ATLAS està format per:

- Detector intern (ID): Aquest component és el més proper al punt de col·lisió i té com a finalitat rastrejar les trajectòries de les partícules carregades. L'ID està format per tres subcomponents:
 - Detector de Píxels: És la part més interna de l'ID. El material de detecció és silici de $250\text{ }\mu\text{m}$ de grossor. Cada mòdul conté 16 xips de lectura i altres components electrònics. La unitat mínima de detecció és el píxel ($50 \times 400\text{ }\mu\text{m}$), dels quals hi ha aproximadament 47000 per mòdul. La mida minúscula dels píxels està dissenyada per rastrejar de manera extremadament precisa prop del punt d'interacció. El detector de píxels està compost per quatre capes dobles de tires de silici i té 6.3 milions de canals de lectura i una àrea total de 61 m^2 .
 - Rastrejador de Semiconductors (SCT): Té un concepte i una funció similar al Detector de Píxels, però amb tires llargues i estretes en lloc de petits píxels, fet que fa possible cobrir una àrea més extensa. Cada tira té una mida de $80\text{ }\mu\text{m} \times 12\text{ cm}$. El SCT és la part més crítica del detector intern per al rastreig bàsic en el pla perpendicular al feix, ja que mesura partícules en una àrea molt més àmplia que el Detector de Píxels.
 - El Rastrejador de Radiació de Transició (TRT): Del l'anglès *Transition Radiation Tracker*, és un rastrejador de tubs de deriva⁴. Cada tub té un diàmetre de 4 mm, una longitud de fins a 144 cm i està ple d'una mescla de gasos (principalment Xenó).
- Imant solenoïdal: És un imant superconductor que envolta l'ID i genera un intens camp magnètic per desviar la trajectòria de les partícules carregades.

⁴Straw chamber en anglès, es tracta d'un tub llarg amb un fil al centre i un gas que s'ionitza quan una partícula el travessa.

- Calorímetre electromagnètic (ECAL): Aquest component té com a funció principal mesurar l'energia de les partícules que interactuen electromagnèticament, com els fotons i els electrons. Està format per capes de cristalls d'escintil·lació que generen senyals de llum quan les partícules interactuen amb ells.
- Calorímetre hadrònic (HCAL): A diferència de l'ECAL, aquest calorímetre està dissenyat per mesurar l'energia de les partícules hadròniques, com els pions i els protons. Està compost per capes de material dens que interactuen amb les partícules, generant una cascada de partícules secundàries que són detectades i mesurades.
- Espectròmetre de muons (MS): Aquest component està dissenyat per mesurar i rastrejar els muons, que tenen una gran capacitat de penetració. L'MS utilitza diferents tecnologies de detectors per identificar i mesurar les trajectòries i les energies dels muons.

2.2 Alineament del detector intern d'ATLAS

Determinar amb precisió la trajectòria d'una partícula dins del detector és una tarea determinant per poder reconstruir diversos observables físics. D'una banda la reconstrucció de les trajectòries permet identificar quines partícules carregades han estat produïdes en una col·lisió. D'altra, la trajectòria d'una partícula carregada en un camp magnètic proporciona informació sobre la seua quantitat de moviment \vec{p} .

La part del detector que s'encarrega de reconstruir les trajectòries és l'ID. Quan les partícules carregades travessen el detector, deixen dipòsits d'energia a cada subdetector (píxel), que permeten la reconstrucció de la trajectòria de la partícula. Amb més de 2 metres d'altura i 6 metres de longitud, l'AIDr té una precisió de mesura de posició millor que una centèsima de mil·límetre. Per aconseguir aquesta precisió, el detector ha d'estar alineat amb una precisió igual o millor.

Un coneixement detallat de la geometria de l'ID és crucial per a una reconstrucció precisa. La posició i forma dels subdetectors no són estacionàries sino que canvien amb el temps. Els elements del detector poden desplaçar-se a causa de fluctuacions de temperatura o de canvis en el camp magnètic. Per tindre en compte aquests moviments, es realitza amb regularitat l'alignament del ID i es determina la geometria real de cada element actiu del subdetector. Aquest alignament s'ha de dur a terme mitjançant mètodes indirectes, perquè el detector és inaccessible durant el funcionament de l'LHC.

L'algoritme d'alignament d'ATLAS està basat en la comparació dels punts d'impacte registrats amb la trajectòria reconstruïda de les partícules. Si la posició d'un subdetector és correctament coneguda, les diferències residuals s'equilibraran. En cas contrari, els residuals mostraran un desviament sistemàtic, indicant el desalineament d'un subdetector. Aquest és un procés iteratiu que segueix l'estructura jeràrquica de l'ID, primer alineant les grans estructures físiques i després els elements individuals del subdetector. La Figura 4 mostra les diferències residuals produïdes pel desalineament de l'ID.

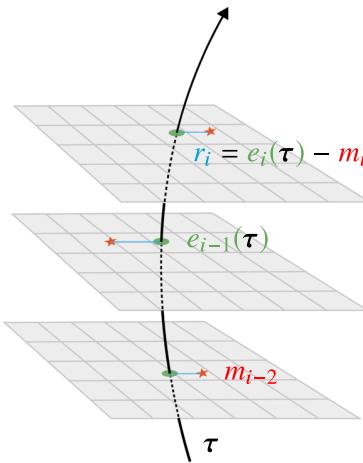


Figura 4: La línia negra és la trajectòria reconstruïda de la partícula que creua l'ID. L'estrella roja representa la mesura en cada capa del detector i les diferències residuals estan representats per la linea blava [148].

2.2.1 Visor Web de Monitoratge de l'Alineament de l'ID

Atés que qualsevol desalineació dels diferents elements de l'ID deteriorarà la qualitat de la reconstrucció de trajectòries i objectes és obligatori un seguiment constant de l'alignació. Durant la producció d'aquesta tesi, he contribuït a l'alignament de l'ID a través del desenvolupament del software de monitoratge dels resultats de l'alignament de l'ID d'ATLAS descrits amunt.

El visor web de Monitoratge de l'alignació de l'ID és una aplicació destinada a la monitoratge dels resultats de l'alignament basats en trajectòries aconseguites en el bucle de calibració per a l'ID. Ajuda a avaluar les correccions d'alignació calculades així com moltes distribucions gràfiques directament relacionades amb el rendiment (per exemple, els residus del detector). A la Figura 5 es mostra la pàgina principal del visor web que he desenvolupat.

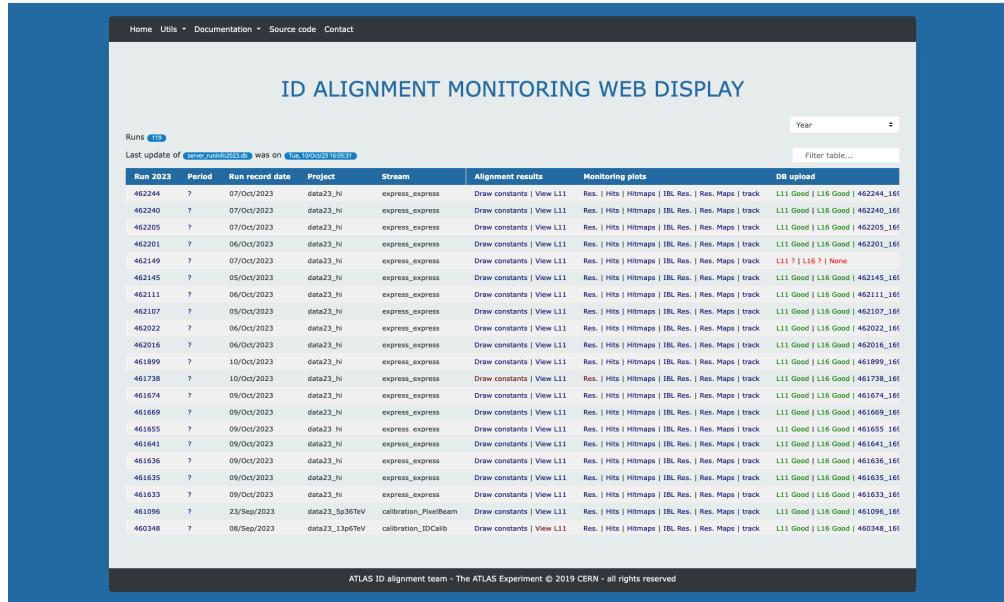


Figura 5: Pàgina principal del Visor Web de Monitoratge de l’Alineació de l’ID.

Les meues responsabilitats incloïen el desenvolupament i implementació del codi tant per a la part client (frontend) com per a la part del servidor (backend) per al Visor Web de Monitoratge de l’Alineació de l’ID. S’han realitzat millores significatives en aquesta eina, incloent-hi la integració amb la configuració estàndard d’Athena, la conformitat amb els estàndards d'estil d'ATLAS, millors en l'eficiència d'execució i actualitzacions dinàmiques, així com una major interactivitat del frontend. A més, s’ha optimitzat el codi per a eliminar redundàncies i s’ha millorat la interfície d’usuari per a facilitar l'accés i la gestió de les dades.

Aquestes millores no només augmenten la velocitat d’execució de l’aplicació, sinó que també proporcionen accés a més informació i permeten gestionar i accedir fàcilment a les dades desitjades. El Visor Web de Monitoratge de l’Alineació de l’ID que he desenvolupat s’està utilitzant activament pels membres de la col·laboració ATLAS per a monitorar l’alineació de l’ID durant la Run 3 així com en el reprocessament de les dades de la Run 2. Això ha simplificat significativament el monitoratge de l’alineació, fent la tasca més eficient i fàcil de fer servir.

3 Recol · lecció de dades, simulació i reconstrucció d'objectes

3.1 Simulació i adquisició d'esdeveniments en ATLAS

La simulació i l'adquisició de dades són també una part fonamental de l'anàlisi, ja que és amb elles amb les quals es realitzarà posteriorment l'estudi del procés tHq .

Per l'anàlisi descrita en aquesta tesi, les dades utilitzades són les recollides pel detector ATLAS durant el Run 2, és a dir, del 2015 al 2018. Les dades van ser produïdes durant col·lisions pp a l'LHC amb una freqüència de 25 ns a $\sqrt{s} = 13$ TeV. La lluminositat total integrada aconseguida durant aquest període va ser de 140 fb^{-1} , amb un error que varia entre el 2,0% i el 2,4%.

La simulació d'esdeveniments, també coneguda com a simulació Monte Carlo (MC), es divideix en diverses etapes que abasten des del càcul de la secció eficaç a escala de partons (quarks i gluons) fins a les cascades de partons i els efectes no pertorbatius del procés, i finalment, la simulació de la interacció de les partícules pel detector i el senyal elèctrica d'aquest. Els diferents passos d'una col·lisió de pp que cal tenir en compte es mostren de forma esquemàtica a la Figura 6.

La simulació de MC inclou diversos processos físics: la dispersió de QCD, també coneguda com la creació dels elements de matriu; la cascada de partons, l'hadronització, els processos de dispersió de QCD secundaris, desintegracions i el pile-up. Cada un d'aquests processos es simula de forma independent i utilitza diferent informació d'entrada.

Hi ha diversos programes al mercat per realitzar una o diverses parts de la simulació. Els utilitzats en algunes de les simulacions a l'anàlisi inclosa en aquesta tesi són: POWHEG BOX, MADGRAPH5_AMC@NLO, SHERPA, PYTHIA i HERWIG. Alguns d'ells han estat utilitzats per la simulació nominal d'un procés, mentre que d'altres s'han utilitzat per obtenir l'error derivat de l'ús d'un programa o l'altre.

L'últim pas de la simulació d'esdeveniments és la simulació del pas de les partícules pel detector, la qual cosa es realitza a través d'Athena [142, 163], un software específic de la col·laboració ATLAS, utilitzant GEANT4 [185].

Les diferents combinacions de generadors de MC utilitzades tant per al procés de senyal com per a tots els fons considerats es mostren en la Taula 1. La influència a l'anàlisi dels processos llistats a la Taula 1 no és uniforme.

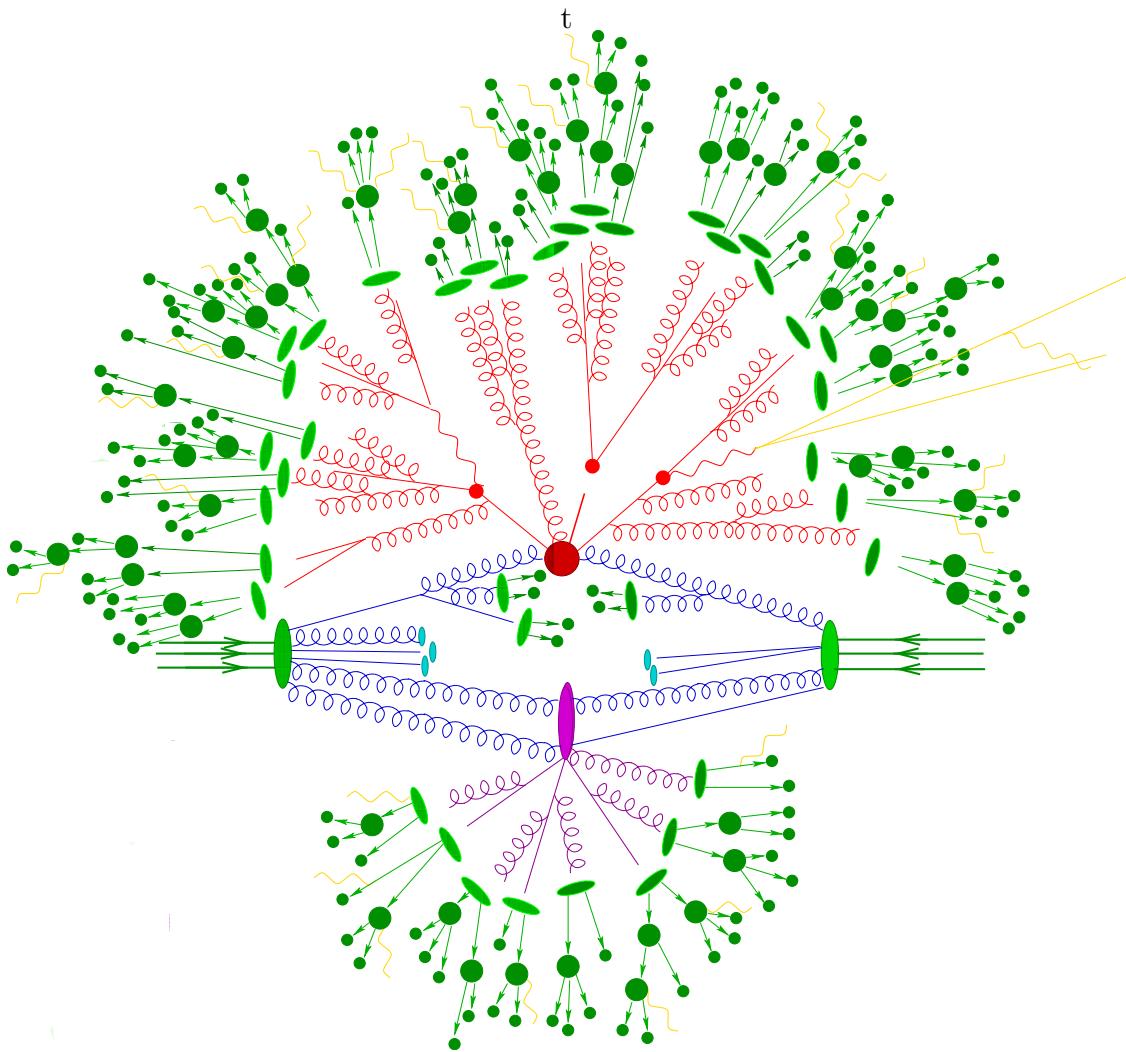


Figura 6: Esquema d'una col·lisió pp per a produir un esdeveniment $t\bar{t}H$. El gran punt roig representa la interacció forta principal, seguida de la desintegració del bosó Higgs i els dos quarks top, representats pels tres punts rojos més petits. Els punts blaus indiquen els partons en el seu estat inicial. Els processos de dispersió forta secundaris es mostren en violeta. Finalment, es mostra el procés d'hadronització (en verd clar) i els estats finals hadrònics (en verd). A més, apareixen unes línies grogues que representen la radiació de fotons.

Procés	Generador de MC	Ordre (esquema)	Joc de PDF	Cascada de partons	Joc de PDF
Signal					
tHq	MADGRAPH5_AMC@NLO 2.6.2	NLO (4FS)	NNPDF3.0NLO nf4	PYTHIA 8.230	NNPDF2.3LO (A14 tune)
Fons					
$t\bar{t}$	POWHEG BOX v2	NLO (5FS)	NNPDF3.0NLO	PYTHIA 8.230	NNPDF2.3LO (A14 tune)
$V+jets$	SHERPA 2.2.1	NLO+LO	NNPDF3.0NNLO	-	-
Diboson	SHERPA 2.2.1-2	NLO+LO	NNPDF3.0NNLO	-	-
Triboson	SHERPA 2.2.2	NLO+LO	NNPDF3.0NNLO	-	-
$t\bar{t}Z$	MADGRAPH5_AMC@NLO 2.3.3	NLO	NNPDF3.0NLO	PYTHIA 8.210	NNPDF2.3LO (A14 tune)
$t\bar{t}W$	SHERPA 2.2.10	NLO	NNPDF3.0NNLO	-	-
$t\bar{t}H$	POWHEG BOX v2	NLO (5FS)	NNPDF3.0NLO	PYTHIA 8.230	NNPDF2.3LO (A14 tune)
t -channel	POWHEG BOX v2	NLO (4FS)	NNPDF3.0NLO nf4	PYTHIA 8.230	NNPDF2.3LO (A14 tune)
tW	POWHEG BOX v2	NLO (5FS, DR)	NNPDF3.0NLO	PYTHIA 8.230	NNPDF2.3LO (A14 tune)
s -channel	POWHEG BOX v2	NLO	NNPDF3.0NLO	PYTHIA 8.230	NNPDF2.3LO (A14 tune)
tZq	MADGRAPH5_AMC@NLO 2.3.3	NLO	NNPDF3.0NLO	PYTHIA 8.230	NNPDF2.3LO (A14 tune)
tWH	MADGRAPH5_AMC@NLO 2.8.1	NLO (5FS, DR)	NNPDF3.0NLO	PYTHIA 8.245p3	NNPDF2.3LO (A14 tune)
tWZ	MADGRAPH5_AMC@NLO 2.3.3	NLO	NNPDF3.0NLO	PYTHIA 8.212	NNPDF2.3LO (A14 tune)
ttt	MADGRAPH5_AMC@NLO 2.2.2	NLO	NNPDF3.1NLO	PYTHIA 8.186	NNPDF2.3LO (A14 tune)
$ttt\bar{t}$	MADGRAPH5_AMC@NLO 2.3.3	NLO	NNPDF3.1NLO	PYTHIA 8.230	NNPDF2.3LO (A14 tune)
ggH	POWHEG BOX v2	NLO	CT10	PYTHIA 8.210	CTEQ6L1 (AZNLO tune)
qqH	POWHEG BOX v1	NLO	CT10	PYTHIA 8.186	CTEQ6L1 (AZNLO tune)
WH	PYTHIA 8.186	LO	NNPDF2.3LO	-	-
ZH	PYTHIA 8.186	LO	NNPDF2.3LO	-	-

Taula 1: Resum de les mostres d'esdeveniments simulats de senyal i fons bàsics utilitzades en els anàlisis de tHq .

3.2 Identificació i reconstrucció d'objectes en ATLAS

Els senyals proporcionats pels diferents components del detector ATLAS són transformats en objectes físics reconstruïts per poder emprar-los en les diferents anàlisis. Un objecte físic fa referència a una partícula o entitat similar a una partícula reconstruïda detectada dins d'ATLAS com a conseqüència d'una col·lisió de partícules. Els principals objectes físics que es detecten i reconstrueixen per a aquesta tesi són:

- **Electrons:** Reconstruïts utilitzant els dipòsits d'energia en el ECAL, coincidents amb trajectòries a l'ID. Sense una bona resolució a l'hora de reconstruir les traces, seria impossible identificar i reconstruir partícules carregades [209]. Per això, un correcte alineament del ID és fonamental, tal com es descriu en la Secció 2.2. A aquesta anàlisi, les condicions fetes servir per etiquetar un objecte com a electró són les recomanades pel *Combined Performance group* i consisteixen a tenir un p_T superior a 10 GeV, estar dins de l'interval de pseudorapidesa de $|\eta^{\text{clust}}| < 2.47$, i superar el nivell d'identificació estricte. S'apliquen condicions sobre l'aillament per a reduir l'impacte dels electrons que no provenen del procés principal (electrons *non-prompt*).
- **Muons:** Es reconstrueixen associant traces en la cambra de muons amb traces en l'ID, complementant aquesta informació amb el senyal dels calorímetres [215]. De nou, es veu la necessitat d'un bon alineament a l'ID. Al

nostra anàlisi es requereix que els muons tinguen $p_T > 7\text{GeV}$, $|\eta| < 2.5$, i que passen un criteri d'identificació mitjà Igual que amb els electrons, es demanda una estricta aïllament dels candidats a muons.

- **Leptons tau:** El leptó τ , amb una massa de 1.78 GeV, és l'únic leptó prou massiu per desintegrar-se en hadrons. En aproximadament un terç dels casos, els leptons τ es desintegren de manera leptònica (τ_{lep}) en un muó o un electró amb dos neutrins. Els τ_{lep} no es reconstrueixen com a tal, sinó que directament s'identifiquen els muons i electrons. En els casos restants, els tau leptons es desintegren de manera hadrònica, en una combinació de mesons carregats i neutres amb un ν_τ . Els leptons τ que es desintegren hadrònicament, conegeuts com a τ_{had} , es reconstrueixen i identifiquen fent servir una xarxa neuronal recurrent. A més, una BDT per a vetar electrons es utilitzada.

Per a acceptar un objecte físic com a τ_{had} es precís que aquest tinga associades una o tres traces i que es trobe en espai definit $|\eta^{\text{clust}}| < 2.5$ excluint $1.37 < |\eta^{\text{clust}}| < 1.52$. També es requereix un p_T mínim de 20 GeV. Una Xarxa Neuronal Recurrent (RNN) és emprada per realitzar l'identificació.

- **Jets:** Formats per la conglomeració de dipòsits d'energia al calorímetre hadrònic, que sorgeixen de l'hadronització de quarks i gluons produïts en les col·lisions. Els jets són feixos de partícules estretament agrupades i altament col·limades que es mouen en la mateixa direcció, seguint la direcció original del quark o gluó original. Aquests objectes es reconstrueixen a partir dels dipòsits d'energia en els calorímetres i les traces de l'ID. Aquesta informació és combinada utilitzant l'algoritme Anti- k_t [221].

Per a aquesta anàlisi, s'aplica el requisit de que el p_T siga superior a 20 GeV i es trobe contigunt en $|\eta| < 4.5$.

- **Jets- b :** Jets que provenen de l'hadronització de quarks bottom. La seua identificació es fa mitjançant diferents algoritmes que exploten els seus trets característics, com els seus paràmetres d'impacte, la presència de vèrtexs secundaris o la seua topologia de desintegració. Particularment, l'algoritme DL1r és utilitzat en aquesta anàlisi. Aquest mètode està basat en una xarxa neural recurrent [276]. També cal que es satisfaça que $p_T > 20\text{GeV}$, $|\eta^{\text{clust}}| < 2.5$.
- **Energía transversa mancant:** També conegeuda com a moment transvers de falta, és una quantitat important que representa el desequilibri del moment transvers en un esdeveniment. La E_T^{miss} es deu a partícules no detectades, que en el cas de la nostra anàlisi són els neutrins.

#	0 τ_{had}	1 τ_{had}	2 τ_{had}
1 ℓ (e/μ)	$tHq(b\bar{b})$ 1 ℓ		$tHq(WW/ZZ/\tau\tau)$ 1 $\ell + 2\tau_{\text{had}}$
2 ℓ (e/μ)	$tHq(WW/ZZ/\tau\tau)$ 2 ℓ SS	$tHq(WW/ZZ/\tau\tau)$ 2 $\ell + 1\tau_{\text{had}}$	
3 ℓ (e/μ)	$tHq(WW/ZZ/\tau\tau)$ 3 ℓ		

Taula 2: Diferents canals per a la producció de tHq segons la presència de leptons de sabor lleuger i taus que es desintegren hadrònicament en l'estat final. Els requisits de multiplicitat sobre els objectes d'estat final, asseguren l'ortogonalitat entre canals per a una futura combinació.

Canal	$H \rightarrow \tau\tau$	$H \rightarrow WW^*$	$H \rightarrow ZZ^*$
2 $\ell + 1\tau_{\text{had}}$	63	32	5
1 $\ell + 2\tau_{\text{had}}$	96	3	1
2 ℓ SS	17	80	3
3 ℓ	14	69	17

Taula 3: La probabilitat percentual dels modes de desintegració del bosó de Higgs dins dels diferents canals multipleptònics ha sigut calculada utilitzant les fraccions de desintegració.

4 Recerca de processos tHq amb un estat final $2\ell + 1\tau_{\text{had}}$

L'objectiu principal d'aquesta anàlisi és establir límits en la secció eficaç de producció del procés tHq . La producció de tHq s'analitza a través de sis canals ortogonals optimitzats de manera independent (presentats a la Taula 2), centrant-se aquesta tesi en l'estat final amb dos leptons lleugers (e/μ) i un leptó tau que es desintegra hadrònic (τ_{had}). El canal $2\ell + 1\tau_{\text{had}}$ es divideix en dues subcategories dependent del signe relatiu de la càrrega elèctrica mostrada pels dos leptons lleugers carregats: El canal 2ℓ OS + $1\tau_{\text{had}}$ si la càrrega és oposada i el 2ℓ SS + $1\tau_{\text{had}}$ si tenen la mateixa càrrega. Els diagrames de Feynman corresponent a aquests dos modes de producció es presenten a la Figure 7.

Per als canals amb estats finals multileptònics, l'anàlisi considera tres modes de desintegració del bosó de Higgs: $H \rightarrow W^+W^-$, $H \rightarrow \tau^-\tau^+$ y $H \rightarrow ZZ$. La Taula 3 mostra l'importància de cada un dels modes de desintegració dins de cada canal de cerca tHq . Es important notar que, en conjunt, aquests tres modes representen únicament el 21% del total de la fracció de dessintegració del bosó de Higgs tal i com s'explica a la Secció 1.3.

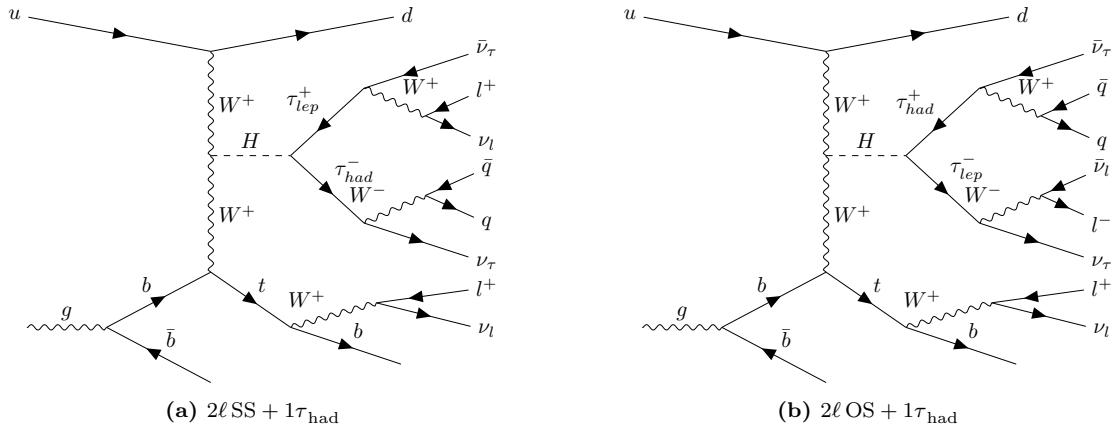


Figura 7: Diagrames a ordre dominant per a la producció tHq ($2\ell + 1\tau_{\text{had}}$) al canal de desintegració $H \rightarrow \tau_{\text{had}}\tau_{\text{lep}}$. Es pot apreciar que el dos leptons lleugers al diagrama (a) tenen la mateixa càrrega elèctrica mentre que en (b) la càrrega és oposada.

4.1 Estudis sobre la senyal

Per tal de poder separar els processos tHq de els fons, cal comprendre les propietats d'aquesta producció. D'aquesta manera podrem definir observables físics (comunament referits com variables) que ens permeten discriminar el senyal per tal crear regions a l'espai de fases enriquides en processos tHq , tal com es fa en la Secció 4.3. El primer pas per fer açò és obtenir la informació del senyal a escala de partons. Després, utilitzant aquesta informació, l'origen del leptons lleugers (és a dir, si venen de la desintegració del bosó de Higgs o de la del quark top) pot ser determinat. Finalment, açò pot ser emprat per reconstruir les propietats cinemàtiques dels processos tHq i, amb aquestes, definir variables.

4.1.1 Validació de les simulacions a nivell de partons

Dins de la col·laboració ATLAS, per tal de generar i administrar la informació a nivells de partons, el paquet de software `TopPartons` és emprat. Una de les contribucions d'aquesta tesi ha sigut el desenvolupament dels codis necessaris per a incloure en aquest paquet els processos tHq . Per tal de validar que el programa elaborat funciona correctament, s'han realitzat càlculs teòrics per comparar-los amb el resultat del paquet.

El paquet elaborat guarda a les NTuples⁵ la informació de totes les partícules que apareixen al diagrama de Feynman dels processos tHq . Aquesta informació consisteix en el PDG-ID⁶, p_T , η , ϕ i massa. Addicionalment, per als leptons τ , es guarda la informació sobre si aquests es desintegren lacònicament o hadrònicament. Aquest desenvolupament es fet servir per tots els canals que participen en la cerca del procés tHq .

Pel que fa als càlculs per validar TopPartons, aquests s'han fet combinant les fraccions de desintegracions de totes les partícules involucrades a les cadenes de desintegració del quark top i el bosó Higgs. Després s'han estudiat totes les combinacions d'aquestes partícules que podien generar els estats finals $2\ell + 1\tau_{\text{had}}$. Gràcies a aquests càlculs podem conéixer les contribucions de cada canal de desintegració a l'estat final i també determinar on es produeix el τ_{had} . Aquests resultats són recollits a la Taula 4.

Canal de desintegració	Origen del τ_{had}		Total
	Quark top	Bosó de Higgs	
$H \rightarrow \tau\tau$	5.06	64.06	69.13
$H \rightarrow WW^*$	9.01	18.01	27.02
$H \rightarrow ZZ^*$	2.22	1.64	3.85
Total	16.29	83.71	100.00

Taula 4: Contribució en percentatge de cada canal de desintegració del bosó de Higgs a l'estat final $2\ell + 1\tau_{\text{had}}$. A més, l'origen del τ_{had} és presentat.

Finalment, els resultats obtinguts a través dels càlculs són comparats amb els del paquet software a través de la següent mètrica:

$$\frac{\text{Nombre d'esdeveniments}(H \rightarrow \text{Canal de desintegració} \rightarrow 2\ell + 1\tau_{\text{had}})}{\text{Nombre d'esdeveniments}(H \rightarrow \text{Canal de desintegració})}.$$

La Taula 5 recull el resultat de la comparació. La discrepància trobada a $H \rightarrow ZZ^*$ no és deguda a una falla en TopPartons sino que el codi que llig el resultat de TopPartons no considera certes combinacions per aconseguir l'estat final $2\ell + 1\tau_{\text{had}}$. Els mateixos tests s'han portat a terme a l'estat final 3ℓ i l'agraïment es donava per als tres canals de desintegració.

Tindre una informació adequada a escala de partons és fonamental per a l'anàlisi, ja que s'utilitza en diverses tasques des de la determinació de l'origen del

⁵Una *NTupla* es refereix a un format de dades utilitzat per emmagatzemar i organitzar informació detallada sobre les col·lisions de partícules i les simulacions. Les nostres NTuples estan al format ROOT [283].

⁶Un nombre que identifica la identitat de la partícula.

Fracció d'esdeveniments	Càlcul teòric	Resultat de TopPartons
$\frac{H \rightarrow \tau\tau \rightarrow 2\ell + 1\tau_{\text{had}}}{H \rightarrow \tau\tau}$	0.1246	0.1232 ± 0.0057
$\frac{H \rightarrow WW^* \rightarrow 2\ell + 1\tau_{\text{had}}}{H \rightarrow WW^*}$	0.0141	0.0151 ± 0.0009
$\frac{H \rightarrow ZZ^* \rightarrow 2\ell + 1\tau_{\text{had}}}{H \rightarrow ZZ^*}$	0.0164	0.0100 ± 0.002

Taula 5: Predicció teòrica comparada amb el resultat de TopPartons. La incertesa a la segona columna correspon a l'error estadístic.

leptó lleuger en el canal $2\ell \text{ SS} + 1\tau_{\text{had}}$ fins a l'estimació de la contribució al fons deguda a la mala identificació de τ_{had} .

4.1.2 Assignació de l'origen dels leptons lleugers

Els dos leptons lleugers a l'estat final del canal $2\ell + 1\tau_{\text{had}}$ poden provenir tant del bosó de Higgs com del quark top. Les ambigüïtats respecte a l'origen d'aquests leptons de sabor lleuger fan extremadament difícil la reconstrucció dels sistemes del quark top i del bosó de Higgs.

Segons els càlculs realitzats en la Secció 4.1.1, el τ_{had} es produeix el 83.7% de les vegades com a producte de la desintegració del bosó de Higgs, en oposició al 16% en què prové de la desintegració del quark top. Llavors, en la majoria del casos podem dir que hi ha un leptó lleuger que prové que bosó de Higgs (ℓ_{Higgs}) i un altre que prové del (ℓ_{top}).

Assumint que el τ_{had} es originat del sistema del bosó de Higgs, l'associació de quin lepton de sabor lleuger prové de la desintegració del quark top i quin del bosó de Higgs es pot fer directament si aquests dos leptons tenen càrrega elèctrica oposada, és a dir, en el canal $2\ell \text{ OS} + 1\tau_{\text{had}}$. Atès que el bosó de Higgs té càrrega neutra, la suma de la càrrega dels seus productes de desintegració hauria de ser zero. Per tant, en el canal $2\ell \text{ OS} + 1\tau_{\text{had}}$, mentre que el leptó lleuger amb càrrega oposada a la del τ_{had} és el que prové del bosó de Higgs, l'altre leptó, és a dir, el que té la mateixa càrrega que τ_{had} , és el que s'origina de la desintegració del quark top.

Per assignar amb precisió l'origen del leptó lleuger en l'escenari $2\ell \text{ SS} + 1\tau_{\text{had}}$, s'ha desenvolupat un mètode fonamentat en BDTs de gradient. La BDT^{OrigenLep} s'implementa utilitzant la biblioteca Toolkit for Multivariate Data Analysis (TMVA) de ROOT [283, 284]. La implementació d'aquesta BDT es pot esquematitzar a través dels següents passos:

1. **Etiquetatge:** La creació d'una etiqueta per a l'entrenament supervisat a través de l'ús d'informació a nivell de partons i l'establiment de categories per a la classificació. Les categories definides per a la classificació són:

- **Tipus 1:** el leptó lleuger amb més p_T (ℓ_1) és ℓ_{top} i el secundari (ℓ_2) és ℓ_{Higgs} .
- **Tipus 2:** ℓ_1 és ℓ_{Higgs} i ℓ_2 és ℓ_{top} .

Donat un esdeveniment simulat, és possible accedir tant a la seua informació a nivell de partons com a nivell de reconstrucció simultàniament. Per als leptons a nivell de partons es coneixen els progenitors, i per als leptons a nivell de reconstrucció, cal identificar els progenitors. Llavors, és possible comparar-los per a crear una associació.

Per a vincular els leptons lleugers a nivell de reconstrucció amb els leptons lleugers a nivell de partó, es construeix un con de $\Delta R < 0.01$ al voltant de cada leptó reconstruït. Quan dins d'aquest con hi ha exactament un leptó lleuger a nivell de partons, es produeix el que s'anomena "una coincidència". Per a identificar correctament l'origen del leptó en un esdeveniment, és necessari que ambdós leptons lleugers a nivell de reconstrucció tinguen coincidències a nivell de partòns. A la Figura 8 es mostren els leptons als dos nivells amb els cons de ΔR .

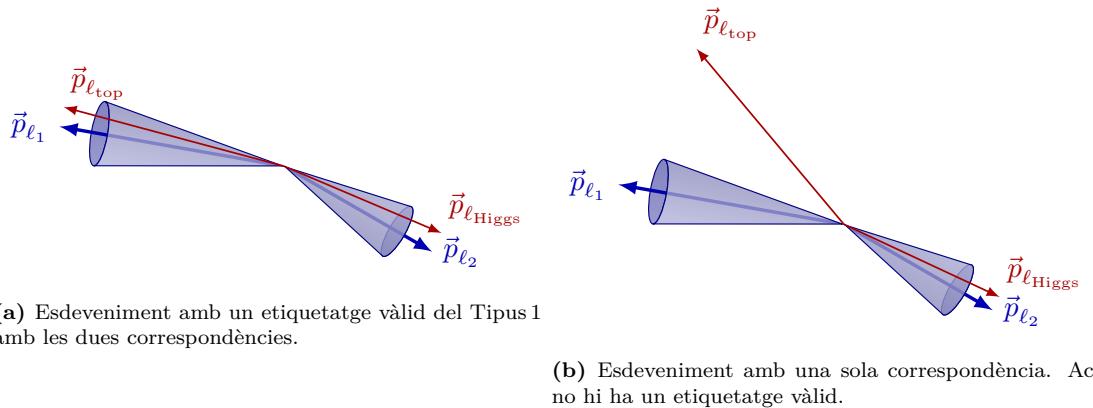


Figura 8: Diferents escenaris per a l'associació de leptons lleugers a nivell de reconstrucció (fletxa blava) i a nivell de partons (fletxa roja). Cal notar que les etiquetes ℓ_{top} i ℓ_{Higgs} només estan disponibles per a les partícules a nivell de partons.

2. **Determinación de variables per la BDT:** La selecció de característiques d'entrada a nivell de reconstrucció amb capacitat discriminatòria entre els Tipus 1 i Tipus 2. Aquestes variables no poden ser utilitzades després com a entrades en els BDTs per a la definició de regions. Les variables utilitzades a

la BDT^{OrigenLep} són les presentades a la Taula 6. Les distribucions d'aquestes variables poden observar-se a l'Appendix C.1.

Variable	Significat	$\langle S^2 \rangle (\times 0.1)$
$m_{\text{vis}}^{\text{opt1}}(H)$	Massa combinada del τ_{had} i el ℓ_1 .	3.734
$\Delta\eta(\tau_{\text{had}}, \ell_1)$	$\Delta\eta$ entre el τ_{had} i el ℓ_1 .	2.457
$m_{\text{vis}}^{\text{opt2}}(H)$	Massa combinada del τ_{had} i el ℓ_1 .	2.025
$\Delta\eta(\tau_{\text{had}}, \ell_2)$	$\Delta\eta$ entre el τ_{had} i el leptó lleuger secundari.	1.864
$m_{\text{pred}}^{\text{opt1}}(t)$	Massa del quark top reconstruïda amb el jet etiquetat b , el ℓ_1 i el ν_{ℓ_1} .	1.596
$\Delta R(b\text{-tagged jet}, \ell_1)$	ΔR entre el jet etiquetat b i el ℓ_1 .	1.142
$m_{\text{pred}}^{\text{opt2}}(t)$	Massa del quark top reconstruïda amb el jet etiquetat b , el ℓ_2 i el ν_{ℓ_2} .	1.104
$\Delta R(b\text{-tagged jet}, \ell_2)$	ΔR entre el jet etiquetat b i el leptó lleuger secundari.	1.009
$\Delta\eta(\text{closest } b\text{-tagged jet}, \ell_1)$	$\Delta\eta$ entre el ℓ_1 i el jet etiquetat b més pròxim a aquest leptó.	0.740

Taula 6: Variables utilitzades a l'entrenament de la BDT^{OrigenLep}. El poder de separació ($\langle S^2 \rangle$) es presenta a la tercera columna.

3. Optimització d'hiperparàmetres del model: L'optimització dels hiperparàmetres d'entrenament és un element clau per a augmentar l'eficàcia del model. En el cas de la BDT^{OrigenLep}, els paràmetres seleccionats es mostren a la Taula 7. Aquests valors òptims s'han determinat mitjançant una cerca en retícula. En referència a la decisió sobre com tractar els pesos negatius (excloure'ls o incloure'ls en l'entrenament), els tests indiquen que un entrenar només pesos positius ofereix un millor rendiment. Els histogrames a la Figura 9 mostren que les distribucions són similars utilitzant tots els esdeveniments o només els esdeveniments amb pesos positius, la qual cosa indica que excloure els pesos negatius de l'entrenament no introduceix cap biaix en el model final.

Hiperparàmetre	Valor	Significat
Type	Gradient	Algoritme per al <i>boosting</i> per als arbres.
MaxDepth	3	Profunditat màxima de l'arbre.
Shrinkage	0.2	Taxa d'aprenentatge per l'algoritme.
NTrees	10^3	Nombre d'arbres en el mètode.
nCuts	40	Nombre de punts a testejar per trobar el tall òptim en la divisió del node.
NegWeights	Ignore neg	Ús dels pesos negatius a l'entrenament.

Taula 7: Valors dels hiperparàmetres utilitzats per gestionar l'entrenament de la BDT^{OrigenLep}. La resta d'hiperparàmetres utilitzen el seu valors per defecte.

4. Entrenament del model: L'entrenament supervisat del model per classificar esdeveniments segons l'origen del leptó lleuger crea un discriminant que

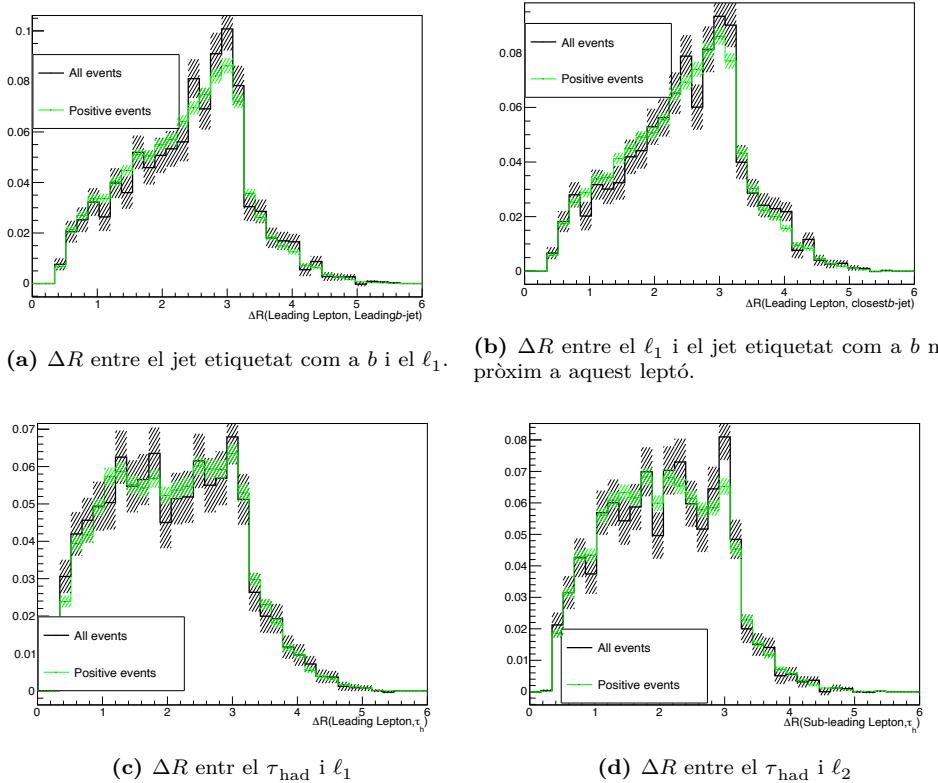


Figura 9: Distribucions normalitzades de variables cinemàtiques utilitzades a la BDT^{OrigenLep} a l’espai de fases corresponent al canal $2\ell \text{SS} + 1\tau_{\text{had}}$. En negre es mostren tots els esdeveniments i en verd només esdeveniments amb pesos positius. Per a cada bin, la banda d’error es calcula com l’arrel quadrada de la suma quadràtica dels pesos.

es presenta a la Figura 10. Per tal de evaluar la capacitat predictiva d’un model estadístic, la validació creuada K -fold és utilitzada amb cinc *folds*.

5. **Injecció del model:** L’últim pas consisteix en l’aplicació del model sobre les dades i en la determinació del llindar de classificació òptim. Una cerca lineal troba que $\text{BDT}^{\text{OrigenLep}} = -0.315$ és el punt òptim per discriminar esdeveniments de Tipus 1 i Tipus 2.

Una vegada establert l’origen del leptó, en comptes de fer servir ℓ_1 i ℓ_2 , les variables per a la discriminació entre senyal i fons fan servir ℓ_{top} i ℓ_{Higgs} . Definir els leptons segons el seu origen, en lloc de la seua p_T , permet crear variables més eficaces per a la discriminació. A la Taula 8, la tècnica basada en la BDT^{OrigenLep} es compara amb un mètode alternatiu que consisteix a aplicar dos requisits cinemàtics. Com es pot apreciar, l’eficàcia de classificació del mètode descrit ací és superior.

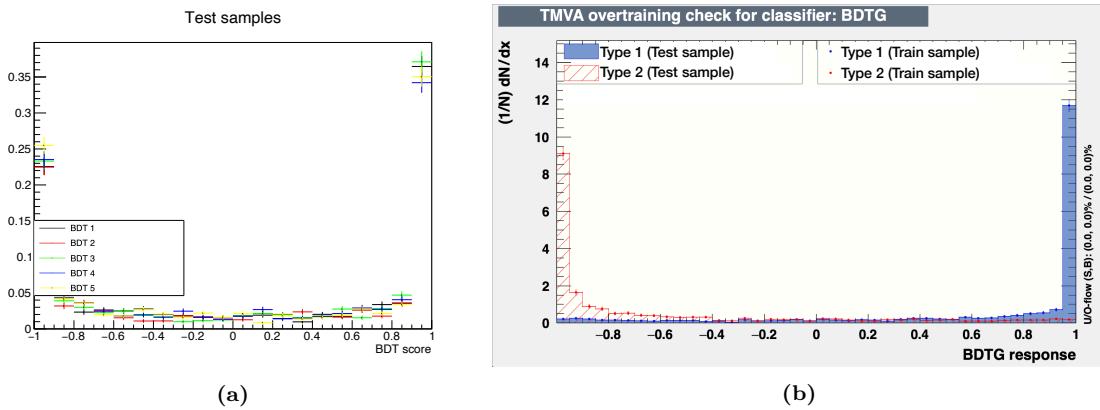


Figura 10: Distribució de la $BDT^{OrigenLep}$. En la figura (a) els cinc models del K -folding són comparats a les mostres de test. En (b) el model final és presentat comparant la distribució d'esdeveniments de Tipus 1 (blau) i de Tipus 2 (roig).

Precisió de l'assignació de l'origen del lepton lleuger			
	Total (100%)	$H \rightarrow \tau\tau$ (83.08%)	$H \rightarrow WW$ (16.92%)
$BDT^{OrigenLep}$	88.39	88.44	88.18
Mètode basat en talls	83.86	84.24	81.80

Taula 8: Precisió calculada comparant la predicció del mètode amb el valor real. El valor real s'ha obtingut utilitzant l'etiquetatge amb l'informació a nivell de partons. Aquest etiquetatge està disponible per a $H \rightarrow \tau\tau$ i $H \rightarrow WW^*$ però no per $H \rightarrow ZZ^*$. Només s'han utilitzat esdeveniments on hi havia una correspondència entre els leptons a nivell de reconstrucció i a nivell de partons.

4.1.3 Reconstrucció del sistema

La reconstrucció precisa del moment i la massa del quark top i del bosó de Higgs és una tasca complicada a causa de la presència de, almenys, quatre neutrins en l'estat final. En aquest treball, l'estrategia utilitzada per a reconstruir completament el top i el bosó de Higgs consisteix primer a reconstruir el sistema del quark top i després la cadena del bosó de Higgs. Com que coneixem el E_T^{miss} total, si l'energia mancant és reconstruïda per al quark top, es pot fer servir:

$$\vec{E}_T^{\text{miss}}(\text{total}) = \vec{E}_T^{\text{miss}}(\text{Higgs}) + \vec{E}_T^{\text{miss}}(\text{top})$$

per a obtindre \vec{E}_T^{miss} del sistema del Higgs.

Llavors, el primer pas és reconstruït el quark top. Com que al seu sistema de desintegració hi ha un neutrí (ν_{top}) com a mínim, cal estudiar la hipòtesi per a aconseguir $\vec{p}(\nu_{\text{top}})$. Les relacions aconseguides a través d'un ajust per a conéixer

$\vec{E}_{\text{T}}^{\text{miss}}$ (top) són:

$$p_{\text{T},\nu_{\text{top}}}^{\text{miss}} = \frac{1615.18 \text{ GeV}^2}{p_{\text{T},\ell_{\text{top}}}},$$

$$\phi_{\nu_{\text{top}}}^{\text{miss}} = \phi_{\ell_{\text{top}}} \pm \frac{\pi}{2},$$

sent $p_{\text{T},\ell_{\text{top}}}$ i $\phi_{\ell_{\text{top}}}$ el moment transvers i l'angle azimutal del ℓ_{top} .

Una vegada obtingut l'energia transversa mancant del sistema del quark top s'utilitza el mètode de la Calculadora de Massa Perduda (MMC), originalment desenvolupat per reconstruir la massa i moment del bosó de Higgs a l'anàlisi $H \rightarrow \tau\tau$ [290].

4.2 Estimació del fons

En la nostra anàlisi, diferenciem els esdeveniments de fons en dos tipus: “reductible” i “irreductible”.

- **Fons irreductibles:** Es refereixen a processos amb estats finals idèntics al senyal. Aquesta mena de fons no es poden distingir dels processos tHq a través de la multiplicitat d'objectes físics reconstruïts. En la nostra anàlisi són Diboson (VV), tW , $t\bar{t}Z$, $t\bar{t}H$, $t\bar{t}W$ i tZq .
- **Fons reductibles:** Originats per errors de reconstrucció, on objectes són incorrectament identificats com a part del senyal. Al canal $2\ell + 1\tau_{\text{had}}$, predominen els esdeveniments identificats consistents en jets erròniament identificats com τ_{had} . Això genera els fons de $t\bar{t}$ i $Z + \text{jets}$. També inclouen processos amb VV i certes produccions de $t\bar{t}X$.

Conèixer amb precisió la fracció esperada de cada procés de fons és fonamental per a obtenir resultats de qualitat. Mentre que per als fons irreductibles, el nombre d'esdeveniments s'estima amb precisió en les simulacions basades en MC, per als fons reductibles no és trivial determinar la seu abundància. Entre els enfocaments per a l'estimació de les taxes de mal identificació, destaca el *mètode d'ajust de plantilla*. Aquesta estratègia basada en dades per a estimar l'abundància dels diferents fons reductibles implica l'aparellament d'objectes reconstruïts en conjunts de dades simulats amb els corresponents objectes a nivell de partó simulat utilitzant cons de $\Delta R = 0.2$. Fent això, es creen les plantilles per a objectes físics ($e, \mu, \tau_{\text{had}}$, jet iniciat per quark, jet iniciat per gluó i desconegut). Una plantilla és un grup de partícules amb la mateixa etiqueta a nivell de partons. Després es realitza un ajust

del MC a les dades per a derivar els factors que reescalaran la mostra de MC. Aquest mètode actua com a nominal. Alternativament, el mètode de recompte s'utilitza per a assignar una incertesa al mètode d'ajust de plantilla. La composició a nivell de partó dels objectes reconstruïts es presenta a la Figura 11. Allí es pot veure que mentre els objectes en el procés tHq típicament estan ben reconstruïts, el τ_{had} reconstruït dels fons no és realment un τ en la majoria dels casos.

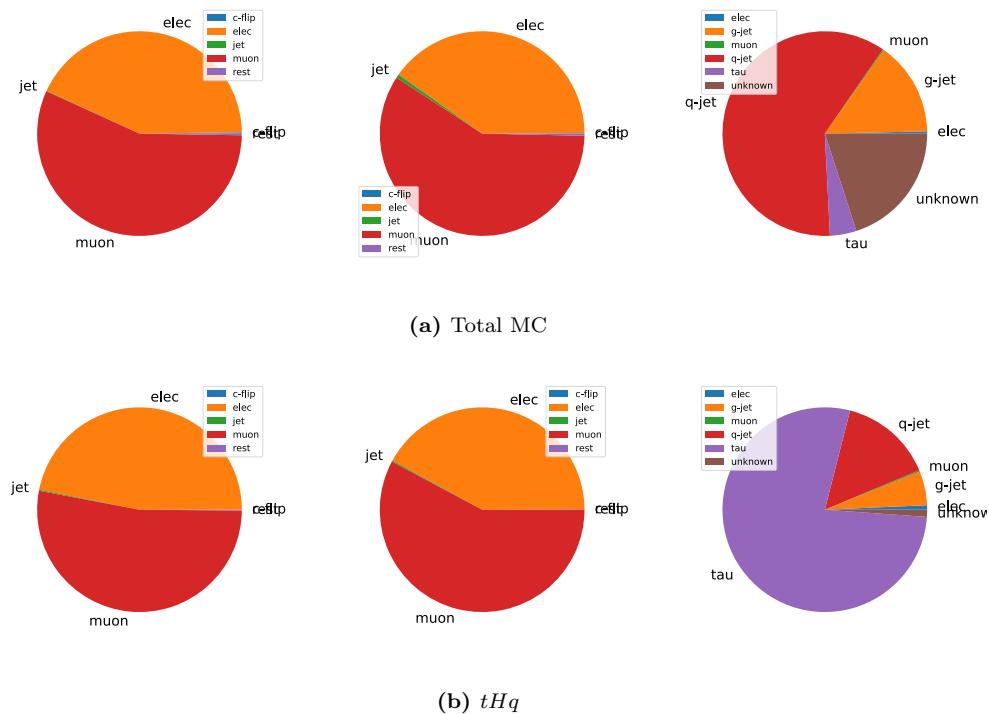


Figura 11: Composition at parton level of the reconstructed leading (left) and sub-leading (middle) leptons, as well as τ_{had} (right) passing the selection cuts in the $2\ell + 1\tau_{\text{had}}$ analysis with oppositely charged light leptons. The results are calculated using all MC samples and tHq production.

4.3 Selecció d'esdeveniments

La selecció d'esdeveniments consisteix en l'aplicació d'un conjunt de condicions que ens permeten separar el senyal del fons. D'aquesta manera, definim una regió de l'espai de fases enriquida amb esdeveniments de senyal. Aquesta regió es denomina regió de senyal (SR).

El punt de partida és la regió de preselecció (PR), on els objectes físics són seleccionats segons l'acceptació del detector. Els requisits de PR estan resumits a

la Taula 10 i el nombre d'esdeveniments a la regió de PR es presenta a la Taula 10 tant per a tot $2\ell + 1\tau_{\text{had}}$ com per als dos canals que el componen.

A continuació, es defineixen variables discriminants i s'utilitzen com a característiques d'entrada per a un BDT que distingí entre esdeveniments de senyal i fons, creant un discriminant.

Objecte	Multiplicitat	Moment lineal	Pseudorapidesa
Leptons lleugers	$n(e/\mu) = 2$	$p_T(e) > 14 \text{ GeV}$ $p_T(\mu) > 14 \text{ GeV}$	$ \eta(e) < 2.47, \eta(e) \notin [1.37, 1.52]$ $ \eta(\mu) < 2.50$
au hadrònic	$n(\tau_{\text{had}}) = 1$	$p_T(\tau_{\text{had}}) > 20 \text{ GeV}$	$ \eta(\tau_{\text{had}}) < 2.50, \eta(\tau_{\text{had}}) \notin [1.37, 1.52]$
Jets	$n(\text{jet}) = [2, 6]$	$p_T(\text{jet}) > 20 \text{ GeV}$	$ \eta(\text{jet}) < 4.5$
Jets etiquetats b	$n(b\text{-jet}) = [1, 2]$	$p_T(b\text{-jet}) > 20 \text{ GeV}$	$ \eta(b\text{-jet}) < 2.5$
E_T^{miss}		$p_T(E_T^{\text{miss}}) \in [5, 800] \text{ GeV}$	

Taula 9: Requeriments de preselecció. Per als leptons principal i secundari s'aplica un tall addicional en p_T : $p_T(\ell_1) > 27, \text{GeV}$ i $p_T(\ell_2) > 20, \text{GeV}$. El signe relatiu de la càrrega elèctrica dels leptons lleugers també s'utilitza en la preselecció per separar els canals $2\ell \text{SS} + 1\tau_{\text{had}}$ i $2\ell \text{OS} + 1\tau_{\text{had}}$.

Finalment, aplicant requisits sobre les sortides del BDT, es defineix la SR. També es poden definir regions de l'espai de fases enriquides amb processos d'un fons específic. Aquestes es coneixen com a regions de control (CR) o de validació (VR). La diferència entre CR i VR és que, mentre les primeres són considerades en els càlculs de l'ajust, les segones no ho són. Les SR, CRs i VRs són totes ortogonals entre si i són subespais de la PR.

Per a distingir el procés de senyal tHq dels fons i crear regions dedicades als fons, es construeixen diversos BDTs de gradient per a la classificació binària:

- BDT($tHq|_{\text{OS}}$): Entrenat per a discriminar el procés tHq de tots els fons simultàniament en el canal $2\ell \text{OS} + 1\tau_{\text{had}}$. Aquest BDT es fa servir després per a definir la SR per a tHq en aquest canal.
- BDT($t\bar{t}|_{\text{OS}}$) Entrenat amb l'objectiu de separar $t\bar{t}$ i $Z + \text{jets}$ en el canal $2\ell \text{OS} + 1\tau_{\text{had}}$. S'utilitza per a definir les VRs per a aquests dos processos.
- BDT($tHq|_{\text{SS}}$): Entrenat per al canal $2\ell \text{SS} + 1\tau_{\text{had}}$ per a discriminar el procés tHq de la resta de processos.

Igual que per a l'assignació de l'origen del leptons lleugers, els passos per a entrenar aquests models són: l'elecció de variables, l'optimització d'hiperparàmetres, l'entrenament del propi model i la injecció de la puntuació del model. Ací també es fa servir k -folding amb $k = 5$ per a la validació creuada.

Procés	$2\ell + 1\tau_{\text{had}}$	$2\ell \text{ OS} + 1\tau_{\text{had}}$	$2\ell \text{ SS} + 1\tau_{\text{had}}$
tHq	3.2 ± 0.8	1.9 ± 0.5	1.22 ± 0.34
tWH	3.4 ± 2.5	2.4 ± 1.8	1.0 ± 0.7
$t\bar{t}$	5710 ± 350	5690 ± 350	25 ± 10
$Z + \text{jets}$	3800 ± 1400	3800 ± 1400	0.7 ± 0.4
$t\bar{t}W$	130 ± 330	90 ± 230	40 ± 100
$t\bar{t}H$	86 ± 29	61 ± 20	25 ± 9
$t\bar{t}Z$	160 ± 29	139 ± 26	21 ± 4
tWZ	22 ± 16	19 ± 14	2.6 ± 2.1
tZq	45 ± 7	40 ± 6	5.4 ± 0.9
tW	260 ± 40	260 ± 40	1.2 ± 1.0
Diboson	190 ± 130	180 ± 130	10 ± 6
Fons menors	16 ± 9	14 ± 8	1.8 ± 1.1
Fons total	10400 ± 1500	10300 ± 1500	140 ± 110
S/B (%)	0.031	0.018	0.89
Significància	0.031	0.019	0.104

Taula 10: Nombre d'esdeveniments al nivell de PR per al canal $2\ell + 1\tau_{\text{had}}$ i els seus dos subcanals. Les incerteses corresponen tant a l'error estadístic com a les incerteses sistemàtiques (descrites a la Secció 4.4).

Els esdeveniments simulats amb pesos negatius representen aproximadament un 30% del total. Per al $\text{BDT}(tHq|_{\text{OS}})$ i $\text{BDT}(t\bar{t}|_{\text{OS}})$, només s'usen els pesos positius. Per al $\text{BDT}(tHq|_{\text{SS}})$, s'usa el valor absolut del pes a l'entrenament. Aquesta elecció és preferida per l'escassetat d'esdeveniments en el canal $2\ell \text{ SS} + 1\tau_{\text{had}}$.

Pel que fa a la selecció de variables, aquestes són escollides per maximitzar el poder de separació de la BDT. Un algoritme iteratiu construeix la llista de variables utilitzades en cada BDT tenint en compte el valor que aporta cada variable al rendiment del model i les seues correlacions amb les altres variables. El resultat d'aquest mètode dona lloc a les tres llistes de variables presentades a la Figura 12. La distribució de la variable més important de cada llista es mostra a la Figura 13.

Una vegada escollides les variables, cal optimitzar els hiperparàmetres de l'entrenament. Aquests s'encarreguen de gestionar el procés d'aprenentatge del model, però no formen part d'ell i seleccionar els adequats és crucial, ja que pot influir significativament en el rendiment del model. L'optimització dels hiperparàmetres del BDT es fa utilitzant un algoritme genètic [294]. El conjunt òptim d'hiperparàmetres trobat es presenta per als tres BDTs simultàniament a la Taula 11.

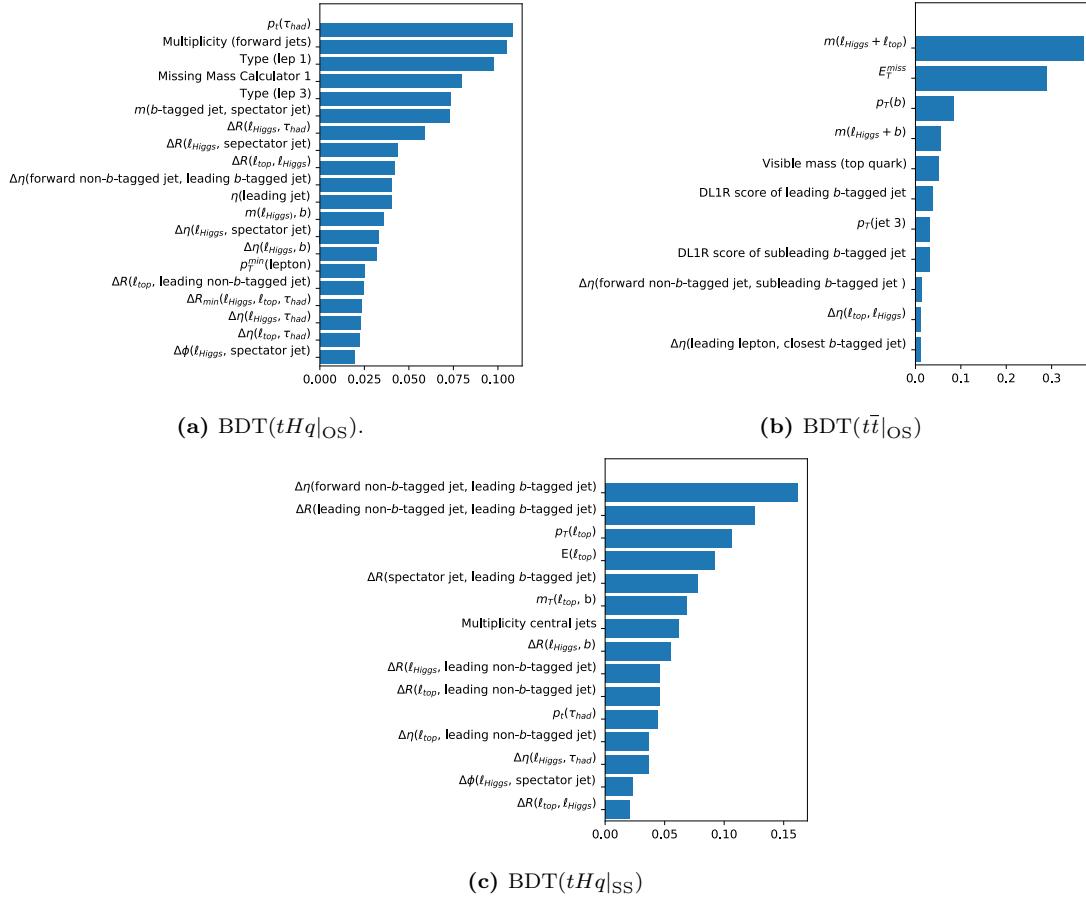


Figura 12: Rànquing de variables per a cada BDT: a) BDT($tHq|_{\text{OS}}$), (b) BDT($t\bar{t}|_{\text{OS}}$) i (c) BDT($tHq|_{\text{SS}}$). Els rànquings s'han obtingut amb l'eina d'importància de característiques d'XGBoost. L'eix x correspon al valor que XGBoost dóna per a avaluar la precisió aportada per la variable al BDT.

El rendiment dels BDTs s'avalua a través de la corba de característiques operatives del receptor (ROC). Per a un model de classificació binària, la corba ROC representa la relació entre la taxa de veritables positius (sensibilitat) i la taxa de falsos positius (1 - especificitat) per a diferents llindars de decisió. Les corbes ROC dels tres BDTs de definició de regió es presenten a la Figura 14. L'àrea sota la corba ROC (AUC per les seues sigles en anglés) és un altre indicador del rendiment del model: un valor d'AUC proper a 1 indica un model molt precís, mentre que un valor d'AUC proper a 0,5 indica un rendiment no millor que l'atzar. Per als models BDT($tHq|_{\text{OS}}$), BDT($t\bar{t}|_{\text{OS}}$) i BDT($tHq|_{\text{SS}}$), les AUC són, respectivament 0.637 ± 0.004 , 0.730 ± 0.003 i 0.6706 ± 0.0013 .

Hiperparàmetre	BDT($tHq OS$)	BDT($t\bar{t} OS$)	BDT($tHq _{SS}$)
Profunditat màxima	4	4	4
Taxa d'aprenentatge	0.1237	0.0334	0.04
Nombre d'estimadors	1500	1500	1500
Pes mínim del fill	0.52	0.077	0.026
Escala de pesos positius	268.838	0.36	83.21
Estratègia de pes neg.	Només positius	Només positius	Valors absoluts

Taula 11: Configuració dels hiperparàmetres utilitzats per a gestionar l'entrenament dels tres BDTs de gradient emprats per a la definició de regions. La resta d'hiperparàmetres es configuren amb els seus valors per defecte.

La distribució de les BDTs en funció del *fold* es mostra a la Figura 15. En aquesta figura també es mostren les fraccions entre els esdeveniments de cada *fold* i es pot apreciar clarament com tots cinc són perfectament compatibles entre si. Això vol dir que el model té bona capacitat per a generalitzar.

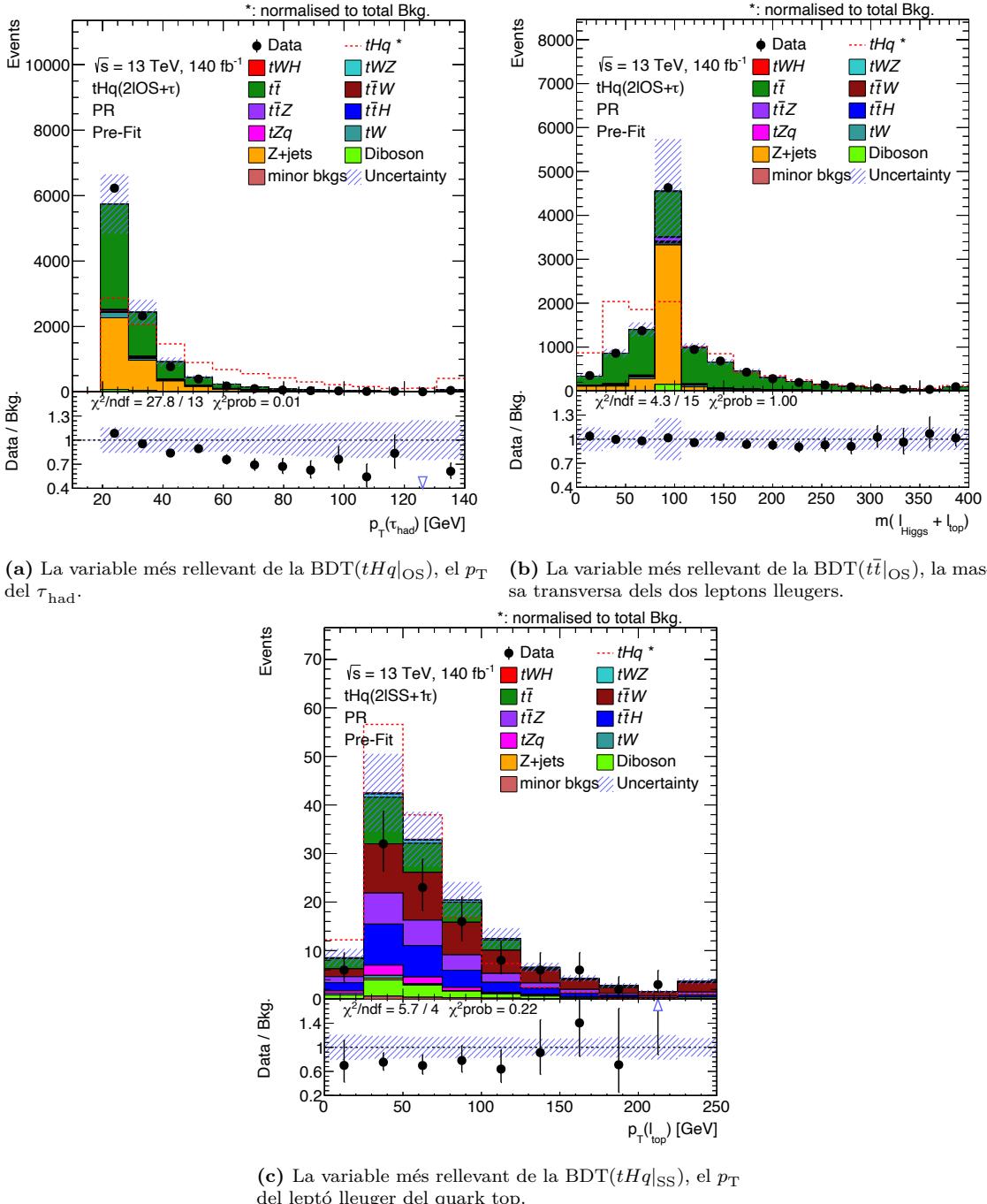


Figura 13: Variables més rellevants per cadascun dels tres models de classificació de processos: (a) BDT($tHq|_{\text{OS}}$), (b) BDT($t\bar{t}|_{\text{OS}}$) i (c) BDT($tHq|_{\text{SS}}$). Les bandes d'incertesa inclouen les incerteses estadístiques i sistemàtiques i el panell inferior presenta la relació entre les dades recollides i la simulació MC. Addicionalment, el χ^2 mesura la concordança entre les dades reals i la mostra d'esdeveniments simulats per MC.

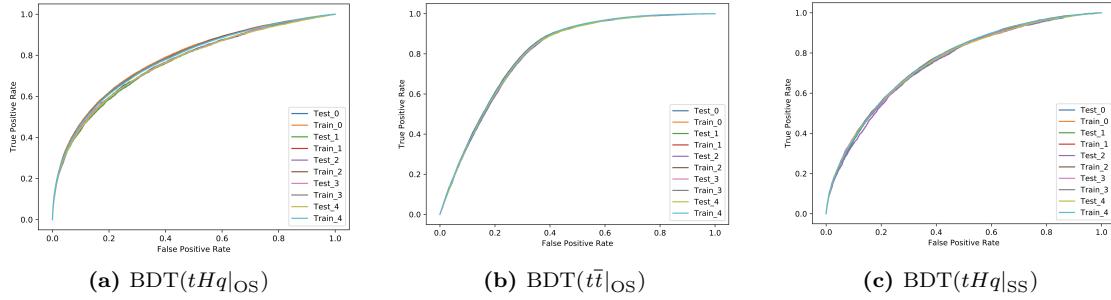


Figura 14: ROC per a tots els grups dels tres models BDT. En cada figura es poden veure cinc parelles de corbes ROC. Cada parella correspon a un grup i, dins d'un grup, la parella tracta les mostres d'entrenament i de prova separatadament.

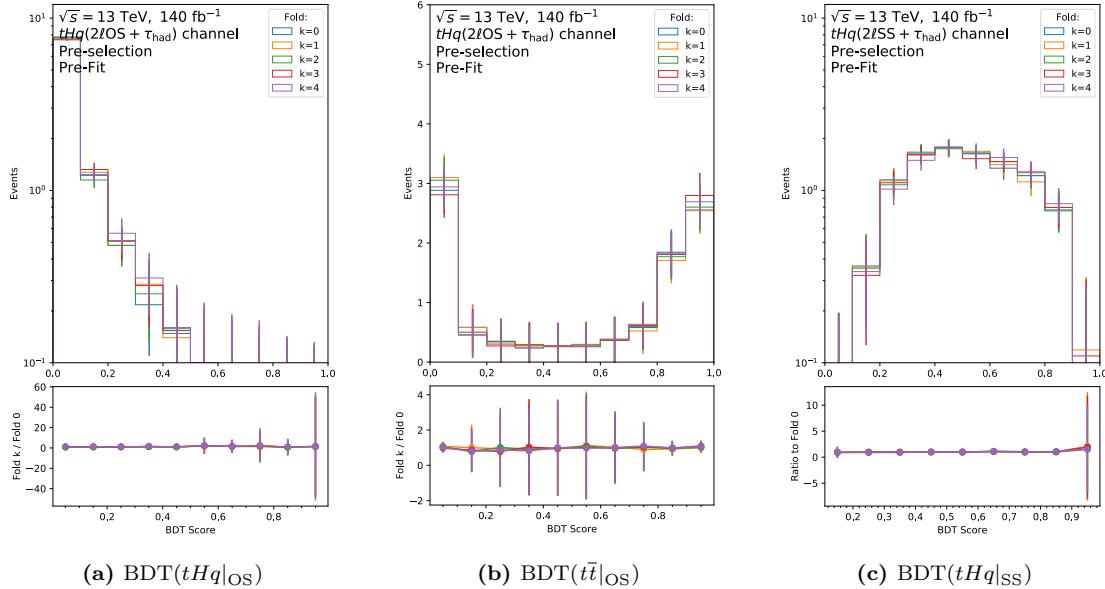


Figura 15: En el panell superior de cada figura, es mostra el perfil del BDT per als cinc *folds* simultàniament. L'error en cada bin del panell superior correspon a la desviació estàndard. En el panell inferior, es presenta la la fracció d'esdeveniments entre el primer grup (Grup 0) i tots els altres. Ambdós panells comparteixen el mateix eix x. Per a tots els *folds* dins d'un BDT, les distribucions són compatibles dins de l'error estadístic. Això indica que el model generalitza adequadament i no es veu afectat pel conjunt específic d'esdeveniments que s'utilitzen per a l'entrenament en cada grup. Els perfils mostrats són dibuixats emprant només les mostres de test.

4.4 Fonts d'incertesa

La incertesa en física es refereix al fet que és impossible mesurar qualsevol quantitat física amb precisió perfecta. Això es deu al fet que tots els instruments de mesura tenen limitacions i estan subjectes a diverses fonts d'error. Aquestes incerteses proporcionen una mesura de l'interval dins del qual es preveu que es trobe el valor real d'una quantitat mesurada. Hi ha dos tipus principals d'incerteses: la incertesa estadística i la incertesa sistemàtica.

La incertesa estadística sorgeix de l'atzar inherent o fluctuacions en les dades finites. Aquestes fluctuacions són típicament descrites per mètodes estadístics, com les distribucions de probabilitat, i són quantificades per mesures estadístiques com la desviació estàndard o intervals de confiança. Les incerteses estadístiques són completament no correlacionades entre mesures successives, la qual cosa significa que cada mesura porta la seua pròpia incertesa estadística independent.

D'altra banda, les incerteses sistemàtiques engloben totes les fonts d'error o variació que no són directament degudes a l'estadística de les dades. Les incerteses sistemàtiques poden ocórrer en qualsevol punt de la cadena d'anàlisi. Estan associades amb diversos factors, incloent-hi l'aparell de mesura, les condicions experimentals, les suposicions fetes en l'anàlisi, els models teòrics emprats, la reconstrucció d'objectes, les tècniques d'estimació del fons i les simulacions MC, entre molts altres. Les incerteses sistemàtiques són completament correlacionades entre mesures successives, la qual cosa significa que afecten consistentment tot el conjunt de dades.

Les incerteses sistemàtiques es classifiquen en dos grans grups depenen del seu origen: incerteses teòriques i incerteses experimentals.

4.4.1 Incerteses teòriques

Aquestes es refereixen a la modelització a simulació de MC i el coneixement teòric sobre els diferents processos. Per a quantificar aquestes incerteses, es comparen diferents prediccions teòriques i variacions en els paràmetres del model de MC per a avaluar el seu impacte en els resultats. Les variacions són comparades amb el mètode principal (referit com a *nominal*), la qual cosa, permet estimar l'exactitud i la fiabilitat dels càlculs teòrics.

Les incerteses de modelatge s'avaluen de tres maneres: comparant la predicció del generador de MC nominal amb la de generadors alternatiu, variant els paràmetres interns de la simulació nominal o variant la secció transversal predicta dins de la incertesa teòrica.

4.4.2 Incerteses experimentals

Les incerteses experimentals juguen un paper crucial en els experiments de física de partícules, ja que sorgeixen del mateix procés de mesura. En el context de les ànalisis de ATLAS, aquestes incerteses provenen principalment de factors relacionats amb el detector i abasten diversos aspectes, com les limitacions de l'aparell de mesura, els procediments de calibració i l'eficàcia de reconstruir objectes físics dins del detector. Algunes fonts comunes d'incerteses experimentals inclouen:

- **Luminositat:** Per a cada any de presa de dades, hi ha una incertesa en la luminositat integrada recollida pel detector ATLAS.
- **Reponderació de Pile-up:** Els esdeveniments de les mostres de simulació MC es repondren per a coincidir amb la distribució observada del nombre mitjà d'interaccions per creuament de feixos en les dades [300]. El valor d'aquesta incertesa s'obté reescalant el valor de $\langle\mu\rangle$ en les dades per 1/0.99 i 1/1.07 al voltant del factor d'escala nominal de 1/1.03 [301].
- **Escala d'energia del jet:** La calibració de l'escala d'energia del jet (JES) corregix l'energia i la direcció dels jets per a coincidir amb la dels jets reconstruïts a nivell de partons [302]. La incertesa associada amb JES s'obté de dades de feix de prova, dades de col·lisions LHC i simulacions MC.
- **Resolució d'energia del jet:** La resolució d'energia del jet (JER) es refereix a la capacitat de l'experiment per a mesurar l'energia dels jets. La JER es mesura per separat per a dades i MC utilitzant les dues tècniques in-situ descrites en les referències [302, 303].
- **Jet vertex tagger:** Les incerteses associades amb l'identificador de vèrtexs provenen de la contaminació residual de jets de pile-up [272, 304].
- **τ_{had} mal identificat:** Les incerteses sistemàtiques associades amb les identificacions de τ_{had} depenen del punt de treball triat a la RNN que s'encarrega de la identificació dels τ_{had} .
- **Taxa de τ_{had} mal identificat:** Per a abordar la incertesa associada amb el mètode de determinació de les probabilitats de τ_{had} mal identificat, el mètode de l'ajust de plantilla es compara amb el mètode de recompte. Ambdós mètodes són descrits a la Secció 4.2.
- **Etiquetatge de sabors pesats i lleugers:** L'eficiència de l'algoritme d'etiquetatge de sabors es mesura per a cada sabor de jet utilitzant mostres de control de dades i simulació. Aquesta evaluació dóna factors de correcció per a ajustar les taxes d'etiquetatge en les simulacions.

- **Eficiència d'Electrons:** S'utilitza un mètode específic per corregir diferències en l'eficiència de reconstrucció d'electrons entre dades reals i simulades. Això es mesura en certes classes d'esdeveniments de partícules.
- **Eficiència de Muons:** De forma anàloga als electrons, es realitzen correccions per ajustar les eficiències d'identificació i aïllament dels muons, basades en diferents tipus d'esdeveniments.
- **Escala d'Energia i Resolució d'Electrons:** Es corregissen petites discrepàncies entre les simulacions i les dades reals respecte a l'energia i la resolució dels electrons.
- **Escala de Moment i Resolució de Muons:** Es fan ajustos en les simulacions dels muons per alinear-les amb les observacions reals, enfocant-se en la seua escala de moment i resolució.
- **Terme Suau de E_T^{miss} :** S'apliquen correccions a l'escala i la resolució d'una part específica de la mesura E_T^{miss} , que no està associada amb objectes calibrats.

4.5 Resultats

4.5.1 Mètode de l'ajust de màxima versemblança

La tècnica utilitzada per mesurar la normalització del senyal ($\mu_{tHq}^{2\ell+1\tau_{\text{had}}}$) i dels processos de fons que es desitgen controlar (k_p) és coneguda com a *mètode de l'ajust de màxima versemblança*, encara que és més habitual referir-se a aquesta pel seu nom en anglès *maximum likelihood fit*. El μ i k_p s'anomenen conjuntament com *paràmetres d'interés del sistema* (POI). El factor de normalització del senyal és també conegit com força del senyal i es defineix com:

$$\mu_{tHq} = \frac{\sigma^{\text{obs}}}{\sigma^{\text{pred}}} .$$

El mètode de màxima versemblança és àmpliament utilitzat als anàlisis d'ATLAS, hem d'estimar desenes o fins i tot centenars de fonts sistemàtiques d'incertesa. Aquestes incerteses són intruïdes a l'ajust de versemblança mitjançant uns paràmetres al model anomenats *paràmetres de molèstia* (NPs).

Aquest mètode funciona maximitzant la funció de versemblança per distribucions en diagrames de barres⁷:

$$\begin{aligned} L(\vec{n}|\mu, \vec{\theta}) &= \prod_{i \in \text{bins}} \mathcal{P}(n_i^{\text{obs}} | \mu \cdot s_i^{\text{exp}}(\vec{\theta}) + b_i^{\text{exp}}(\vec{\theta})) \\ &\times \prod_{j \in \text{syst}} \mathcal{G}(\theta_{0,j} | \theta_j, \Delta\theta_j) \times \prod_{s \in \gamma} \mathcal{P}(\theta_{0,s} | \theta_s, \Delta\theta_s). \end{aligned} \quad (1)$$

Ací, l'índex i es refereix al bins dels histogrames que s'utilitzen al fit. L'índex j corre sobre els NPs que es refereixen a les fonts d'incertesa sistemàtica i l'índex s als NPs associats amb la incertesa estadística de la mostra de simulacions de MC (anomenades *gammas* o γ). \mathcal{P} i \mathcal{G} són, respectivament, les funcions de les distribucions Poissonianes i Gaussianes. El nombre d'esdeveniments observat és n_i^{obs} i l'esperat n_i^{exp} , que es calcula:

$$n_i^{\text{exp}} = \mu \cdot s_i^{\text{exp}}(\vec{\theta}) + \sum_{p \in \text{Bkg}} k_p \cdot b_i^{\text{exp}}(\vec{\theta}),$$

on μ i k_p són les normalitzacions de la senyal i els diferents fons. A l'equació 1, els θ son els NPs, mentre que θ_0 i $\Delta\theta$ són el valor nominal del NP i la variació d'una desviació estàndard aplicada per valorar el seu impacte.

A més del factor de normalització del senyal, també es proporciona un límit superior mitjançant una prova estadística. La prova depén del valor de μ , i el límit superior s'estreu utilitzant el mètode descrit a la referència [308] per a establir un nivell de confiança del 95%. La interpretació de μ , així com el límit superior, també pot ser escrita en termes de la secció eficaç de producció del procés de senyal de la següent manera:

$$\sigma^{\text{obs}} = \mu(tHq) \times \sigma^{\text{pred}}, \quad (2)$$

on σ^{pred} és el valor de la secció de producció predit per la teoria.

4.5.2 Estratègia d'ajust

Experiment cec

A l'hora de formular l'estratègia de l'anàlisi és crucial evitar examinar les regions de dades que es preveu que tinguen una alta concentració d'esdeveniments de senyal tHq . Aquesta pràctica de cegar l'experiment s'anomena *blinding* i juga un paper fonamental a l'hora de protegir l'anàlisi de potencials biaix o predisposicions que podrien ocórrer si els investigadors es veuen influenciats per variacions estadístiques observades.

⁷Cada barra del diagrama de barres és anomenada *bin* i al diagrama en si ens referim com a *histograma*.

En aquesta recerca, el punt corresponent a les dades en un bin només es mostrat si la fracció de senyal esperada era menor que 0.3% i tots els bins en les SRs estaven completament cegats. Només en la Secció 4.5.4 la recerca es fa amb dades des-cegades.

Tipus d'ajust

Primerament, fem un ajust utilitzant únicament les mostres de simulacions de MC. En aquest ajust cap dada real és emprada. Aquesta configuració es coneix com a hipòtesis d'Asimov. Els resultats d'aquesta mena d'ajust es presenten a la Secció 4.5.3. El propòsit d'aquesta mena d'ajust és avaluar la sensitivitat de l'anàlisi i comprovar l'estabilitat de l'ajust amb la configuració escollida. Si l'ajust d'Asimov presenta cap mena d'inestabilitat podem canviar la definició de les regions fetes servir a l'ajust o els paràmetres que deixem flotar als càlculs.

Si l'ajust d'Asimov no funciona correctament, les dades reals són parcialment influïdes als càlculs. Primer a les regions de control, on fem un ajust per a mesurar la normalització del fons que desitgem controlar (k_p). Per fer això assumim la hipòtesis de senyal nul · la (és a dir, $\mu_{tHq}=0$). Aquesta configuració coneix com ajust de les CR fons únicament.

Després les dades s'utilitzen a totes les regions de l'anàlisi amb la hipòtesi nul · la i, de nou, es calculen el k_p . La diferència ací és que podem determinar l'impacte de la SR en la determinació dels factors de normalització. Encara que la SR s'utilitza en aquest ajust, a les gràfiques continua cegada per a no produir cap biaix. Per simplicitat, els resultats d'aquests dos tipus d'ajust no estan inclosos en el resum.

Finalment, si tots els ajustos preliminars funcionen correctament. L'ajust de màxima versemblança per a distribucions amb bins es realitza emprant totes les regions i tenint també en compte el MC del senyal. Aquest ajust final es du a terme amb dades des-cegades. Els resultats de l'ajust final es presenten a la Secció 4.5.4.

Poda de the NPs

En principi, introduïm un NP per a cada incertesa sistemàtica en la funció de versemblança. A causa del gran nombre de paràmetres d'ajust, aquesta tècnica és molt costosa computacionalment i pot trigar hores o dies per fer un sol ajust. El que ens preocupa és quines fonts sistemàtiques tenen una gran contribució a la incertesa dels POI. Si identifiquem les fonts d'incertesa que no tenen una contribució significativa, podem llevar-les de l'ajust i agilitzar el procés. Aquesta tècnica és referida com a *poda*.

Configuracions de l'ajust de màxima versemblança

En aquest cas, per configuracions, ens referim a quines regions de l'espai de fases incloem en l'ajust de màxima versemblança i a quins factors de normalització deixem flotar com a paràmetres lliures. Les diverses estratègies que sigut explorades es presenten a la Taula 12 per al canal $2\ell \text{OS} + 1\tau_{\text{had}}$ i a la Taula 13 per al $2\ell \text{SS} + 1\tau_{\text{had}}$.

	$t\bar{t}$ i $Z + \text{jets}$	Paràmetres lliures	Pesos corregits
Utilitzat	CR	$\mu_{tHq}, k_{t\bar{t}}, k_{Z+\text{jets}}$	✓
Alternativa 1	VR	μ_{tHq}	✓
Alternativa 2	CR	μ_{tHq}	✓
Alternativa 3	CR	$\mu_{tHq}, k_{t\bar{t}}, k_{Z+\text{jets}}$	✗

Taula 12: Diferents configuracions d'ajust de versemblança provades per al canal $2\ell \text{OS} + 1\tau_{\text{had}}$. La columna “ $t\bar{t}$ i $Z + \text{jets}$ ” indica si les regions de l'espai de fases dedicades a aquests dos processos són utilitzades als càlculs (CR) i si només s'utilitzen per a la validació de l'ajust (VR). Totes aquestes configuracions han estat explorades (amb dades cegues) per a comprovar l'impacte de cada opció als resultats.

	CRs	k_p lliures
Utilitzat	Tots els fons	$k_{t\bar{t}, t\bar{t}X}$
Alternativa 1	$t\bar{t}, t\bar{t}X$	$k_{t\bar{t}}, k_{t\bar{t}W}$
Alternativa 2	$t\bar{t}, t\bar{t}X$	$k_{t\bar{t}}, k_{t\bar{t}X}$
Alternativa 3	$t\bar{t}, t\bar{t}X$	$k_{t\bar{t}}, k_{t\bar{t}W}, k_{t\bar{t}H}, k_{t\bar{t}Z}$

Taula 13: Diferents configuracions de l'ajust provades per al canal $2\ell \text{SS} + 1\tau_{\text{had}}$. La columna CR fa referència a quines regions de control s'utilitzen a l'ajust i la columna k_p lliures es refereix a quins factors de normalització es calculen juntament amb la $\mu_{tHq}^{2\ell \text{SS} + 1\tau_{\text{had}}}$. Totes aquestes han estat explorades amb dades cegues per a comprovar com la configuració influencia els resultats.

Com es pot veure a l'última columna de la Taula 12, pel canal $2\ell \text{OS} + 1\tau_{\text{had}}$ s'ha explorat no utilitzar els factors d'escala que corregeixen les imprecisions en l'estimació del fons a causa d'objectes físics erròniament identificats⁸. La idea consisteix en controlar la freqüència d'aquesta mena de processos fent ús dels factors de normalització $k_{t\bar{t}}$ i $k_{Z+\text{jets}}$. Els resultats preliminars amb aquesta opció no eren massa prometedors. Finalment, s'ha fet servir la configuració en la que $k_{t\bar{t}}$ i $k_{Z+\text{jets}}$ són calculats. Encara que es podria pensar que no cal controlar $t\bar{t}$ i $Z + \text{jets}$ perque ja disposem dels factors d'escala que corregissen⁹ la quantitat d'esdeveniments de

⁸Principalment quarks o gluons mal identificats com a τ_{had} .

⁹En aquest context corregir significa fer que el MC represente correctament la física d'ATLAS, incloent-hi el nombre d'esdeveniments de processos mal identificats.

MC que tenim d'aquest tipus, la realitat és que l'ajust millora quan $k_{t\bar{t}}$ i $k_{Z+\text{jets}}$ són paràmetres lliures del sistema.

Pel que respecta al canal $2\ell \text{SS} + 1\tau_{\text{had}}$, el principal problema que hi ha és que hi ha un excés generalitzat d'esdeveniments de simulats amb MC respecte a les dades recollides i això provoca que l'ajust siga prou inestable quan tracta de controlar els processos de fons. Inicialment, s'ha tractat de controlar $k_{t\bar{t}}$ i $k_{t\bar{t}W}$ fent servir una regió dedicada per a $t\bar{t}$ i una altra per a $t\bar{t}X$ ¹⁰. El motiu per tractar de fitar $t\bar{t}W$ en comptes de $t\bar{t}X$ és que hi ha una lleugera tensió entre les prediccions teòriques sobre la producció de $t\bar{t}W$ i els resultats experimentals. Tant ATLAS com CMS han mostrat certa sobreabundància en les dades en comparació amb les prediccions teòriques [309, 310]. En qualsevol cas, ajustar únicament $t\bar{t}W$ i $t\bar{t}$ no és suficient per a corregir el modelatge. Si tractem de controlar els $t\bar{t}$, $t\bar{t}W$, $t\bar{t}Z$ i $t\bar{t}H$ de forma independent, el fit resulta molt inestable perquè són massa paràmetres lliures per a una mostra estadística insuficient. Llavors, els dos plantejaments amb millors resultats per al canal $2\ell \text{SS} + 1\tau_{\text{had}}$ són o bé ajustar $k_{t\bar{t}}$ i $k_{t\bar{t}X}$ en les dues regions que tenim, o bé juntar tots els fons en una sola regió ortogonal al SR i obtindre un únic factor de normalització, $k_{t\bar{t}, t\bar{t}X}$.

4.5.3 Ajust d'Asimov

L'ajust Asimov consisteix a utilitzar només els esdeveniments generats per MC com si aquests fossin les dades recollides. Les dades Asimov es construeixen com a conjunts de dades distribuïdes en bins, en els quals el n_i^{obs} de cada bin es fixa a n_i^{exp} . Sota la hipòtesi Asimov, la força del senyal en l'Equació 1 es fixa a la unitat. En aquesta mateixa expressió, els diferents NPs es fixen als seus valors nominals ($\theta = \theta^0$).

A la Figura 16 es mostren les correlacions entre les NPs, le factors de normalització i la μ_{tHq} .

A la Taula 14 descomposem l'incertesa total en la mesura de μ_{tHq} per grups d'incerteses sistemàtics sota la hipòtesi d'Asimov. La influència de cada NP en la determinació μ_{tHq} es presenta a la Figura 17.

La sensibilitat esperada a la força del senyal i als factors de normalització en el canal $2\ell \text{OS} + 1\tau_{\text{had}}$ és:

$$\Delta\mu_{tHq} = {}^{+31.52}_{-32.66} (\text{tot.}) {}^{+15.87}_{-14.85} (\text{stat.}) \quad (3)$$

$$\Delta k_{t\bar{t}} = \pm 0.04 (\text{tot.}) \pm 0.02 (\text{stat.}) \quad (4)$$

$$\Delta k_{Z+\text{jets}} = \pm 0.10 (\text{tot.}) \pm 0.02 (\text{stat.}) \quad (5)$$

¹⁰Cal notar que és complicat discriminar entre $t\bar{t}W$, $t\bar{t}Z$ i $t\bar{t}H$.

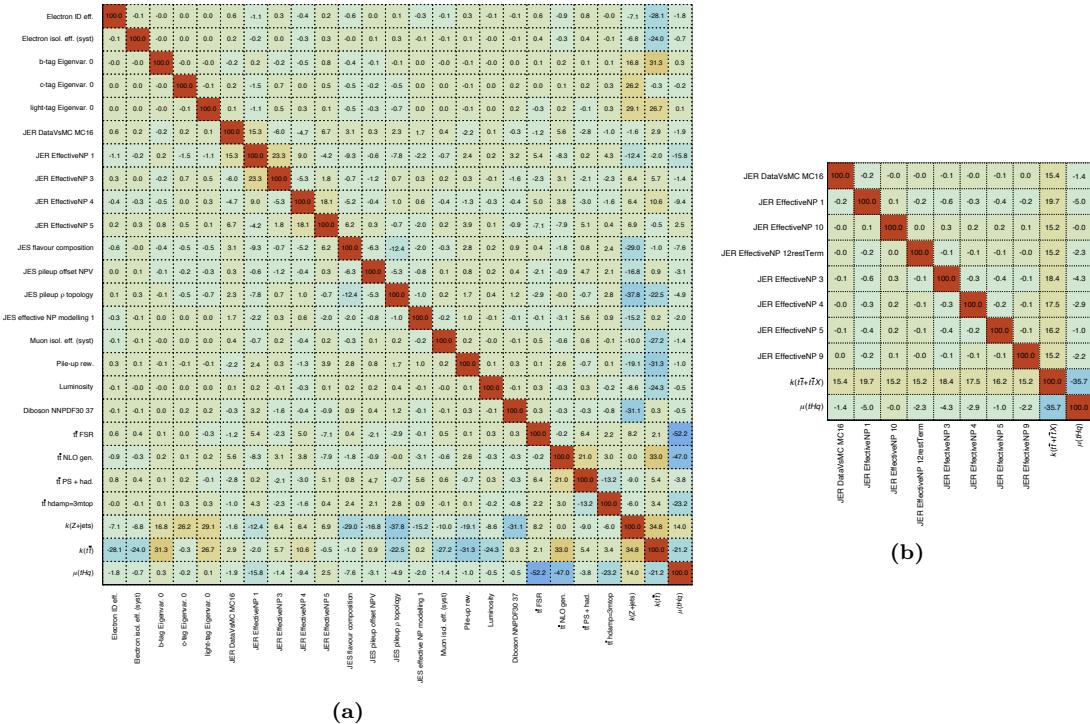


Figura 16: Correlació entre els diferents NP's i els POIs en el canals (a) $2\ell \text{OS} + 1\tau_{\text{had}}$ i (b) $2\ell \text{SS} + 1\tau_{\text{had}}$ la sota la hipòtesi d'Asimov. Només es mostren els NP's amb almenys una correlació superior al 15%.

L'incertesa total (tot.) inclou efectes estadístics i sistemàtics. També es mostra separadament l'incertesa estadística (stat.). Tant les incerteses estadístiques com les sistemàtiques són comparables (± 14.39 per a l'estadística i ± 15.65 per a la sistemàtica).

La sensibilitat esperada a la força del senyal de tHq i al factor de normalització $k_{t\bar{t}, t\bar{t}X}$ en el canal $2\ell \text{SS} + 1\tau_{\text{had}}$ és:

$$\Delta\mu_{tHq} = \pm 6.51(\text{tot.}) \pm 6.06(\text{stat.}) \quad (6)$$

$$\Delta k_{t\bar{t}, t\bar{t}X} = \pm 0.16(\text{tot.}) \pm 0.12(\text{stat.}) . \quad (7)$$

L'error estadístic és el component dominant en la incertesa total de $\mu^{2\ell \text{SS} + 1\tau_{\text{had}}} tHq$. La sensibilitat a la producció de tHq millora significativament respecte al canal $2\ell \text{OS} + 1\tau_{\text{had}}$ (Equació 3). Quant a la incertesa en el fons, és major que en el canal $2\ell \text{OS} + 1\tau_{\text{had}}$. Això es deu parcialment al fet que la contribució del fons es redueix notablement en el canal $2\ell \text{SS} + 1\tau_{\text{had}}$. Tot i això, la major part de la incertesa sobre $k t\bar{t}, t\bar{t}X$ es deu a les contribucions sistemàtiques.

Font d'incertesa	$2\ell \text{ OS} + 1\tau_{\text{had}}$	$2\ell \text{ SS} + 1\tau_{\text{had}}$
Incertesa del MC	± 7.845	± 1.376
Modelatge		
Incerteses teòriques	± 11.917	± 1.262
Incerteses a les PDFs	± 7.895	± 0.837
Experimental		
Instrumental	± 4.917	± 0.793
Instrumental: etiquetatge de sabor	± 0.439	± 0.163
Instrumental JES i JER	± 10.851	± 2.877
Factors de normalització	± 10.493	± 2.725
Total d'incertesa sistemàtica	± 28.511	± 4.614

Taula 14: Incerteses sistemàtiques en la mesura de μ_{tHq} en els canals $2\ell + 1\tau_{\text{had}}$ sota la hipòtesi d'Asimov.

4.5.4 Adjust amb totes les dades

Un cop explorats els resultats esperats utilitzant només simulacions MC en l'ajust Asimov (vegeu la Secció 4.5.3), el següent pas en l'anàlisi és incorporar les dades observacionals reals. En el cas de l'ajust a les dades, no s'apliquen condicions ad-hoc a la força del senyal, els factors de normalització o els NPs. Per tant, en contrast amb l'ajust Asimov, els valors de μ_{tHq} i els k_p poden divergir d'una, i els θ s poden ser diferents dels seus θ^0 s (és a dir, desviació o "pull"). La significància dels pulls es presenten a Figura 18.

A la Taula 15 descomposem l'incertesa total en la mesura de μ_{tHq} per grups d'incerteses sistemàtiques utilitzant totes les dades.

El resultat de l'ajust perfil-semblança amb compartiments en el canal $2\ell \text{ OS} + 1\tau_{\text{had}}$ produeix la següent força del senyal i factors de normalització:

$$\mu_{tHq} = -22.11^{+29.22}_{-34.08} (\text{tot.})^{+14.76}_{-13.72} (\text{stat.}) \quad (8)$$

$$k_{t\bar{t}} = 0.97 \pm 0.03 (\text{tot.}) \pm 0.02 (\text{stat.}) \quad (9)$$

$$k_{Z+\text{jets}} = 1.03 \pm 0.11 (\text{tot.}) \pm 0.02 (\text{stat.}) \quad (10)$$

L'incertesa total (tot.) inclou efectes estadístics i sistemàtics. L'incertesa estadística (stat.) també es mostra per separat. Els resultats obtinguts són compatibles amb el Model Estàndard. Mentre que les normalitzacions de $t\bar{t}$ i $Z + \text{jets}$ gairebé no són escalades, estant prop de la predicció del SM, el procés de tHq és escalat negativament per un factor de -22.11 . La gran incertesa estadística en el resultat

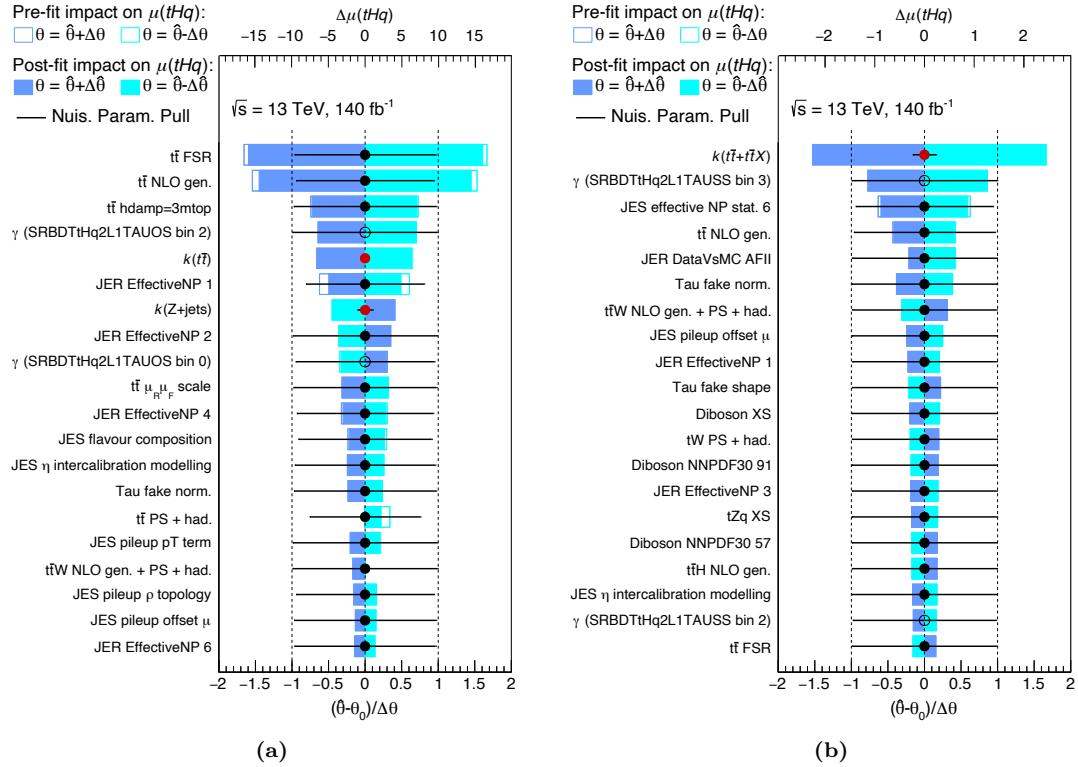


Figura 17: Rànquing dels NPs més impactants en l'ajust Asimov en el canals (a) 2ℓ OS + $1\tau_{\text{had}}$ i (b) 2ℓ SS + $1\tau_{\text{had}}$. Els NPs estan ordenats per l'impacte en la determinació μ_{tHq} en ordre decreixent. Les caixes blaves i cianes fan referència a l'eix x superior i mostren l'impacte en μ_{tHq} . Els rectangles buits mostren l'impacte pre-ajust i els omplerts el post-ajust. A més, els valors dels NPs i les seues incerteses també s'inclouen com a punts i línies, respectivament. La incertesa dels NPs es mesura amb l'eix x inferior

de $\mu_{tHq}^{2\ell \text{OS}+1\tau_{\text{had}}}$ (el $^{+14.76}_{-13.72}$ en l'Equació 6.9) podria ser reduïda si la mostra estadística fos més gran.

La força del senyal i $k_{t\bar{t}, t\bar{t}X}$ obtinguts com a resultat de l'ajust perfil-semblança amb compartiments en el canal 2ℓ SS + $1\tau_{\text{had}}$ són:

$$\mu_{tHq} = -2.67 \pm 5.44(\text{tot.}) \pm 5.03(\text{stat.}) \quad (11)$$

$$k_{t\bar{t}, t\bar{t}X} = 0.73 \pm 0.14(\text{tot.}) \pm 0.10(\text{stat.}) . \quad (12)$$

El $\mu^{2\ell \text{SS}+1\tau_{\text{had}}} tHq$ és escalat a un valor negatiu que és compatible amb, tenint en compte la incertesa, la predicción del SM (com també és el cas per al $\mu^{2\ell \text{OS}+1\tau_{\text{had}}} tHq$ en l'Equació 6.9). El factor de normalització dels fons amb un parell de quarks top són escalats pels factors $k_{t\bar{t}, t\bar{t}X}$. Aquest factor de normalització és proper a la unitat i és compatible amb la predicción del SM.

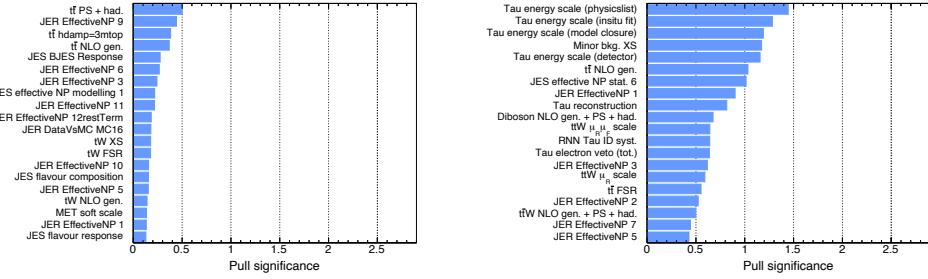


Figura 18: Les 20 desviacions més grans per als canals (a) $2\ell \text{OS} + 1\tau_{\text{had}}$ i (b) $2\ell \text{SS} + 1\tau_{\text{had}}$ ordenades de manera que la desviació més significativa es mostren en primer lloc.

Font d'incertesa	$2\ell \text{OS} + 1\tau_{\text{had}}$	$2\ell \text{SS} + 1\tau_{\text{had}}$
Incertesa del MC	± 7.136	± 0.983
Modelatge		
Incerteses teòriques	± 24.041	± 1.105
Incerteses a les PDFs	± 7.895	± 0.815
Experimental		
Instrumental	± 5.408	± 0.716
Instrumental: etiquetatge de sabor	± 0.455	± 0.126
Instrumental JES i JER	± 11.377	± 2.727
Factors de normalització	± 10.527	± 2.235
Total d'incertesa sistemàtica	± 28.543	± 4.081

Taula 15: Incerteses sistemàtiques en la mesura de μ_{tHq} en els canals $2\ell + 1\tau_{\text{had}}$ utilitzant totes les dades.

A més de la força del senyal, també es determina un límit superior mitjançant un test estadístic unidireccional. Aquest test depén del valor de μ_{tHq} , i el límit superior s'obté utilitzant el mètode CLs [308] per a establir un nivell de confiança del 95% (95% CL). La interpretació de la força del senyal, junt amb el límit superior, també es pot expressar en termes de la secció de producció del procés de senyal com es mostra a la Taula 16.

Un límit superior donat amb interval 95% CL significa que només hi ha una possibilitat del 5% que el resultat obtingut siga degut a una fluctuació estadística.

Però, què passaria si la física fos diferent de la descrita pel SM i l'acoblament de Yukawa amb el top tingués un signe oposat al del SM? A continuació, els mateixos estudis fets fins ara es realitzen sota l'acoblament de Yukawa invertit. Això es fa utilitzant mostres alternatives per a la producció completa de tH (és a dir, els

Límit superior per μ_{tHq}				
Canal	Mesurat	Esperat	1σ CL ₉₅	2σ CL ₉₅
2ℓ OS + $1\tau_{\text{had}}$	47.51	61.2	[41.95, 93.48]	[30.56, 148.6]
2ℓ SS + $1\tau_{\text{had}}$	13.83	16.94	[10.33, 30.22]	[6.991, 52.24]

Taula 16: Límits superiors observats i esperats per al μ_{tHq} per a ambdós canals $2\ell + 1\tau_{\text{had}}$.

processos $tHq + tWH$). Aquestes mostres alternatives han estat produïdes amb $y_t = -1$. Els resultats obtinguts són:

$$\begin{aligned}\mu_{tHq, y_t = -y_t^{\text{SM}}}^{2\ell \text{OS}+1\tau_{\text{had}}} &= -10.7^{+11.6}_{-39.4} (\text{tot.})^{+5.0}_{-4.8} (\text{stat.}) \\ \mu_{tHq, y_t = -y_t^{\text{SM}}}^{2\ell \text{SS}+1\tau_{\text{had}}} &= -0.3 \pm 1.2 (\text{tot.}) \pm 1.1 (\text{stat.}).\end{aligned}$$

Els límits superiors a la força de senyal es presenten a la Taula 17.

Límit superior per μ_{tHq}				
Canal	Mesurat	Esperat	1σ CL ₉₅	2σ CL ₉₅
2ℓ OS + $1\tau_{\text{had}}$ ($y_t = -y_t^{\text{SM}}$)	77.95	119.6	[46.97, 284.9]	—
2ℓ SS + $1\tau_{\text{had}}$ ($y_t = -y_t^{\text{SM}}$)	20.79	24.07	[7.478, 112.5]	[3.197, 342.4]

Taula 17: També s'inclouen els límits superiors observats i esperats per al μ_{tHq} per a ambdós canals $2\ell + 1\tau_{\text{had}}$ derivats amb el mètode CL sota la hipòtesi de l'acoblament de Yukawa invertit.

5 Conclusions

Aquesta tesi presenta l'estudi per a mesurar la producció directa d'un bosó de Higgs en associació amb un quark top simple, centrant-se en estats finals amb dos leptons de sabor lleuger i un lepton τ que decau hadrònicament, utilitzant el detector ATLAS. Atenent a la càrrega relativa entre el lepton lleuger, aquesta investigació es divideix en dos canals: 2ℓ OS + $1\tau_{\text{had}}$ i 2ℓ SS + $1\tau_{\text{had}}$.

La recerca d'un procés tan rar es motiva per la interacció complexa entre dues partícules fonamentals: el bosó de Higgs i el quark top. D'una banda, el bosó de Higgs juga un paper crític en la nostra comprensió de l'adquisició de massa per les partícules a través del mecanisme de Trencament Espontani de la Simetria. D'altra banda, el quark top, notable per ser la partícula més massiva en el Model Estàndard i l'única que decau abans de la seua hadronització. Per tant, s'espera

que l'acoblament de Yukawa entre aquestes dues partícules siga el més gran en el SM i es pot mesurar a través d'aquesta interacció. Aquesta mesura és central en el programa experimental del LHC i podria indicar una possible violació de CP, influenciant la secció de producció de tHq .

En aquesta tesi es discuteixen els fonaments teòrics de la física del quark top i del bosó de Higgs, es fa una revisió del detector ATLAS i el seu rendiment, i es descriuen la cadena de simulació i la reconstrucció d'objectes. Després, es detalla amb cura la recerca de la producció de tHq .

Aquesta anàlisi utilitza col·lisions protó-protó a $\sqrt{s} = 13$ TeV del detector ATLAS durant la Run 2 del LHC amb una lluminositat integrada total de 140 fb^{-1} . S'implementa i utilitza la informació a nivell de partó per a reconstruir el procés tHq . L'origen del lepton lleuger s'avalua mitjançant l'ús de BDT. Després s'aborda la taxa de partícules mal identificades utilitzant el mètode d'ajust de plantilles per a corregir els rendiments de MC.

Posteriorment, utilitzant diversos BDTs, es defineixen les SRs i CRs de l'anàlisi. Utilitzant aquestes regions, es realitza un ajust de la versemblança amb dades d'Asimov per determinar la sensibilitat de l'anàlisi. Després, fent servir totes les dades, es calculen els factors de normalització del fons dominants. Finalment, s'obté per a cada canal, els valors de la força del senyal de tHq . Aquests es mostren a la Figura 19. El valor esperat s'aconsegueix emprant dades d'Asimov i l'observat s'aconsegueix amb l'ajust del MC a les dades reals.

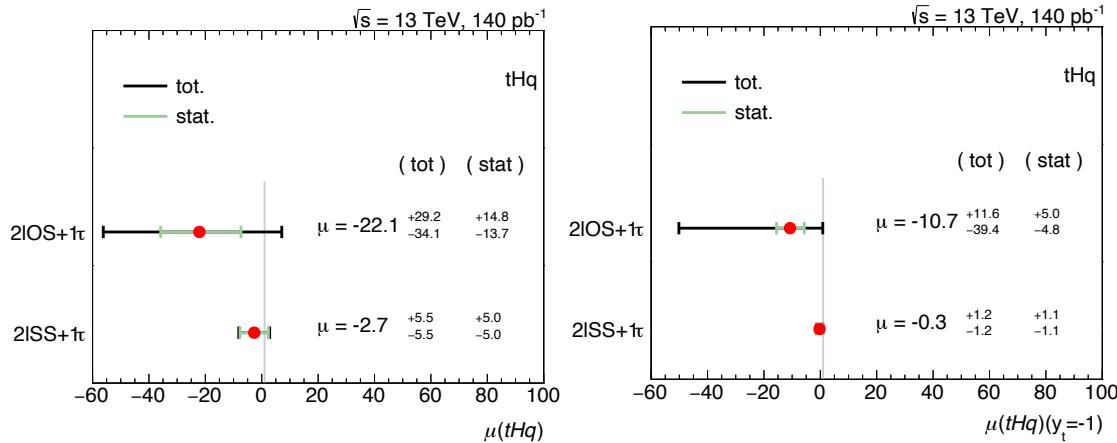


Figura 19: Valors de la força del senyal tant per al canal $2\ell \text{OS} + 1\tau_{\text{had}}$ com per al $2\ell \text{SS} + 1\tau_{\text{had}}$ sota (a) el SM i (b) la hipòtesi de l'acoblament de Yukawa invertit. La incertesa total (tot) inclou efectes estadístics i sistemàtics. L'incertesa estadística (stat) es mostra per separat.

Aquest resultat és completament compatible amb el Model Estàndard. Els límits al 95% CL en la secció de tHq es troben a la Taula 16 i a la Figura 20.

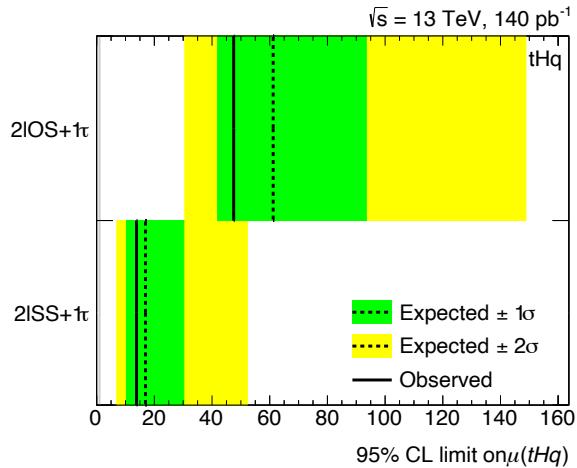


Figura 20: Límits superiors observats i esperats per al μ_{tHq} per a ambdós canals $2\ell + 1\tau_{\text{had}}$. Les àrees verdes i grogues representen les variacions de 1σ i 2σ del límit superior esperat.

Mirant cap al futur, les perspectives per a millorar els resultats d'aquesta anàlisi són prometedores, impulsades per diversos desenvolupaments anticipats:

- La propera Run 3 d'alta lluminositat promet una major significació estadística de l'anàlisi. Això serà molt beneficiós ja que en el canal $2\ell SS + 1\tau_{\text{had}}$ (el que té millor sensibilitat) la incertesa estadística és la dominant. En el canal $2\ell OS + 1\tau_{\text{had}}$, la incertesa deguda a la mostra estadística és similar a la incertesa sistemàtica i més esdeveniments de dades enriquiran sens dubte aquesta recerca.
- Les tècniques actuals per a estimar les taxes de fons deguts a objectes mal identificats també es beneficiaran d'un augment en la mostra estadística, ja que aquestes són mètodes basats en dades.
- Amb més estadístiques podríem explorar la divisió de les regions utilitzades en els càlculs de l'ajust segons la multiplicitat de trajectòries del τ_{had} (1-prong o 3-prong). Com que els jets amb diferent origen imiten de manera diferent el τ_{had} dependent de la seua "prongness", pot ser beneficiós explorar aquesta classificació quan es defineixin les regions utilitzades en l'ajust de semblança perfilada.
- Actualment hi ha una lleugera tensió entre la simulació i els resultats experimentals per a $t\bar{t}W$. Per tant, el progrés en les tècniques de simulació o una millor comprensió d'aquests processos refinaran més l'anàlisi. Això serà útil en el canal $2\ell SS + 1\tau_{\text{had}}$, on aquest és el segon fons més important.

A més, aquesta anàlisi es pot ampliar de diverses maneres. Primer, la definició de regions pot ser millorada combinant els BDTs descrits en aquesta tesi amb les xarxes neuronals que també estan sent utilitzades per les anàlisis en curs d'ATLAS. En segon lloc, incorporar els processos tWH en el senyal permetria realitzar una recerca de tH més completa i, per tant, es podria provar adequadament la hipòtesi de l'acoblament de Yukawa invertit del top.

Dels dos canals explorats en aquesta tesi, el $2\ell SS + 1\tau_{had}$ és el que pot afegir més sensibilitat a futures combinacions amb altres canals de tHq .

