Physics with b-jets using the ATLAS Run 2 data : from calibration to Higgs boson couplings measurements in the V H, H → bb channel

Romain Bouquet

Sorbonne Université

École Doctorale des Sciences de la Terre et de l'Environnement
et Physique de l'Univers de Paris, ED 560

Laboratoire Astroparticule et Cosmologie &
Laboratoire de Physique Nucléaire et de Hautes Énergies

Paris

Physics with b-jets using the ATLAS Run 2 data: from calibration to Higgs boson couplings measurements in the $VH, H \to b\bar{b}$ channel

Presented by:
Romain Bouquet

Submitted in fulfillment of the requirements for the degree of:
Docteur des Sciences Sorbonne Université

Supervised by:
Giovanni Marchiori

Defended on the 24th of January 2023 in front of the committee:

Mr.	*Eli*	*Ben-Haïm*	*President*
Mr.	*Andrea*	*Rizzi*	*Referee*
Ms.	*Viola*	*Sordini*	*Referee*
Ms.	*Hannah*	*Arnold*	*Examiner*
Mr.	*Jean-François*	*Grivaz*	*Examiner*
Mr.	*Luca*	*Scotto Lavina*	*Examiner*
Mr.	*Jean-Baptiste*	*de Vivie de Régie*	*Examiner*
Mr.	*Giovanni*	*Marchiori*	*Supervisor*

“Now is the time to understand more,
so that we may fear less”
Marie Curie

“Et loin des beaux discours, des grandes théories
À sa tâche, chaque jour, on pouvait dire de lui
Il changeait la vie”
Jean-Jacques Goldman

Thesis summaries

Abstract

Many signatures of interest for Standard Model (SM) measurements and for searches of beyond-SM (BSM) phenomena include at least one b-jet in their final state. For instance, the main decay of the Higgs boson (H) is to a pair of b-quarks with a SM predicted branching fraction of 58%. The Higgs boson is the cornerstone particle of the SM as it is responsible for the mass of the other elementary particles and for the spontaneous symmetry breaking of the electroweak interaction. $H \to b\bar{b}$ decays are thus essential to probe the quark mass generation mechanism and constrain BSM theories in the Higgs sector. Moreover, other processes such as top-quark production and heavy flavour decays of Z boson or of hypothetical new resonances can produce b-jet(s).

This thesis focuses on calibration and physics measurements of processes involving b-jets exploiting proton-proton collisions collected at a center-of-mass-energy of 13 TeV between 2015 and 2018 by the ATLAS experiment at the Large Hadron Collider.

The first aspect of the thesis work concerns the first ATLAS in situ differential calibration of the energy scale of b-jets, referred to as b-JES. Events with a $t\bar{t}$ lepton+jets signature are selected and the b-JES is derived thanks to a new software framework developed for that purpose. The b-JES measurement has been performed inclusively for b-tagged jets transverse momenta above 20 GeV, and thanks to the large size of the control sample under study, differentially as a function of the b-tagged jet transverse momentum between 30 and 500 GeV. The specific b-tagged jet energy correction in data is consistent with unity, with an uncertainty of 2.5% for the inclusive measurement, and decreasing from 4% to 1.9% for b-tagged jet transverse momenta increasing between 30 and 500 GeV for the differential measurement. The results have been published in a dedicated ATLAS conference note and have been presented at the Moriond Electroweak 2022 conference during a YSF talk session.

The second part of the thesis is about a study of the properties of the Higgs boson. The Higgs-to-bottom Yukawa coupling, the SM predictions and kinematic properties of Higgs boson production are probed looking for Higgs bosons produced in association with a weak vector boson ($V = W, Z$) and decaying to a pair of b-quarks. The vector bosons are reconstructed through their leptonic decays. The $V(\to \text{leptons})H(\to b\bar{b})$ final state provides the best sensitivity to the measurement of the VH production cross-section and the $H \to b\bar{b}$ decay rate. Thanks to improved b-jet and c-jet identification algorithms and other refinements in the analysis procedures such as a more extensive use of multivariate techniques, an increase of the order of 20% of the sensitivity is expected for the $VH, H \to b\bar{b}$ analysis with respect to the previous ATLAS publication based on the same data. The expected precision on the signal strength, defined as the ratio between the observed signal yield and the corresponding SM prediction, is equal to $\mu_{VH}^{H\to b\bar{b}} = 1.00^{+0.15}_{-0.14}$. Furthermore, the expected 95% confidence level upper limit on the signal strength of the $VH, H \to c\bar{c}$ process is expected to be of 16 times the prediction of the Standard Model and is a factor 2 lower than in the previous published analysis of the same data. A publication of those results is foreseen for early 2023.

Résumé

Plusieurs signatures d'intérêts pour des mesures du Modèle Standard (MS) et pour des recherches de phénomènes au delà du MS incluent au moins un b-jet dans leur état final. Par exemple, la désintégration principale du boson de Higgs (H) est en paire de quarks-b avec une prédiction, par le MS, pour le rapport d'embranchement de 58%. Le boson de Higgs est la pierre angulaire du MS car il est responsable de la masse des autres particules élémentaires et de la brisure spontanée de symmétrie de l'interaction électrofaible. Les désintégrations $H \to b\bar{b}$ sont donc essentielles pour tester le mécanisme de création de masse des quarks et pour constraindre les théories au delà du MS dans le secteur de la physique du Higgs. De plus, d'autres processus comme la production de paires de quarks-top et les désintégrations du boson Z en saveurs lourdes ou d'hypothétiques nouvelles résonances peuvent produire un ou des b-jet(s).

Cette thèse se focalise sur la calibration et les mesures physiques des processus impliquant des b-jets en exploitant des collisions proton-proton collectées à une énergie de 13 TeV dans le centre de masse entre 2015 et 2018 par l'expérience ATLAS au Grand Collisionneur de Hadrons.

Le premier aspect du travail de thèse porte sur la première calibration différentielle in situ de l'échelle d'énergie des b-jets au sein d'ATLAS, appelée b-JES. Les événements présentant les caratéristiques des processus $t\bar{t}$ lepton+jets sont sélectionnés et la b-JES est obtenue grâce à un nouveau programme développé à cet effet. La b-JES a été mesurée de manière inclusive pour des jets b-étiquetés dont l'impulsion transverse est supérieure à 20 GeV, et également, grâce à la taille importante de l'échantillon de contrôle, en fonction de l'impulsion transverse des jets b-étiquetés entre 30 et 500 GeV. La correction spécifique pour l'échelle d'énergie des jets b-étiquetés est en adéquation avec l'unité, l'incertitude associée à la mesure inclusive est de 2.5%, et décroît de 4% à 1.9% pour des jets b-étiquetés dont l'impulsion transverse croît entre 30 et 500 GeV pour la mesure différentielle. Les résultats ont été publiés dans une note de conférence d'ATLAS et ont été présentés à la conférence Électrofaible de Moriond 2022 pendant une session YSF.

La seconde partie de la thèse est dédiée à l'étude des propriétés du boson de Higgs. Le couplage de Yukawa du Higgs avec le quark-bottom, les prédictions du MS et les propriétés cinématiques de la production du boson de Higgs sont investigués par le biais de la recherche de bosons de Higgs produits en association avec un boson vecteur de l'interaction faible ($V = W, Z$) et se désintégrant en une paire de quarks-b. Les bosons vecteurs sont reconstuits par le biais de leur désintégration leptonique. L'état final $V(\to \text{leptons})H(\to b\bar{b})$ offre la meilleure sensibilité pour la mesure de la section efficace de la production VH et du rapport d'embranchement $H \to b\bar{b}$. Grâce à l'amélioration des algorithmes d'identification des jets-b et jets-c, une augmentation de 20% de la sensibilité est attendue pour l'analyse $VH, H \to b\bar{b}$ par rapport à la précédente publication basée sur les mêmes données. La précision attendue pour la force du signal, définie comme le ratio entre le signal observé et celui prédit par le MS, est égal à $\mu_{VH}^{H\to b\bar{b}} = 1.00^{+0.15}_{-0.14}$. Par ailleurs, le niveau de confiance à 95% attendu pour la limite supérieure de la force du signal du processus $VH, H \to c\bar{c}$ est 16 fois la prédiction du Modèle Standard et est un facteur 2 inférieur à la précédente analyse publiée pour les mêmes données. Une publication de ces résultats est prévue pour début 2023.

Acknowledgements

Through those acknowledgements, I would like to emphasize the crucial role that many persons or groups[1] have played prior and/or during my thesis work.

First, I would like to thank ***all the members of the jury*** for accepting being part of that committee, and for all the feedback I received from them which for sure has improved the quality of this manuscript. It has been an honor defending my thesis in front of such experts from different experiments and nationalities which is something I truly admire about the community of Particle Physics and of Science in general since it gathers persons from different horizons. I also really appreciated that all the members could be present at the defense as in the recent past, due to the Covid pandemic, it would not have been possible.

I would like to express my entire gratitude to the two referees, ***Andrea Rizzi and Viola Sordini***, for having reviewed my thesis manuscript[2]. I also wanted to address a special thanks to ***Eli Ben-Haïm*** for having accepted being the president of the jury.

I am deeply thankful to my supervisor, ***Giovanni Marchiori***, who accepted me as a PhD student. I am grateful for his support, for all the advice and knowledge he provided me throughout those years while at the same time letting me the liberty of exploring my own ideas. I greatly valued the clarity of his explanations, his broad and precise knowledge, his availability, and his always constructive and positive feedback. It has been a pleasure working with him, I learned much more than I could ever imagine thanks to him. Finally, I also enjoyed our non physics related discussions[3].

In a more general way, I would like to thank ***all the persons from the APC and LPNHE laboratories*** from physicists, post-docs, PhDs and trainee students to computing, technical and administrative services who have clearly contributed to the smooth running of my thesis and made it enjoyable both from a scientific but also extra-professional point of view. I would like to thank in particular ***the persons from the ATLAS teams of the two laboratories*** with whom I interacted a lot and shared nice moments with.

This work would not have been possible without the support of ***the Jet/ETmiss, the In situ jet calibration and the $VH(b\bar{b}/c\bar{c})$ ATLAS groups***. Here I want to highlight that Science is about a collaborative work that requires connecting various knowledge and skills, and ATLAS is a perfect example for that. I benefited from the expertise, advice and the work of the past and present members of the ATLAS collaboration especially from those teams: all this "legacy" and guidance have been precious and helpful.

The ***ATLAS $VH(b\bar{b}/c\bar{c})$ analysis team*** played a key role in my thesis. Working with this team was an incredible experience and an immense chance: thanks to everyone for the fruitful discussions and for always providing pertinent insights. I also enjoyed our social events and meeting such nice persons was delightful[4]. I am very much looking forward the legacy analysis results which will reflect the large amount of work and dedication of the group. I owe a large tribute to the analysis contacts, ***Hannah Arnold, Yanhui Ma and Elisabeth Schopf***, for all the discussions that I had with them, all their advice, for taking the time to answer my multiple questions and for their kindness. Thank you as well for having entrusted me as responsible for the analysis framework code. I have also to thank ***Francesco Armando di Bello and Elisabeth Schopf*** for all their guidance for the multivariate studies I performed: interacting with them has been enlightening. Lastly, I wanted to thank ***Giulia di Gregorio*** as her various remarks, during conversations I had with her

1. All the names mentioned in the acknowledgements are alphabetically ordered and are not traducing a personal preference.
2. Maybe the proper wording would be "my a bit too long thesis manuscript". Sorry for that!
3. Thank a lot Giovanni for always having brought that positive energy whatever the circumstances!
4. To the entire $VH(b\bar{b}/c\bar{c})$ analysis team, I am wishing to all of you a bright and successful future!

while writing this manuscript, have been more than once useful.

Reaching a PhD degree would not have been possible without the recommendations of my previous internship supervisors, ***Florian Beaudette, Mateusz Dyndal, Beate Heinemann and David Rousseau***, while applying for internships, masters and ultimately for this thesis: I would like to thank them for their support. Thanks as well to ***all the professors*** I had, especially in Physics and Mathematics, for the quality of their teaching and having encouraged me pursuing this path.

Particle Physics analyses require powerful, shared computing resources[5], and as well sophisticated softwares[6]. I have to admit I am extremely amazed by the computing infrastructures, such as the grid, of CERN and of its associate institutes. Thanks to ***all the persons ensuring the quality and development of such infrastructures and softwares***: the results presented in this thesis are partially thanks to your work. I have asked you many questions over those years, your help has been decisive. In the same spirit, many thanks to the ***Stack Overflow community***.

Carrying out a career as a physicist has been and is still a long term goal for me since at least 15 years now. The road is still long but obtaining a PhD degree is for sure a step forward. It has not always been simple[7] to get there, but luckily I met incredible people along the way, and I really feel privileged for that! To ***all the persons mentioned below***, from the bottom of my heart thanks for being by my side since years (or even decades for some of you)! There are not strong enough words to express you how much I am grateful to know you!

I would like to begin with my childhood friend, ***Anis***[8], thanks for your friendship over so many years, all the adventures and emotions we had together especially when playing or watching football. To my closest friends, ***Cyrielle, Margaux and Zoé***[9], thanks for always bringing that craziness and funny vibes whenever we see each other. To ***Alessandro***[10] ***and Giorgos***[11], you are amazing guys, this trip in Greece was just incredible. To ***Alessandro and Sara***[12], I will always keep in mind our nights at each other places during the curfew period. To ***Angèle, Baptiste, Raphaël, Roxane***[13] and later ***Félix*** joined us, those Friday night sessions mainly at La Bulle à Bière or sometimes at Les Arènes were an essential part of my life. To all persons I shared my office with[14], ***Ang, Giulia***[15]***, Jad, Jean-Phillipe, Joao, Keerthi, Louis, Luis***[16]***, Marton, Qiuping, Renaud***[17]***, Tong, Valentina and Yulei***, thanks for all the funny moments we shared together, to me you were much more than office mates. To my old time friends from the south of France, the ***Pompelup and Lesron***[18] ***teams***[19] thanks for always being easy going and chill persons. To the ***Joffre team***[20], it is always a pleasure seeing you. To the ***NPAC crew***[21], all the fun indeed started one year before beginning our thesis thanks to you. To ***all the APC and LPNHE PhD students and post-docs***[22], thanks for having made my thesis years extraordinary.

5. It is not a surprise that the World Wide Web was invented at CERN!
6. For instance, the ROOT software functionnalities never cease to impress me!
7. Especially in 2016!
8. As we use to say "On est là!" either on or off a football pitch!
9. Thanks for staying the same even if one of us is often 10 000 km away from us... I will not give a name here though!
10. Alessandro, thanks for everything! From our chats at the lab to the countless nights and also parties we spent together!
11. Thanks captain Zorz for constituting that franco-italo-greco crew and leading us to Greece. Yamas to you! Raki Raki Rumba!
12. Sara, I will continue to tap my beer on the table before drinking as a famous Spanish proverb recommends to!
13. Also refered to as the polar bear or le Vieux Cèdre fan club, aka the best team of the CDD. I am glad I met you (or already know you for some of you) because you are awesome persons!
14. The list is pretty long... since I have been working in 2 laboratories and my offices were big!
15. Giulia, history will remember that for some reason you left our office at some point... I really liked talking with you!
16. Luis, I am including you in our office since you were always visiting us. Thanks a lot for all the interesting conversations I had with you whether it was about science or not! Everyday I was waiting you to come, it was great moments!
17. Renaud, I will remember you took my place as soon as I left, enjoy the rest of your thesis at the best desk!
18. I cannot put the entire name of that group in this manuscript for obvious reasons.
19. We use to say our IQ decreases whenever we see each other, but I want to keep seeing you!
20. My "not true friends team" is more accurate according to you... already 10 years that we know each other!
21. The Friday nights were a must! Special thanks to Sullivan that positronium TL and playing football with you were memorable!
22. That international atmosphere on a daily basis was incredible! I will forever remember our card games, the Lebanese market sandwiches, those raclettes, fondues or even crepes nights, and for sure that post-Covid break to Cannes and Nice!

Last but not least, to ***my family in particular my parents and sister***[23], your love and support has always been a pillar. If I made it this far it is without a doubt thanks to you. Thank you so much to have allowed me pursuing that dream and cheered me up during tough times. Even if now it has been an eternity you are not able to follow what I am doing, I just want you to know that your education and the values you passed on to me have been one of the most precious things I could ever receive from someone. My final thanks go to ***Cookie***[24], the cat of my parents, who joined our family in the middle of my thesis. Cookie, not that you were much interested by my work, but thanks for your support when I could see you.

I will finish these acknowledgements mentioning ***you***, the reader of those lines, if you got to that point it probably means you are one of the persons mentioned earlier or you were interested[25] by what I wrote so far: in both cases, thanks. If you continue reading this manuscript, I hope you will appreciate its content. I just want you to know that for my entire thesis work "I have no regrets because I gave my best every day no matter what was my level" as my father would say.

23. Élodie, even after this graduation, the answer is yes I will continue to bother you especially with reels or whatever the social networks will come out in the future! That is my role of older brother...
24. Cookie, as Élodie would say "Enjoy your life you deserved it!"
25. Or I should say "at least I hope you were interested..."

Table of contents

Thesis summaries **5**
Abstract . . . 5
Résumé . . . 6

Acknowledgements **7**

Table of contents **11**

Introduction **15**
Personal contributions . . . 16
Structure of the thesis . . . 17

1 Theoretical framework **19**
1.1 The Standard Model . . . 20
1.1.1 Quantum field theory and Standard Model Lagrangian . . . 23
1.1.2 Quantum electrodynamics . . . 24
1.1.3 Quantum chromodynamics . . . 25
1.1.4 Electroweak unification . . . 26
1.2 The Higgs mechanism . . . 28
1.2.1 The Brout-Englert-Higgs mechanism . . . 28
1.2.2 The fermion Yukawa couplings . . . 30
1.3 Higher order corrections . . . 31
1.4 Conceptual and experimental tools for probing the Standard Model predictions . . . 32
1.4.1 Cross-section . . . 33
1.4.2 Decay rate and branching ratio . . . 34
1.4.3 Signal strength . . . 34
1.4.4 Simplified template cross-section . . . 35
1.4.5 Coupling-strength modifiers . . . 37
1.4.6 Effective field theory . . . 37
1.5 Event simulation . . . 40
1.6 Higgs boson measurements at the LHC . . . 43
1.7 Limitations of the Standard Model . . . 53

2 The LHC and the ATLAS detector **57**
2.1 The Large Hadron Collider . . . 57
2.1.1 Luminosity and number of interaction per bunch crossing . . . 58
2.1.2 Production and acceleration of particles . . . 60
2.1.3 Experiments of the LHC . . . 61
2.2 The ATLAS detector . . . 62
2.2.1 System of coordinates and physics related definitions . . . 63
2.2.2 Particle energy loss and shower development . . . 65
2.2.3 Magnets system . . . 67
2.2.4 Inner detector and tracking system . . . 69
2.2.4.1 The pixel detector . . . 71
2.2.4.2 SemiConductor Tracker . . . 72
2.2.4.3 Transition Radiation Tracker . . . 72
2.2.5 Electromagnetic and hadronic calorimeters . . . 73
2.2.5.1 Electromagnetic calorimeter . . . 73
2.2.5.2 Hadronic calorimeter . . . 75

2.2.6 Muon spectrometer . . . 76
2.2.7 Trigger and data acquisition . . . 77

3 Reconstruction and identification of objects in ATLAS 81
3.1 Vertex and track reconstruction . . . 82
3.2 Electrons . . . 86
3.2.1 Electron reconstruction . . . 86
3.2.2 Electron identification and isolation criteria . . . 89
3.2.3 Electron calibration . . . 91
3.3 Muons . . . 93
3.3.1 Muon reconstruction . . . 93
3.3.2 Muon identification and isolation criteria . . . 94
3.4 Taus . . . 96
3.5 Photons . . . 97
3.6 Jets . . . 97
3.6.1 Reconstruction of jets with the anti-k_t algorithm . . . 98
3.6.2 Small-R jets . . . 99
3.6.2.1 PFlow and EMTopo reconstruction algorithms . . . 100
3.6.2.2 Calibration of small-R jets . . . 100
3.6.3 Large-R jets . . . 106
3.6.3.1 Reconstruction of large-R jets . . . 106
3.6.3.2 Calibration of large-R jets . . . 107
3.6.4 VR track-jets and ghost association . . . 110
3.6.5 Flavour tagging . . . 112
3.6.6 Jet cleaning . . . 116
3.7 Overlap removal . . . 116
3.8 Missing transverse energy . . . 117

4 Energy scale calibration of b-tagged jets using $t\bar{t}$ lepton+jets events 121
4.1 Introduction to the b-JES measurement . . . 122
4.2 Data and simulation samples . . . 127
4.3 Event selection . . . 129
4.4 In situ b-JES measurement with a template method . . . 132
4.4.1 $\chi^2(\alpha)$ curves determination . . . 133
4.4.2 b-JES statistical uncertainty . . . 138
4.4.3 b-JES systematic uncertainties and total uncertainty . . . 138
4.5 Results of the b-JES measurement . . . 142
4.5.1 Template method validation with pseudo-data . . . 143
4.5.2 Inclusive measurement in data . . . 146
4.5.3 Differential p_{T} measurement in data . . . 149
4.6 Conclusion . . . 153

5 Higgs decay to a pair of b- or c-quarks in association with a vector boson 155
5.1 Introduction to $VH \to b\bar{b}/c\bar{c}$ events and Higgs decay topologies . . . 157
5.2 Data and simulation samples . . . 161
5.3 $VH, H \to b\bar{b}/c\bar{c}$ event reconstructions and selections . . . 165
5.3.1 Overview of the $VH, H \to b\bar{b}/c\bar{c}$ event combination . . . 169
5.3.2 Common selections for the resolved and boosted topologies . . . 173
5.3.3 Specific selections for the resolved topology . . . 175
5.3.4 Specific selections for the boosted topology . . . 177
5.3.5 Specific selections for the 0-lepton channel . . . 178
5.3.6 Specific selections for the 1-lepton channel . . . 179

5.3.7 Specific selections for the 2-lepton channel . . . 179
5.3.8 Specific $VH, H \to c\bar{c}$ selections . . . 180
5.4 Jet mass scale calibration for small-R jets . . . 182
5.5 Flavour tagging methods . . . 185
5.6 Multivariate discriminants in the $VH, H \to b\bar{b}/c\bar{c}$ analyses . . . 189
5.7 Fit formalism . . . 195
5.7.1 Hypothesis testing, p-value and significance . . . 195
5.7.2 Likelihood function, signal strength and observed significance . . . 196
5.7.3 Asimov dataset and expected significance . . . 199
5.7.4 Pulls, impacts of nuisance parameters and correlations in fit . . . 200
5.7.5 Breakdown of uncertainties . . . 201
5.7.6 Asimov significance . . . 201
5.7.7 Transformation D . . . 202
5.8 Background normalizations and systematic uncertainties . . . 203
5.8.1 Background normalizations and modelling uncertainties . . . 203
5.8.2 Experimental uncertainties . . . 207
5.9 Results from the previous publications . . . 209
5.9.1 $VH, H \to b\bar{b}$ resolved and boosted combination . . . 210
5.9.1.1 Combination strategy . . . 210
5.9.1.2 STXS division of phase space . . . 211
5.9.1.3 Results of the STXS measurement . . . 212
5.9.1.4 EFT interpretation . . . 219
5.9.1.5 Comparison with the CMS latest results . . . 222
5.9.2 $VH, H \to b\bar{b}/c\bar{c}$ combination . . . 224
5.9.2.1 $VH, H \to c\bar{c}$ only likelihood fit . . . 224
5.9.2.2 Combination of $VH, H \to b\bar{b}$ and $VH, H \to c\bar{c}$ . . . 227
5.9.2.3 Comparison with CMS latest results . . . 228
5.9.3 $VH, H \to b\bar{b}/c\bar{c}$ cross-check analyses . . . 229
5.9.3.1 Dijet invariant mass cross-check . . . 229
5.9.3.2 Diboson cross-check . . . 230
5.10 Expected results for the Legacy Run 2 analysis . . . 235
5.10.1 Expected $VH, H \to b\bar{b}$ results . . . 235
5.10.2 Expected $VH, H \to c\bar{c}$ results . . . 242
5.11 Future improvements and prospects for the $VH, H \to b\bar{b}/c\bar{c}$ analysis . . . 246
5.11.1 Run 3 of the LHC . . . 246
5.11.2 High luminosity phase of the LHC . . . 249
5.11.3 Future Circular Collider . . . 251
5.12 Conclusion . . . 252

6 MVA studies for the $VH, H \to b\bar{b}$ analysis 255
6.1 $VH, H \to b\bar{b}$ resolved MVA study in the 0-lepton channel . . . 256
6.1.1 Training, evaluation and input variables . . . 256
6.1.2 Hyperparameters . . . 260
6.1.3 Overtraining checks . . . 260
6.1.4 BDT performances . . . 263
6.1.5 Conclusion of the MVA training . . . 266
6.2 Combination study of the resolved and boosted topologies in the 1-lepton channel . 266
6.2.1 Event selection and combination strategies . . . 266
6.2.1.1 Combination strategies . . . 266
6.2.1.2 Multivariate discriminants . . . 268
6.2.2 Significances and transformation D . . . 268
6.2.3 Combination results . . . 269

6.2.3.1 Results for the different strategies . . . 269
6.2.3.2 Impact of the transformation D parameters . . . 277
6.2.3.3 Impact of the b-tagging strategy for the boosted topology . . . 280
6.2.4 Conclusion of the combination study . . . 287
6.3 Conclusion . . . 287

General conclusion **289**

Bibliography **291**

Appendices **309**
A Penalized spline: some technical details . . . 309
A.1 Cross-validated residual sum of squares tests . . . 309

Introduction

The Standard Model (SM) of elementary particles is a theory that describes three of the four fundamental interactions in Nature: the weak, the electromagnetic and the strong forces respectively mediated by the weak vector bosons ($V = W^{\pm}, Z^0$), the photon (γ) and the gluons (g). The gravitational force is the only interaction not described by that theory.

Within the Standard Model, a key role is played by the Higgs field, which is responsible for the spontaneous symmetry breaking between the electromagnetic and weak interactions and for the origin of the masses of the elementary particles. For this reason, since its theoretical prediction in 1964, the Higgs boson - *i.e.* the quantum excitation of the Higgs field - has been searched for by many experiments.

In 2012, the ATLAS and CMS collaborations announced the observation a particle [1, 2] with a measured mass $m_H = 125$ GeV presenting the expected characteristics of the Higgs boson as predicted by the Higgs mechanism. Such discovery has been one of the main reason of the construction of the Large Hadron Collider (LHC) as the Higgs boson was a missing part of the SM and is an essential particle of this theory.

Since then, ATLAS and CMS observed several of the Higgs production modes (gluon-gluon fusion, vector-boson fusion, associated production with a W, Z boson or a $t\bar{t}$ pair) and decay modes ($H \to \gamma\gamma, ZZ, W^+W^-, b\bar{b}, \tau^+\tau^-$) predicted by the SM. Those two collaborations have now entered a precision measurement era were the Higgs properties are measured in the context for instance of differential or simplified template cross-section (STXS) approaches or reinterpreted with an effective field theory (EFT). The Higgs boson could for instance be a portal for discovering physics phenomena beyond the SM (BSM) or dark matter candidates as in case of massive dark particles, they should strongly couple with the Higgs and would induce deviations from the SM predictions. Thus, precise Higgs measurements allow constraining BSM theories. Probing possible deviations from the SM especially in high energy regime is essential as some hints indicate that the SM could be a low energy limit of a broader theory.

In 2018, the main decay mode of the Higgs ($H \to b\bar{b}$) to a pair of b-quarks, which has a branching fraction of 58%, was observed both by ATLAS and CMS [3, 4]. Such observation was made possible thanks to the so called Higgs-strahlung mechanism (denoted VH) where a Higgs boson is produced in association with a weak vector boson ($V = W, Z$) in particular targeting leptonic decays of the vector boson ($Z \to \nu\bar{\nu}, W \to l\nu$ and $Z \to l^+l^-$ with $l = e, \mu$) respectively classified in the 0-, 1- and 2-(charged) lepton channels. Such production mode, despite its low cross-section allows for a significant reduction of the multijet background, and hence benefit from an increase of the signal to background ratio compared to the gluon-gluon fusion (ggF), vector boson fusion (VBF) and $t\bar{t}H$ mechanisms. As a result, the $VH, V \to \text{leptons}$ production mode offers the best inclusive sensitivity to $H \to b\bar{b}$ and $H \to c\bar{c}$ decay modes to date.

For a $H \to b\bar{b}$ decay, depending on the energy of the Higgs boson, the two b-hadrons can either be reconstructed as two small-radius jets for low energy-Higgs bosons known as the resolved topology or captured in a single large-radius jet for high-energy Higgs bosons referred to as boosted topology.

The two previous ATLAS VH publications were a resolved-boosted $VH, H \to b\bar{b}$ combination [5], and a resolved $VH, H \to b\bar{b}$ and $VH, H \to c\bar{c}$ analyses combination [6]. During my thesis, I contributed to the activities of the ATLAS $VH, H \to b\bar{b}/c\bar{c}$ analysis team, working on a "legacy" publication of a joint measurement of the $VH, H \to b\bar{b}$ (resolved and boosted topologies) and the $VH, H \to c\bar{c}$ (resolved topology only) processes with the full Run 2 ATLAS dataset. This analysis comes with major updates and changes with respect to the previous publications, in particular the

multivariate analysis (MVA) approach has been extended to the boosted $VH, H \to b\bar{b}$ and the resolved $VH, H \to c\bar{c}$ analyses. Moreover, the b-tagging (respectively and c-tagging) performances have been improved thanks to the use of more advanced deep-learning algorithms with better light- and c-jet (respectively light- and b-jet) tagging rejections. A significant improvement in the sensitivity to the $VH, H \to b\bar{b}/c\bar{c}$ processes is thus expected. The expected results of the current state-of-the-art of this analysis, which is still being optimized by means of a "blind-analysis" procedure, will be extensively presented. The unblinding of the data and the publication of the results are foreseen in early 2023.

Personal contributions

In this thesis, when relevant, the details of my personal contributions are reminded at the beginning of the chapters. All figures, schemes and tables which are not mentioning a reference in their caption, such as "taken from Ref…", have been produced by myself.

My main personal contributions are related to the physics of b-jets which is the guideline of this entire thesis.

The first aspect of this thesis, concerns the* in situ *calibration of the b-jet energy scale referred to as the b-JES which is a residual correction on the energy of b-jets applied only in data. Currently, the *in situ* jet energy calibration in ATLAS is flavour-inclusive even though the energy response of b-jets and the radiation pattern of their showers are expected to be different from the light-, c- and gluon initiated jets since the mass of b-quarks is larger. As a result, ATLAS prescribes an additional uncertainty to b-jets based on differences in response between different simulated samples to cover possible differences between the modelling of the response of b-jets, but no measurement of such energy scale has been performed with data to date.

During my PhD, and in particular in the first half of the three years of the thesis work, I performed a feasibility study, where for the first time, a b-JES inclusive and differential measurement in data exploiting events with a $t\bar{t}$ lepton+jets signature is carried out. This novel and never performed study is entirely based on my work since I implemented *ex nihilo* the dedicated framework to achieve such measurements. In addition, the results I have obtained, have been published in a dedicated ATLAS conference note [7] and I have presented [8, 9] them at the Moriond 2022 Electroweak Interactions & Unified Theories conference [10] during a Young Scientist Forum (YSF) session. This work allowed me to fully qualify as an author of the ATLAS publications.

The second aspect of this manuscript, concerns the study of the Higgs boson decay to a pair of b-quarks in the VH production mode looking for leptonic decays of the vector boson. Since a bit more than 2 years now, I am involved in the ATLAS $VH, H \to b\bar{b}/c\bar{c}$ *Legacy Run 2* analysis.

My work was focused on the $VH, H \to b\bar{b}$ analysis. In particular, I took part in the multivariate analysis (MVA) studies since this approach enables to enhance the signal sensitivity and thus plays a leading role. In the 0-lepton channel, for the resolved topology, I was in charge of the optimization of the multivariate discriminant (hyperparameters of the training procedure and choice of the input variables) used to separate signal from background after the event selection. A new training, with respect to that of the multivariate discriminant used in the previous publications, was needed in order to fully exploit the improvements in sensitivity provided by the updated b-tagging and jet reconstruction algorithms.

For the 1-lepton channel, I performed an optimization study for the combination of the resolved and boosted topologies exploring more complex strategies than previously probed. This study was of high interest since for the first time the MVA approach was used for the boosted topology.

I am also one of the three main editors of an ATLAS internal note which documents all the MVA studies performed for the $VH, H \to b\bar{b}/c\bar{c}$ Legacy Run 2 publication. This internal note is a

major update with respect to the previous round since the MVA is now used in all regions of the $VH, H \to b\bar{b}/c\bar{c}$ analysis.

From a technical point of view, since a bit more than one year (October 2021), I am the responsible for the analysis framework of the ATLAS $VH(b\bar{b}/c\bar{c})$ group. That code is a central part of the workflow since it is producing the inputs needed for fits and multivariate analysis studies: it is thus used on a daily basis by several persons. In that context, I was part of the $VH(b\bar{b}/c\bar{c})$ coordination team that takes the important decisions concerning the analysis. Since two years, I have been one of the main code developer and have been deeply involved in developing new functionalities, implementing analysis related updates, improving and keeping up to date the code and its documentation in order to provide a tutorial for beginners but also for much more advanced usage of the framework, I took care of the CI/CD pipelines, helped other persons of the group on feature implementations, providing advice, code debugging tips and so on. I was also one of the main reviewer of the code and was in charge of releasing "reference" versions of the code for our group.

I have also strongly participated in the improvement effort of the analysis workflow, enhancing its stability to transient errors, automatizing steps such as the summation of outputs, the verifications of outputs and very recently implementing the submission of jobs to the grid which should in principle help our group to get results more quickly. Moreover, some of the developments that I made were not only useful to our group, as part of our framework is shared among several Higgs analyses and thus can benefit to other ATLAS groups. Some of the key improvements included have been a reduction by a factor of two of the size of the output files thanks to a more clever way to store information and also identifying two critical bugs in the ATLAS analysis software and in ROOT [11, 12] that improved the efficiency and stability of the analysis code, such as a reduction by an order of magnitude of the time taken by the summation of outputs which was one of the most time-consuming steps of the analysis chain.

Structure of the thesis

This thesis is organized as follows. Chapter 1 presents the theoretical framework relevant for the content of this thesis in particular related to the Higgs physics. In Chapter 2, an overview of the LHC and a detailed description of the ATLAS detector are provided. The reconstruction and identification of the different objects with the ATLAS detector is then explained in Chapter 3 with an emphasis for objects of interest for this thesis (tracks, leptons and jets). A feasibility study of the b-JES inclusive and differential measurement with $t\bar{t}$ lepton+jets events is discussed in Chapter 4. Chapter 5 is a general presentation of the $VH, H \to b\bar{b}$ and $VH, H \to c\bar{c}$ analyses with a special attention on $VH, H \to b\bar{b}$ since my work was mainly focused on this channel. The expected results based on the full ATLAS Run 2 dataset and the latest ATLAS improvements are shown in that chapter. A study that I carried out about the jet mass scale calibration of small radius jets in the $VH, H \to b\bar{b}$ resolved topology is also presented in this chapter. Chapter 6, is dedicated to MVA studies for the $VH, H \to b\bar{b}$ analysis that I performed for the analysis concerning the boosted decision tree training in the 0-lepton channel for the resolved topology and an optimization of the resolved and boosted $VH, H \to b\bar{b}$ topology combination strategy in the 1-lepton channel.

Chapter 1

Theoretical framework

1.1 The Standard Model . . . 20
1.1.1 Quantum field theory and Standard Model Lagrangian . . . 23
1.1.2 Quantum electrodynamics . . . 24
1.1.3 Quantum chromodynamics . . . 25
1.1.4 Electroweak unification . . . 26
1.2 The Higgs mechanism . . . 28
1.2.1 The Brout-Englert-Higgs mechanism . . . 28
1.2.2 The fermion Yukawa couplings . . . 30
1.3 Higher order corrections . . . 31
1.4 Conceptual and experimental tools for probing the Standard Model predictions . . . 32
1.4.1 Cross-section . . . 33
1.4.2 Decay rate and branching ratio . . . 34
1.4.3 Signal strength . . . 34
1.4.4 Simplified template cross-section . . . 35
1.4.5 Coupling-strength modifiers . . . 37
1.4.6 Effective field theory . . . 37
1.5 Event simulation . . . 40
1.6 Higgs boson measurements at the LHC . . . 43
1.7 Limitations of the Standard Model . . . 53

The Standard Model (SM) of particle physics is currently the theory describing the interactions, the characteristics and properties of the elementary particles known to date. It formalizes, with a Lagrangian approach, three of the four fundamental interactions: the electromagnetic, the strong and the weak forces. This theory also allows to explain the origin of the mass of particles due to their coupling to the Higgs boson which is the result of the quantum excitation of the Higgs field. The SM predictions have been intensively and precisely probed experimentally since several decades at colliders including for instance particle discoveries such as the Z and W bosons [13–15] by the UA1 and UA2 experiments at the Super Proton Synchrotron, the top quark [16, 17] at the Tevatron by the CDF and DØ experiments and more recently the Higgs boson [1, 2] by the ATLAS and CMS collaborations with the Large Hadron Collider (LHC).

Despite its highly predictive power, some hints (such as for instance the presence of dark matter in the Universe) indicate that the SM could be an approximation of a more fundamental theory. This led to the development of many alternatives or extensions referred to as Beyond Standard Model (BSM) theories though no experimental evidence is currently favoring one of them nor significant deviations from the SM predictions are observed. Higgs related measurements are an important aspect to probe the SM as for instance dark matter candidates could be heavy particles and thus would strongly couple with the Higgs leading to potential deviations of the SM predictions that could be observed in direct or indirect measurements.

This chapter first describes the SM in Section 1.1 while focusing in particular on the Higgs mechanism in Section 1.2. It then explains, in Section 1.3, why higher order corrections for SM predictions are important to properly predict and simulate experimental results. The Section 1.4 is highlighting some methods used to probe the SM predictions such as performing cross-sections measurements with the simplified template cross-section approach or reinterpreting experimental results in the context of an effective field theory. The Section 1.5 details how the SM Lagrangian can be used to predict experimental results and the different steps needed to obtain a Monte Carlo event simulation. The Section 1.6 presents some Higgs boson measurements performed at the LHC. Finally, the Section 1.7 emphasizes the main limitations of the SM.

1.1 The Standard Model

The fundamental components of the Universe which are presented in Figure 1.1 are called elementary particles and can be divided in two categories, the matter and forces components.

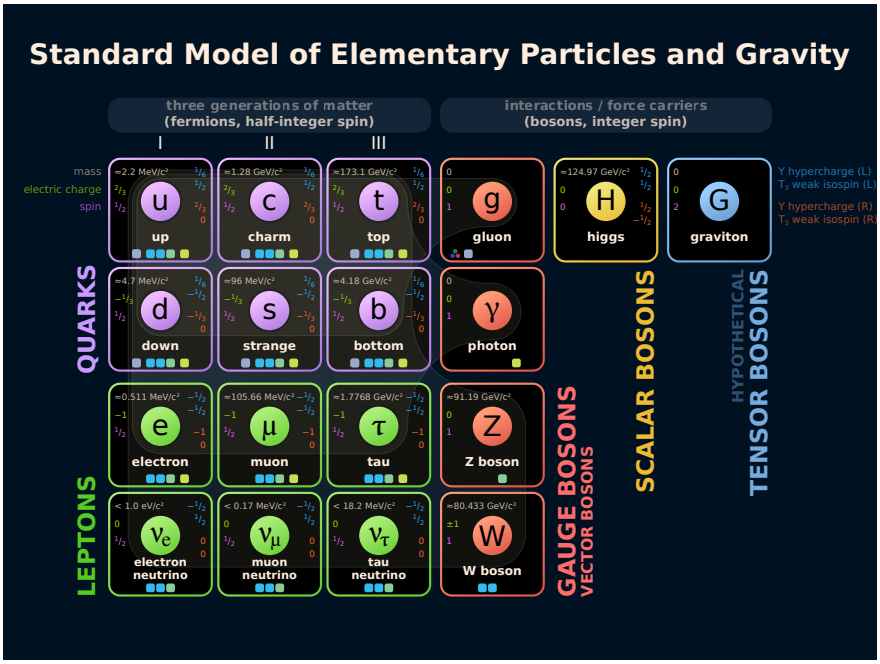
- **The matter components** are composed of 12 spin-1/2 particles referred to as fermions divided into three *generations* (also called *families*) with increasing masses. Each family includes a negatively charged lepton (e^-, μ^-, τ^-) and their corresponding neutrino (neutral lepton) $(\nu_e, \nu_\mu, \nu_\tau)$. The generations are completed with up (u, c, t) and down (d, s, b) type-quarks which are respectively carrying a fractional electric charge of $+2/3$ and $-1/3$. There exist three replicas of each quark that differ by their color charge (R, G, B) corresponding to red, green and blue. The colour charge is a property of the strong interaction formalized by the quantum chromodynamics (QCD) described later in this section. For every matter particle, it exists an antimatter one with all opposite sign charges.
- **The force components**: particles interact thanks to spin-1 vector bosons referred to as *mediators of the interactions* or *force carriers*. The electromagnetic interaction is mediated by the photon (γ, electrically neutral and massless particle), the strong interaction by the gluon (g, neutral massless particle), while the electroweak interaction by 3 massive bosons: the Z^0 boson which is electrically neutral and the $W^\pm$ bosons which carry a ± 1 electric charge. The Higgs boson is the only spin-0 (hence scalar) boson. This neutral massive particle enables to explain the mass of the other elementary particles as presented in the Section 1.2.

The charged leptons are interacting via the electromagnetic and weak forces while neutrinos are only involved in weak interactions as they are electrically neutral. As leptons do not carry a colour charge, they do not interact via the strong interaction. Both the muons and the tau leptons are unstable particles with respective lifetimes of $2.2\ \mu s$ and $290\ fs$ [18], and thus decay into lighter particles. The mass of the neutrinos from the three families and their ordering remains unknown however the observation of neutrino oscillations proved that they are massive particles as predicted by Bruno Pontecorvo [19]. The most stringent limit to date was obtained by the KATRIN collaboration which determined that $m_\nu < 0.8\ eV$ with a 90% confidence level [20] thanks to studies of the tritium β-decay spectrum. The neutrino flavour eigenstates are related to the mass eigenstates by the Pontecorvo-Maki-Nakagawa-Sakata (PMNS) matrix [21].

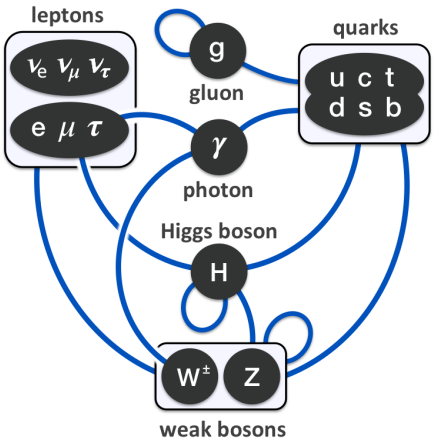
The quarks are the only particles that interact though all three forces. The quark flavour is conserved by the strong and electromagnetic interactions while it is violated for the weak interaction, more precisely it is violated in weak charged current processes *i.e.* mediated by the W boson whereas for weak neutral current decays, *i.e.* mediated by a Z boson, the flavour is conserved. The Cabibbo-Kobayashi-Maskawa (CKM) matrix [22, 23] relates the transition between the mass eigenstates and the weak interaction eigenstates. This CKM matrix is a complex unitary matrix basically providing the probability of a transition between two quarks when the mediator is a W boson: probability which is proportional to $\left|V_{qq'}\right|^2$. This probability is the highest for quarks of the

same family *e.g.* $|V_{ud}| \approx |V_{cs}| \approx |V_{tb}| \approx 1$ while it largely decreases, as summarized in Figure 1.2, for quarks not from the same family. The transition probability is the smallest between the first and third families as for instance $|V_{td}| \approx 0.0085 \approx \lambda^3$ [24] with $\lambda \approx 0.226$ being one of the four Wolfenstein parameters, parameters that fully describe the CKM matrix. Due to its large mass ($m_t \approx 172.5$ GeV) and weak coupling, the top quark decay time is of the order of 5×10^{-25} s which is around 20 times shorter than the time needed for hadronization. As a result, top quarks do not form bound states and always decay to other quarks, predominantly to b-quarks ($t \to W^+ b$) as $|V_{tb}| \approx 0.999$ [24].

Concerning the characteristics of the fundamental interactions, the electromagnetic and (low energy) weak interaction strengths are respectively a factor 10^{-2} and 10^{-13} smaller compared to the strong interaction. The typical lifetime of particles that decay through strong interaction is of the order of 10^{-23} to 10^{-20} s, is ranging from 10^{-20} to 10^{-15} s for the electromagnetic interaction and is spanning over 15 orders of magnitude for the weak interaction as the typical lifetime of B and D mesons is of the order of 10^{-12} s while the neutron lifetime is 878 s [18]. Finally, the range of the strong interaction is typically 1 fm for the strong interaction, it is 10^{-3} fm for the weak interaction (smaller than the proton scale) and is infinite for the electromagnetic interaction.



(a)



(b)

Figure 1.1: (a) Elementary particles known to date as described by the Standard Model along with their properties and (b) known interactions between them: closed loops symbolize a particle self interaction (taken from Refs. [25, 26]).

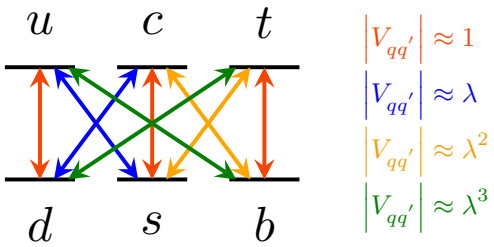


Figure 1.2: Order of magnitude of the absolute value ($|V_{qq'}|$) of the CKM matrix elements as a function of $\lambda \approx 0.226$.

1.1.1 Quantum field theory and Standard Model Lagrangian

The SM is based on a renormalizable quantum field theory (QFT) *i.e.* particles are treated as excited states of their underlying quantum fields, quantum fields which are defined as the minimum (Equation (1.1.4)) of a (density) Lagrangian associated to a particle. Fields are basically operators that create or destroy particles.

More precisely, the treatment of a discrete system of particles depending on n generalized coordinates q_i can be transformed to a continuous system by replacing the Lagrangian (L) with its Lagrangian density ($\mathcal{L}$) [27]:

$$L\left(q_i, \frac{dq_i}{dt}\right) \rightarrow \mathcal{L}\left(\phi_i, \partial_\mu \phi_i\right), \tag{1.1.1}$$

with $\phi_i = \phi_i(t, \boldsymbol{x})$ the continuous fields depending on space and time that replace the generalized coordinates q_i, and

$$\partial_\mu \phi_i := \frac{\partial \phi_i}{\partial x^\mu} \tag{1.1.2}$$

the derivative of those fields with respect to space or time coordinates ($\mu = 0, 1, 2, 3$). The Lagrangian can be retrieved from the spatial integral of the density Lagrangian:

$$L = \int \mathcal{L}\, d^3\boldsymbol{x}. \tag{1.1.3}$$

Following the least action principle, the Euler-Lagrange equation (left-hand side of the equation below) for the Lagrangian is equivalent for the Lagrangian density to the right-hand side of the equation implying constraints on each continuous field ϕ_i

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = 0 \Leftrightarrow \partial_\mu \left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi_i)}\right) - \frac{\partial \mathcal{L}}{\partial \phi_i} = 0. \tag{1.1.4}$$

For simplicity, in the following, the Lagrangian always refers to the density Lagrangian.

For a quantum field theory, the invariance properties of the Lagrangian under a group of continuous local transformations are called local gauge transformations or gauge symmetries. Such gauge invariance implies that the physics of the system considered remain the same under such transformation. From Noether's theorem [28, 29] it follows that a physical quantity is conserved for each local gauge symmetry and vice-versa, for every physical quantity conserved in a system, there exists an associated local gauge symmetry. For instance, if the Lagrangian is respectively invariant under translation in space it translates into momentum conservation of the system, the rotation in space invariance is associated to the kinetic momentum conservation, the time translation invariance to the energy conservation, and if under a change of phase of the system the Lagrangian remains unchanged, then it implies that the electric charge is conserved.

The SM Lagrangian ($\mathcal{L}_{\text{SM}}$) can be expressed as a sum of different terms:

$$\mathcal{L}_{\text{SM}} = \mathcal{L}_{\text{EW}} + \mathcal{L}_{\text{QCD}} + \mathcal{L}_{\text{Yukawa}} + \mathcal{L}_{\text{Higgs}}, \tag{1.1.5}$$

where $\mathcal{L}_{\text{EW}}$ is the Lagrangian associated to the electroweak (EW) interactions *i.e.* the combination of the electromagnetic and the weak interactions, $\mathcal{L}_{\text{QCD}}$ is for the strong interaction and finally $\mathcal{L}_{\text{Yukawa}} + \mathcal{L}_{\text{Higgs}}$ are related to the Higgs mechanism. Each of those terms is going to be discussed in the coming sections. The SM theory is based on the

$$SU(3)_C \times SU(2)_L \times U(1)_Y$$

symmetry group with $SU(3)_C$ the group of 3-dimensional rotations in the colour space associated to the colour symmetry of the strong interaction, and $SU(2)_L \times U(1)_Y$ are associated to the conservation of the weak isospin (T) and hypercharge (Y) for the EW interaction, quantities which are reported for the elementary particles in Figure 1.1a (T_3 is the third component of the weak isospin).

Let's define a simple example that will be useful to understand the following sections. For all spin-1/2 massive particles for which parity is a symmetry, such as fermions, the Dirac equation [30] describes their behaviour for the relativistic QFT:

$$(\not{p} - m)\psi(x) = 0, \tag{1.1.6}$$

with $x = (t, \boldsymbol{x})$ the spacetime coordinates of the particle, m its mass, $\not{p}$ is the Feynman notation equal to $\not{p} := p_\mu \gamma^\mu = i\partial_\mu \gamma^\mu$ adopting the Einstein summation notation, ψ the (four components complex) Dirac spinor of the particle (*i.e.* its wavefunction) and γ^μ the Dirac matrices [27]. By definition for a^μ, b^ν both contravariant vectors then $a^\mu b_\mu := a_\mu b^\mu = g_{\mu\nu} a^\mu b^\nu$ with the diagonal tensor matrix $g_{\mu\nu} = \mathrm{diag}(-1, 1, 1, 1)$ and $\not{a} := a_\mu \gamma^\mu = a^\mu \gamma_\mu$. This equation corresponds to the motion of a free non-interacting particle and is the generalization of the Schrödinger equation as it takes into account relativistic effects. It can be derived as the equation of motion corresponding to the following Lagrangian:

$$\mathcal{L} = \bar{\psi}(\not{p} - m)\psi. \tag{1.1.7}$$

1.1.2 Quantum electrodynamics

The simple example of the free moving particle from the Dirac Equation (1.1.6) and the corresponding Dirac Lagrangian Equation (1.1.7) can be refined in a more realistic scenario to obtain the Lagrangian of the EM interaction involving charged fermions and electromagnetic fields. It can be expressed as the sum of the Dirac Lagrangian and an EM term that are equivalent to the Maxwell equations [31]:

$$\mathcal{L}_{\mathrm{EM}} = \bar{\psi}(\not{p} - m)\psi - \frac{1}{4} F_{\mu\nu} F^{\mu\nu}, \tag{1.1.8}$$

with $F_{\mu\nu} := \partial_\mu A_\nu - \partial_\nu A_\mu$ the EM field strength tensor. The term involving the tensor ψ correspond to the non-interacting spin-1/2 charged fermion while A_μ represents the spin-1 EM carrier (the photon) with $F_{\mu\nu} F^{\mu\nu}$ being the kinetic energy term for the photon. It is possible to show that $\mathcal{L}_{\mathrm{EM}}$ is invariant under the global continuous symmetry:

$$\psi(x) \to \psi'(x) = e^{i\alpha} \psi(x), \tag{1.1.9}$$

where α is a constant phase. This symmetry can be transformed to a local symmetry performing the $U(1)_Q$ transformation

$$\psi(x) \to \psi'(x) = e^{iq\alpha(x)} \psi(x), \tag{1.1.10}$$

with q the charge of the fermion ($q = -|e|$ for an electron) and in that case the phase $\alpha(x)$ is local *i.e.* it can be different at all points in spacetime. To preserve the gauge invariance of the EM Lagrangian from Equation (1.1.8), the derivative of the Dirac term must be replaced by the covariant derivative (D_μ) defined as follows:

$$\partial_\mu \to D_\mu := \partial_\mu + iqA_\mu, \tag{1.1.11}$$

and

$$A_\mu \to A'_\mu = A_\mu - \partial_\mu \alpha, \tag{1.1.12}$$

which leads to the quantum electrodynamics (QED) Lagrangian [27]:

$$\mathcal{L}_{\mathrm{QED}} = \bar{\psi}(\not{p} - m)\psi - \frac{1}{4} F_{\mu\nu} F^{\mu\nu} - q\bar{\psi}\not{A}\psi = \mathcal{L}_{\mathrm{EM}} - q\bar{\psi}\not{A}\psi., \tag{1.1.13}$$

This Lagrangian contains the EM Lagrangian term where particles are free and an additional term coming from the covariant derivative which encodes the interaction between two Dirac spinors *i.e.* two fermions and a photon. It can be interpreted in terms of Feynman diagrams with a vertex and a propagator [32], and it formalizes the EM interactions, the coupling strength of the interaction being the charge q.

1.1.3 Quantum chromodynamics

The quantum chromodynamics is a non-abelian[1] gauge theory that describes the strong interaction. It allows to explain the observed spectrum of hadrons and was developed after the quark model. The QCD theory is based on a $SU(3)$ gauge symmetry associated to the colour charge conservation (red, green, blue) for strong interaction processes. The associated local $SU(3)_C$ phase transformation is [27, 33]

$$\psi(x) \rightarrow \psi'(x) = e^{ig_S \boldsymbol{\alpha}(x)\cdot \boldsymbol{T}}\psi(x), \tag{1.1.14}$$

with g_S the strong interaction coupling constant, $\boldsymbol{T}$ the eight generators of the $SU(3)$ symmetry group which are related to the Gell-Mann matrices by $T^a = \lambda^a/2$ $(a = 1, \ldots, 8)$ and $\alpha^a(x)$ are eight functions of the spacetime coordinate x. Since $SU(3)$ is represented by 3×3 matrices, the wavefunction ψ must now include the 3 charge color degrees of freedom. The local phase transformation from Equation (1.1.14) corresponds to a rotation in the colour space around an axis whose direction is different for every point in spacetime.

The associated QCD Lagrangian for quarks of mass m_q can be written as:

$$\mathcal{L}_{\text{QCD}} = -\frac{1}{4}G^a_{\mu\nu}G^{\mu\nu}_a + \sum_q \bar{\psi}_{qj}(i\not{D} - m_q\delta_{jk})\psi_{qk}, \tag{1.1.15}$$

with

$$D_\mu := \partial_\mu\delta_{jk} + ig_S(T_aA^a_\mu)_{jk}, \tag{1.1.16}$$

$$G^a_{\mu\nu} := \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + g_S f^{abc}A^b_\mu A^c_\nu, \tag{1.1.17}$$

and ψ_{qk} is the Dirac spinor associated to the quark q, k represents its colour charge, δ_{jk} is the function imposing the colour charge to be conserved in the free propagation of quarks (this function is implied in Equation (1.1.15)), D_μ is the covariant derivative, A^a are the gluon fields and f^{abc} represents the structure constant of the $SU(3)$ group (a, b and c are index running from 1 to 8 and correspond to different color states of the gluons). As a result only quarks carrying a colour charge interact via the strong force through the exchange of gluons.

$G^a_{\mu\nu}G^{\mu\nu}_a$ is the kinetic term of the gluons. When expanded using the $G^a_{\mu\nu}$ definition from Equation (1.1.17), 3 terms appear in the $G^a_{\mu\nu}G^{\mu\nu}_a$ product which are respectively interpreted as the free propagation of gluons, the three gluon self-interaction vertex and the four gluon interaction vertex also referred to as trilinear and quartic self interactions. This is a main difference with respect to QED: unlike photons, gluons can interact with themselves.

Therefore, there are in total 8 gluons which are massless colour charged particles. For short distances (equivalently high energies $E \gg \Lambda_{QCD} \approx 200$ MeV), interactions between quarks and gluons are small and the particles can be considered as asymptotically free: it is the perturbative regime. In that case, the prediction of the QCD Lagrangian from Equation (1.1.15) can be treated with a perturbative approach *i.e.* thanks to approximations which are referred to as perturbative QCD (pQCD). However, for lower energies the interaction between partons largely increases and cannot be treated with the pQCD as the coupling between quarks (and gluons) becomes large: this principle is called the confinement of quarks and gluon hence explaining that quarks are never observed alone and are indeed confined inside hadrons (either mesons or baryons respectively composed of two and three valence quarks and/or antiquarks) that are colourless combinations of quarks and antiquarks. In that regime, the potential energy between quarks is proportional to their distance. At some point, as the separation between them increases, the potential energy reaches the pair of quark-antiquark (and gluons) production threshold. Such pair is hence created and form new hadrons from their combination with the initial quarks as schematized in Figure 1.3. Such combination process is called the hadronization and is the reason why only jets are observed at

1. A non-abelian gauge symmetry group is a group for which gauge transformations do not commute

colliders instead of single hadron particles as at high energy hadrons fragment into several less energetic hadrons and so on forming a collimated stream of hadrons.

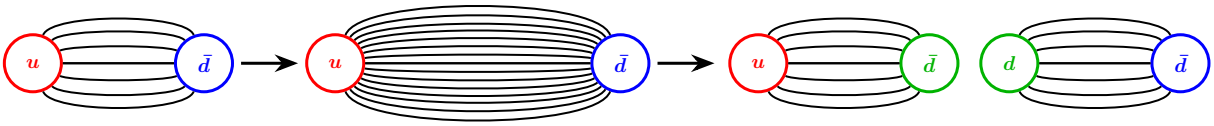


Figure 1.3: Scheme of a quark-antiquark splitting forming two new hadrons thanks to a quark-antiquark pair creation. This representation is called the *Lund string model* where the interactions (via strong force mediated by gluons) are symbolized by flux tubes. As the separation between the initial quarks increases (middle scheme), the potential energy increases and eventually reaches the quark-antiquark pair creation threshold.

1.1.4 Electroweak unification

A common theory for the electromagnetic and weak interactions, referred to as electroweak unification, was formalized by Sheldon Glashow, Steven Weinberg and Abdus Salam [34–36] (also known as the GWS model). It is based on the local $SU(2)_L \times U(1)_Y$ gauge symmetry. The electroweak unification allows to describe the two different types of weak interaction which are respectively the charged current interactions such as, for instance, β decays and are mediated by the (charged) $W^{\pm}$ bosons while on the other hand the neutral current interactions are mediated by the (neutral) Z boson. Flavour violation only occurs for charged current interactions while neutral current interaction were observed to conserve this quantity which led to postulate the existence of 2 different bosons.

Experimentally the weak interaction is found to be strongly dependent on the chirality of particles. Particles can either be left- or right-handed. For massless (or ultra-relativistic *i.e.* $E \gg m$) particles the chirality is equivalent to the helicity which is the sign of the projection of the spin vector on the momentum vector of the particle: right helicity is positive while left helicity is negative. It was observed that the weak charged current interaction, *i.e.* interactions with a W boson, only couples left-handed particles to right-handed antiparticles. The left-handed fermions are grouped in weak isospin 1/2 doublets *e.g.*

$$\begin{pmatrix} \nu_e \\ e^- \end{pmatrix}_L, \quad \begin{pmatrix} u \\ d \end{pmatrix}_L,$$

while the right-handed particles form singlets. On the other hand, the weak neutral current interactions, *i.e.* with a Z boson, couple both left-handed and right-handed particles. It was also found that the weak interaction preserves the weak isospin (T) and the weak hypercharge (Y) which are related to the electric charge (Q) by the Gell-Mann-Nishijima formula [37, 38]:

$$Y = 2(Q - T_3), \tag{1.1.18}$$

with T_3 the third component of the weak isospin.

Since the two quantities Y and T are conserved it follows from the Noether's theorem that there exists 2 gauge symmetries in the Lagrangian describing the electroweak unification [33, 39]. The $SU(2)$ transformation associated to the weak charged current interaction is related to a weak isospin rotation

$$\psi(x) \rightarrow \psi'(x) = e^{ig\boldsymbol{\alpha}(x)\cdot\boldsymbol{T}}\psi(x), \tag{1.1.19}$$

where g is the W boson coupling constant, $\boldsymbol{\alpha}(x) = (\alpha_1(x), \alpha_2(x), \alpha_3(x))$ is a vector of three real parameters, and $\boldsymbol{T} = \boldsymbol{\sigma}/2$ with $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)$ the vector of Pauli matrices. The local gauge

invariance requirement can only be satisfied introducing three gauge fields, W_μ^k with $k = 1, 2, 3$ corresponding to three gauge bosons denoted W^1, W^2 and W^3. Since the three generators ($\boldsymbol{T}$) of $SU(2)$ are 2×2 matrices, the wavefunction ψ should have 2 components. $\psi(x)$ is thus a complex scalar of $SU(2)$ referred to as the weak isospin doublet

$$\psi(x) = \begin{pmatrix} \psi_1(x) \\ \psi_2(x) \end{pmatrix}, \tag{1.1.20}$$

with $(\psi_1(x), \psi_2(x)) \in \mathbb{C}^2$. Moreover, as the weak charged current interaction couples together different fermions (due to the observed flavour violation), the weak isospin doublet should be composed of different flavours that are differing by one unit in terms of electric charge since the interaction mediator is a $W^\pm$ boson. As a result, $\psi(x) = (\nu_e(x), e^-(x))$ for instance and the gauge transformation from Equation (1.1.19) should only affect left-handed particles and right-handed antiparticles to respect the experimental observations. This explains the notation $SU(2)_L$ as right-handed particles and left-handed antiparticles are placed in singlets with $T = 0$ *i.e.* they are unaffected by the $SU(2)$ transformation and do not couple to the W boson.

The charged current can then be expressed as a linear combination of W_μ^1, W_μ^2 associated to the $W^\pm$ bosons. However, since the Z boson couples left- and right-handed particles it cannot correspond to the W^3 boson of the $SU(2)_L$ local gauge symmetry. The exact same statement holds for the photon. In the GSW model, the $U(1)$ gauge symmetry from QED is replaced by the $U(1)_Y$ symmetry with the hypercharge transformation which is simply a local phase change:

$$\psi(x) \to \psi'(x) = e^{i\boldsymbol{\beta}(x)Y}\psi(x), \tag{1.1.21}$$

with $\boldsymbol{\beta}(x)$ a real parameter. To the $U(1)_Y$ symmetry is associated a field denoted B_μ.

The corresponding Lagrangian of the EW interaction, for massless particles, can be written as:

$$\mathcal{L}_{EW} = i\bar{\psi}\not{D}\psi - \frac{1}{4}W_{\mu\nu}^a W_a^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu}, \tag{1.1.22}$$

where

$$D_\mu := \partial_\mu + igW_\mu^a T^a + \frac{i}{2}g'B_\mu Y, \tag{1.1.23}$$

$$W_{\mu\nu}^a := \partial_\mu W_\nu^a - \partial_\nu W_\mu^a - g\varepsilon^{abc}W_\mu^b W_\nu^c, \tag{1.1.24}$$

$$B_{\mu\nu} := \partial_\mu B_\nu - \partial_\nu B_\mu, \tag{1.1.25}$$

with ε^{abc} the totally antisymmetric tensor and $a = 1, 2, 3$.

In total, the combination of $SU(2)_L \times U(1)_Y$ implies that the 4 gauge fields W_μ^1, W_μ^2, W_μ^3 and B_μ describe the electroweak interaction. It is possible to show that with a rotation of those fields such as presented below in Equation (1.1.26), the fermions have an electromagnetic-like interaction (affecting both the left-handed and right-handed spinors, with coupling proportional to the charge) with the A_μ field, which is thus interpreted as the photon, while neutral and charged currents involving only left-handed particles couple to the $W^\pm$ and Z^0 bosons. All four mediator bosons are massless in this theory.

$$\begin{pmatrix} W_\mu^+ \\ W_\mu^- \\ Z_\mu \\ A_\mu \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} & 0 & 0 \\ 0 & 0 & \cos(\theta_W) & -\sin(\theta_W) \\ 0 & 0 & \sin(\theta_W) & \cos(\theta_W) \end{pmatrix} \begin{pmatrix} W_\mu^1 \\ W_\mu^2 \\ W_\mu^3 \\ B_\mu \end{pmatrix}, \tag{1.1.26}$$

where θ_W is the weak mixing angle (also known as the Weinberg angle) measured to be $\theta_W \approx 30°$ [40] and which is related to the electromagnetic coupling constant (e which is also the charge of the electron) by

$$e = g\sin(\theta_W) = g'\cos(\theta_W) \tag{1.1.27}$$

with g and g' respectively the coupling constant of the $SU(2)_L$ and $U(1)_Y$ local gauge transformations.

The electroweak unification is valid at high energies where the B_μ and the W^a_μ bosons can be considered as massless as otherwise the renormalization of the theory would be spoiled [33]. However, at low energy the electroweak symmetry is spontaneously broken leading to a decoupling of the weak and electromagnetic forces. The weak interaction Z and W bosons carriers are endowed masses while the photon which is the electromagnetic carrier remains massless: such phenomenon is called the Higgs mechanism and is described in the next section.

1.2 The Higgs mechanism

The electroweak theory and the EW Lagrangian presented in Equation (1.1.22) are only valid assuming massless mediator particles as otherwise the gauge invariance would be spoiled and would also break the renormalization of the theory. As a result, the mass of the W and Z bosons have to be explained with another mechanism. Similarly, the observed mass of the other elementary particles, *i.e.* all the fermions, is incompatible with the gauge invariance of the EW Lagrangian.

In 1964, the *Brout-Englert-Higgs mechanism* [41–44] was one of the theories developed which allows for a spontaneous breaking of the $SU(3)_C \times SU(2)_L \times U(1)_Y$ symmetry group which then becomes $SU(3)_C \times U(1)_Q$ conferring masses to the weak bosons while the photon remains massless as it is associated to the $U(1)_Q$ group symmetry. Moreover, in this theory, the masses of the fermions are explained via the Yukawa interaction of the fermions with the Higgs field. Such theory was then tested and is still under study at the LHC at CERN where a particle with a mass of $125\ \mathrm{GeV}$ presenting the expected properties of the Higgs boson was observed by the ATLAS and CMS experiments thus proving the validity of the Higgs mechanism theory and completing that important missing part of the SM.

1.2.1 The Brout-Englert-Higgs mechanism

The Brout-Englert-Higgs (BEH) mechanism introduces a scalar Higgs term to the SM Lagrangian

$$\mathcal{L}_{\text{Higgs}} = (D^\mu\phi)^\dagger(D_\mu\phi) - V(\phi), \tag{1.2.1}$$

where ϕ is 2-dimensional complex scalar (Higgs) field

$$\phi = \begin{pmatrix}\phi^+ \\ \phi_0\end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix}\phi_1 + i\phi_2 \\ \phi_3 + i\phi_4\end{pmatrix}, \tag{1.2.2}$$

D_μ is the covariant derivative defined in Equation (1.1.23). The term $(D^\mu\phi)^\dagger(D_\mu\phi)$ is the kinetic term of the Lagrangian while $V(\phi)$ is the Higgs potential defined as

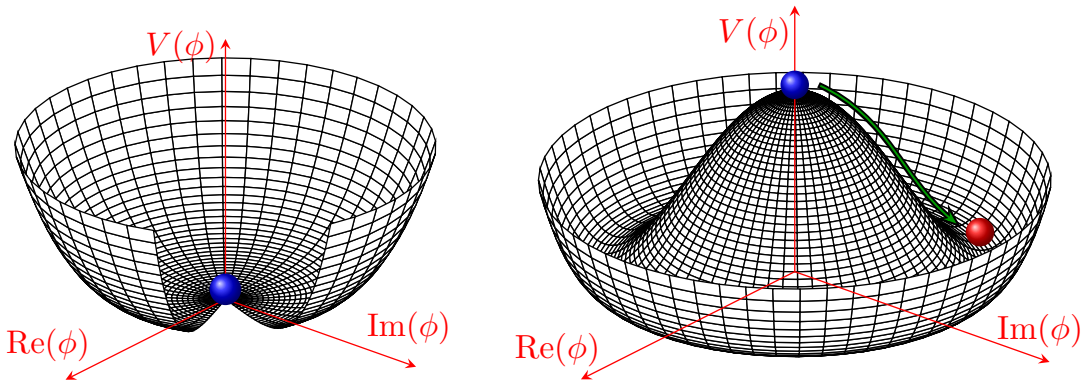
$$V(\phi) := \mu^2\phi^\dagger\phi + \lambda(\phi^\dagger\phi)^2 = \mu^2\left|\phi\right|^2 + \lambda\left|\phi\right|^4, \tag{1.2.3}$$

where μ^2 and λ are two constant real numbers (μ can be complex). The μ^2 part is a mass-type term while the λ part describes the quadratic self-interaction among the scalar fields. The vacuum state is the lowest energy state of the field ϕ and corresponds to the minimum of the potential $V(\phi)$. Indeed, in order to have an energetically bounded theory *i.e.* a finite minimum for the potential, λ

must be positive while μ^2 can either be positive or negative which affects the shape of the potential as shown in Figure 1.4. If $\mu^2 \geq 0$ then $V(\phi)$ has only one minimum in $\phi = 0$ while if $\mu^2 < 0$ then $\phi = 0$ is not a minimum and the potential has an infinite set of minima corresponding to:

$$|\phi|^2 = -\frac{\mu^2}{2\lambda} = \frac{v^2}{2}, \tag{1.2.4}$$

where v is called the vacuum expectation value (abbreviated vev).



(a) (b)

Figure 1.4: Higgs potential $V(\phi) = \mu^2 |\phi|^2 + \lambda |\phi|^4$ with $\lambda > 0$ and either (a) $\mu^2 \geq 0$ or (b) $\mu^2 < 0$ (schemes highly inspired by Refs. [45, 46]).

The spontaneous $SU(2)_L \times U(1)_Y$ symmetry breaking can only be achieved in the scenario where $\mu^2 < 0$: in that case nature "is choosing" a particular vacuum (ground) state out of all possible states minimizing the potential hence spontaneously breaking the symmetry as shown in Figure 1.4b. There exists a given basis for which the chosen vacuum state ϕ_0 can be expressed as:

$$\phi_0 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}. \tag{1.2.5}$$

It is then possible to expand the Higgs field around its ground state as:

$$\phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}, \tag{1.2.6}$$

with $h(x)$ a real scalar field. Injecting such field in the kinetic term from Equation (1.2.1), one can identify the kinetic terms associated to the W, Z bosons and photons hence deducing their mass as predicted by the BEH mechanism which leads to the following equalities [47]

$$m_W = \frac{gv}{2}, \quad m_Z = \frac{\sqrt{g'^2 + g^2}v}{2}, \quad m_\gamma = 0, \quad \frac{m_W}{m_Z} = \cos(\theta_W), \tag{1.2.7}$$

and $v \approx 246$ GeV is measured in muon decays. No kinetic term concerning the field A_μ is found as expected for the photon which is thus predicted to be massless. Moreover, the other terms depending on the Higgs scalar field $h(x)$ in the expansion of the covariant derivative correspond to the interaction between the Higgs and the gauge bosons:

$$\mathcal{L}_{VVH} = \frac{2m_W^2}{v} W_\mu^+ W^{-\mu} h(x) + \frac{2m_Z^2}{v} Z_\mu Z^\mu h(x), \tag{1.2.8}$$

$$\mathcal{L}_{VVHH} = \frac{2m_W^2}{v^2} W_\mu^+ W^{-\mu} h^2(x) + \frac{2m_Z^2}{v^2} Z_\mu Z^\mu h^2(x). \tag{1.2.9}$$

Finally, injecting the field expansion in the potential $V(\phi)$, the Higgs Lagrangian becomes (taking only into account terms involving h):

$$\mathcal{L}_H = \mu^2 h^2 - \lambda v h^3 - \frac{\lambda}{4} h^4 = \frac{1}{2} m_H^2 h^2 - \sqrt{\frac{\lambda}{2}} m_H h^3 - \frac{\lambda}{4} h^4, \tag{1.2.10}$$

and so by identification the mass of the Higgs is equal to

$$m_H = \sqrt{2}\,|\mu| = \sqrt{2\lambda} v. \tag{1.2.11}$$

The Higgs boson mass is not predicted nor constrained by the Higgs mechanism theory: it is a free parameter which needs to be determined experimentally ($\lambda \approx 0.13$ [47] experimentally). The corresponding interaction vertices associated to the terms found in the Lagrangians $\mathcal{L}_H$, $\mathcal{L}_{VVH}$ and $\mathcal{L}_{VVHH}$ are represented in Figure 1.5 with as well the interaction coupling factors. All coupling factors are hence found to be proportional to the mass squared of the Higgs boson or vector bosons ($V = W, Z$).

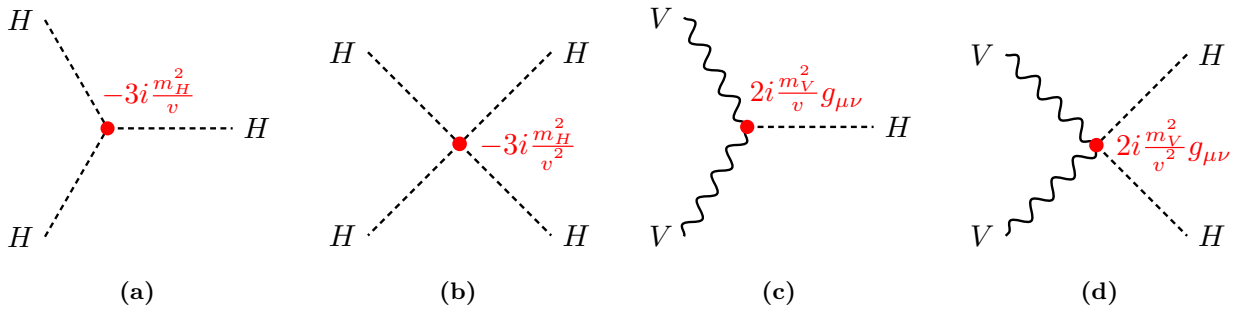


Figure 1.5: Leading order interaction vertices of the Higgs as predicted by Brout-Englert-Higgs mechanism. (a) Trilinear and (b) quartic Higgs self-coupling respectively as predicted by the h^3 and h^4 terms of the $\mathcal{L}_H$ Lagrangian. (c) and (d) interaction of the Higgs with the vector boson ($V = W, Z$) in the $\mathcal{L}_{VVH}$ and $\mathcal{L}_{VVHH}$ Lagrangians. The values of the coupling factors reported (in red) are obtained from the Lagrangian themselves, but their signs are taken from Ref. [48].

1.2.2 The fermion Yukawa couplings

The BEH mechanism allows to explain the mass of the W and Z bosons but cannot explain the fermion masses. As a result, an additional term in the SM Lagrangian needs to be introduced. It is so-called *Yukawa* interaction Lagrangian defined as:

$$\mathcal{L}_{\text{Yukawa}} = -y_f(\bar{\psi}_L \phi \psi_R + \bar{\psi}_R \phi^\dagger \psi_L), \tag{1.2.12}$$

with y_f the Yukawa coupling for the fermion f, ψ_R and ψ_L are the singlet right- and doublet left-handed fermion isospins and ϕ is again the Higgs complex scalar field. For instance in the case of the electron $\psi_R = e_R$ and $\psi_L = (\nu_e, e)_L$. For an electron, injecting the Higgs field expansion from Equation (1.2.6) in the Lagrangian leads to:

$$\mathcal{L}_e = -\frac{y_e v}{\sqrt{2}}(\bar{e}_L e_R + \bar{e}_R e_L) - \frac{y_e h}{\sqrt{2}}(\bar{e}_L e_R + \bar{e}_R e_L). \tag{1.2.13}$$

The Yukawa coupling of the electron (y_e) is not predicted, but it is determined interpreting the multiplicative factor of the first term as the mass (m_e) of the electron while the second term corresponds to the coupling (g_{eeH}) between the electron and the Higgs boson itself:

$$m_e = -\frac{y_e v}{\sqrt{2}}, \quad g_{eeH} = -\frac{y_e}{\sqrt{2}} = \frac{m_e}{v}. \tag{1.2.14}$$

The coupling of the leptons with the Higgs is hence proportional to their mass and the mechanism is the same for quarks. The Figure 1.6 presents the Feynman diagram of the Higgs-to-fermion interaction vertex.

The formalism described above only provides the masses to the lower component of the doublet *i.e.* only the mass of the charged leptons and the down-type quarks are explained. For the mass of the up-type quarks, the introduction of the Hermitian conjugate of the Higgs field ϕ^c is required:

$$\phi^c = -i\sigma_2\phi^* = -\frac{1}{\sqrt{2}}\begin{pmatrix} v+h \\ 0 \end{pmatrix}. \tag{1.2.15}$$

The same transformation properties for the conjugate Higgs field are obtained as for the Higgs field which introduces the masses of the up-type quarks.

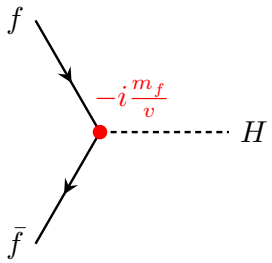


Figure 1.6: Leading order fermion couplings to the Higgs and interaction vertex factor as predicted by the Yukawa theory.

1.3 Higher order corrections

The theory of the Standard Model explained so far does not include corrections *i.e.* only simple Feynman diagrams, referred to as Born or tree level diagrams, are taken into account. However, such assumption is not valid as in reality radiation, vertex, propagator and box corrections need to be considered to have a correct description and predictions of physics processes, corrections which are basically originating from quantum vacuum fluctuations and probabilistic effects. Applying such corrections is not trivial for instance when computing cross-sections from the matrix element amplitudes divergences may appear as it requires to integrate the additional terms over all possible momenta which can result in infinite values for low or high energies (called respectively infrared or ultraviolet divergences). To circumvent this issue, the so-called *renormalization* techniques are used to basically deal with infinite quantities adding terms to the Lagrangian which cancels out those infinite values when computing matrix element amplitudes while improving the predictions.

Mathematically there is an infinite number of additional loop corrections which can be seen as a "Taylor expansion" of the prediction to improve the accuracy of the theory. For instance the cross-section of a process can be written:

$$\sigma = \sigma_0\left(1 + \sum_{i=0}^{+\infty}\alpha_s^i\right), \tag{1.3.1}$$

with σ_0 the tree-level cross-section and α_s^i are additional corrections terms with $i = 0$ corresponding to the leading order (LO) correction, $i = 1$ the next-to-leading-order (NLO) term which is a smaller correction than the LO one, $i = 2$ the next-to-next-to-leading-order (NNLO) and so on. For order of corrections above the NLO or NNLO level, computations are more and more complex as the number of extra loop terms to take into account growths exponentially. As a result, usually

corrections are derived with an NLO or NNLO precision. The (NLO/LO, NNLO/NLO...) *k-factor* quantifies the changes between two consecutive orders of calculation and can for instance change the value of the cross-section of a process by 10 to 20% for NLO corrections while those corrections tend to zero for higher orders. It is hence important to apply such corrections when using simulated samples.

Since counter balancing terms are added to the Lagrangian for renormalizable theories (*e.g.* EW, QCD, QED) to cancel divergences originating from loop computations beyond the LO, those terms needs to be reinterpreted by modifying the definition of the Lagrangian constant terms (such as the electric charge, masses and coupling constants). Such modifications imply that the predictions become dependent on the energy scale (μ) of the processes considered. The previously constant terms are now depending on μ as for instance the strong coupling constant $\alpha_s = \alpha_s(\mu^2)$. Those couplings are referred to as *running coupling constants* since they evolve with μ. For instance for one-loop corrections only, then the strong coupling constants can be expressed as

$$\alpha_s(\mu^2) = \frac{b_0}{\ln\left(\frac{\mu^2}{\Lambda^2_{\text{QCD}}}\right)}, \tag{1.3.2}$$

with b_0 a constant of the theory, $\Lambda_{\text{QCD}} \approx 200$ MeV is the energy scale for QCD and corresponds to the infrared cutoff *i.e.* an estimate for the energy threshold below which perturbative QCD is not valid anymore (perturbative QCD is not valid below 1 GeV). With the renormalization of the theories, the QCD and weak interaction running coupling constants are found to be decreasing with the energy scale while the QED constant increases as shown in Figure 1.7.

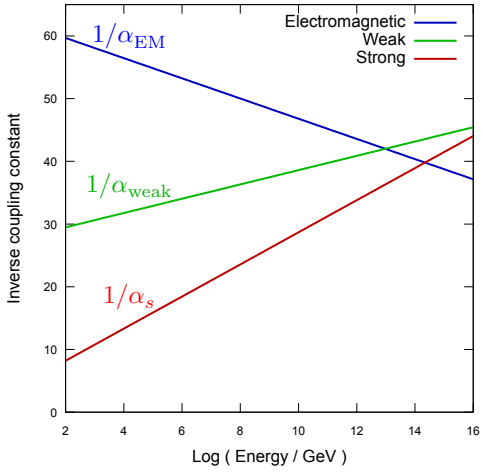


Figure 1.7: Evolution of the inverse of the electromagnetic (α_{EM}), the weak (α_{weak}) and the strong (α_s) interaction running coupling constants as a function of the energy scale (μ) as predicted by the Standard Model (slightly modified figure taken from Ref. [49]).

1.4 Conceptual and experimental tools for probing the Standard Model predictions

The SM Lagrangian allows for (probabilistic) predictions of the behaviour of particles and enables to deduce physical quantities which can be measured experimentally such as the decay rate (Γ) and the cross-section (σ) or other properties which can be specific to the process under study such as the angular distributions of particles, their polarization and so on. This section briefly describes how

some of those physical quantities (non-exhaustive list) can be obtained from the SM Lagrangian and compared to experimental results to probe the validity of the SM.

1.4.1 Cross-section

To calculate the cross-section of a process $ij \to k$, where two particles i and j interact with each other to produce a given final state particle k, then two quantities need to be computed: the so-called *amplitude*, which contains the full dynamical information of the process, the second information being the phase space available which is purely kinematic. The cross-section of a process can be interpreted as the probability of a process to occur.

The matrix element amplitudes are computed with Fermi's golden rules thanks to the Feynman diagrams of the process, with the propagators and the vertex coupling constants coming from the Lagrangian:

$$\sigma_{ij\to k} \propto \int \left|\mathcal{M}_{ij\to k}(\mathbf{\Phi_k})\right|^2 d\mathbf{\Phi_k}, \tag{1.4.1}$$

with $\mathbf{\Phi_k}$ the phase space allowed for the final state k, $\left|\mathcal{M}_{ij\to k}\right|$ is the transition amplitude which is proportional to the so-called matrix element. The matrix element is often calculated perturbatively *i.e.* including additional corrections as explained in Section 1.3.

For hadron colliders, the composite nature of hadrons needs to be taken into account as indeed the collisions occur between partons from the hadrons (either valence quarks, or from the sea of quarks and gluons) and not between hadrons themselves due to the large energy. The momentum of those partons is unknown as they only carry a fraction of the hadron momentum. To circumvent that lack of knowledge, a probabilistic approach was developed: *probability density functions* called *parton distribution functions* (PDFs) predicting the momenta distribution of the partons inside hadrons are used instead as shown in Figure 1.8. The PDFs ($f_i^A(x, Q^2)$) are functions of the fractional momentum (x) carried by a parton i with respect to the momentum of the mother particle A and of Q^2 which is the square of the momentum transferred during the collision between the two interacting partons. Such PDFs cannot be derived analytically since in the low energy regime the QCD theory is not well understood, but they can be measured experimentally by studying for instance Deep Inelastic Scattering (DIS) collisions. Extrapolation to other energy scale Q are possible thanks to the DGLAP equations [50].

For hadronic interaction cross-sections, the QCD factorization theorem states that the hadronic cross-section $\sigma_{AB\to k}$ (with A, B two colliding hadrons) is a convolution of the partonic cross-sections $\sigma_{ij\to k}$ whose PDFs are denoted respectively $f_i^A(x_1, Q^2)$ and $f_j^B(x_2, Q^2)$:

$$\sigma_{AB\to k} = \int \sum_{i\in A, j\in B} f_i^A(x_1, Q^2) f_j^B(x_2, Q^2) \sigma_{ij\to k}\, dx_1 dx_2, \tag{1.4.2}$$

where the sum runs over all partons from the hadrons A and B that can result in the $ij \to k$ process.

With increasing momentum of the proton *i.e.* increasing Q^2, the momentum fraction of sea quarks and gluons increases much more than for the valence quarks as shown in Reference [52]. As a result, processes involving gluons in the initial state are dominant as for instance $t\bar{t}$ events are predominantly produced by gluon fusion at the LHC, but also the main production mode of the Higgs is due to gluon fusion (so called ggH mechanism $gg \to H$) whose Feynman diagram is presented in Figure 1.13a.

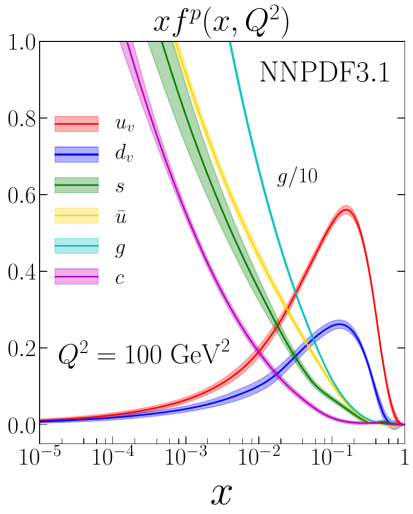


Figure 1.8: Parton distribution functions for partons inside the proton as a function of the parton momentum fraction x as predicted by the next-to-next-to-leading-order (NNLO) NNPDF3.1 for an energy scale of $Q = 10$ GeV. u_v and d_v refers to the PDFs of the up and down valence quarks of the proton while the other PDFs are for partons from the sea of quarks and gluons showing that valence partons carry a larger momentum fraction compared to the other partons (taken from Ref. [51]).

1.4.2 Decay rate and branching ratio

The decay rate (or decay width) of a process (Γ) is the probability per unit of time that a given particle will decay through that process. The lifetime of a particle can be expressed as ($\tau = 1/\Gamma_{\text{tot}}$). The computation of the decay rate is similar to the cross-section. Since usually several final states are possible, the computation are performed separately for each case. The total width of the particle is obtained from the sum of all the individual decays. The probability for a particle (denoted A) to decay into a specific final state X is the so-called *branching ratio* ($\mathcal{B}$) defined as the ratio of the partial decay width of the specific final state (X) over the total decay width of the particle (A) under study:

$$\mathcal{B}(A \to X) = \frac{\Gamma(A \to X)}{\sum_i \Gamma_i(A \to X)}. \tag{1.4.3}$$

1.4.3 Signal strength

Measuring the decay width of a process is technically difficult at hadron colliders as for instance to measure the branching ratio for the Higgs decay to a pair of b-quarks ($H \to b\bar{b}$) at the LHC requires first to know the cross-section $\sigma(pp \to H)$ which can only be assessed through theoretical assumptions.

Let's take a concrete example the so called $VH, H \to b\bar{b}$ processes where a Higgs boson is produced in association with a (weak) vector boson $V = W, Z$ (the VH Feynman diagram is shown in Figure 1.13c) and the Higgs boson decays to a pair of b-quarks. Such processes, which are either quark-initiated ($qq' \to WH$ and $q\bar{q} \to ZH$) or gluon initiated ($gg \to ZH$), are of interest to probe the coupling of the Higgs to the bottom quark, and will be extensively presented and studied in the Chapter 5 and 6 of this thesis. The probability of observing $VH, H \to b\bar{b}$ events is thus proportional to $\sigma_{qq/gg \to VH} \cdot \mathcal{B}_{H \to b\bar{b}}$ where $\sigma_{qq/gg \to VH}$ is the inclusive cross-section for quark and gluon initiated VH processes and $\mathcal{B}_{H \to b\bar{b}}$ is the branching ratio for the $H \to b\bar{b}$ decays (the missing

multiplicative term being related to the luminosity and the signal acceptance). As a result:

$$\sigma_{qq/gg\to VH} \cdot \mathcal{B}_{H\to b\bar{b}} \propto \sigma_{qq/gg\to VH} \cdot \frac{\Gamma_{H\to b\bar{b}}}{\Gamma_H}, \tag{1.4.4}$$

with $\Gamma_{H\to b\bar{b}}$ is the Higgs decay width for $H \to b\bar{b}$ events and Γ_H the total Higgs width.

From Equation (1.4.4), a comparison of the observed measurements of $VH, H \to b\bar{b}$ processes with respect to the predictions of the SM can be performed thanks to the *signal strength* (μ) defined as ratio of observed versus SM predicted probabilities:

$$\mu := \frac{\sigma \cdot \mathcal{B}}{\sigma_{\text{SM}} \cdot \mathcal{B}_{\text{SM}}}. \tag{1.4.5}$$

The measurement of the signal strength eventually boils down to deriving the ratio of the number of $VH, H \to b\bar{b}$ signal events found in data compared to the one predicted with event simulation using the SM. If the signal strength is compatible with unity within uncertainties then the experimental observations are compatible with the SM predictions otherwise if μ is significantly smaller (respectively greater) than unity it implies that the expected number of events as predicted by the SM is larger (respectively smaller) than observed in data which would indicate possible new physics phenomena that would either mimic, enhance or suppress some processes.

1.4.4 Simplified template cross-section

The simplified template cross-section (STXS) framework [53–55] has been developed in order to perform differential cross-section and signal strength measurements *i.e.* dividing phase space usually as a function of the transverse momentum (p_{T}) of a particle from the process under study and/or additionally as the number of jets/particles in the final state as shown in Figure 1.9. Such division of phase space enables to enhance sensitivity since the SM for instance could only be a theory valid in the low energy limit hence deviation from the SM predictions could be expected in high transverse momentum regions that correspond to small parts of the total phase space. For example for the $VH, H \to b\bar{b}$ STXS measurement presented in Section 5.9.1.2, the phase space is among other criteria divided as a function of the transverse momentum (p_{T}^V) of the vector boson. For other Higgs STXS measurements generally the phase space is divided as a function of the transverse momentum of the Higgs (p_{T}^H) and possibly other criteria.

There exist several STXS *schemes* also referred to as *stages* which are basically different "versions" for the division of the phase space *e.g.* the stage 0 bin was used for the Run 1 of the LHC, stage 1.0 and 1.1 for the full Run 2 measurements and currently the stage 1.2 which is an update of the stage 1.1 is used for the Run 2 legacy STXS measurements. The differences between those versions is the number of separately measured bins which has evolved over time and became more fine-grained as the size of the available dataset collected at the LHC has increased.

The STXS formalism is essential as the combination of different results associated to different decays of the Higgs boson can easily be performed thanks to the division of phase space into mutually exclusive regions based only on properties of the Higgs production (and not relying on the Higgs decay products) which are referred to as STXS bins. It also allows for combination of results between different collaborations such as ATLAS and CMS. In the context of a signal strength measurement, only the signal normalization (controlled by μ) can be changed while the shape of the distribution(s) of interest are set to the SM predictions. On the other hand, in the context of a STXS measurement, only the sum of the templates is forced to be identical to the SM prediction while measurements in each bin can show some deviations from the SM.

As a summary, the STXS framework is aiming at:

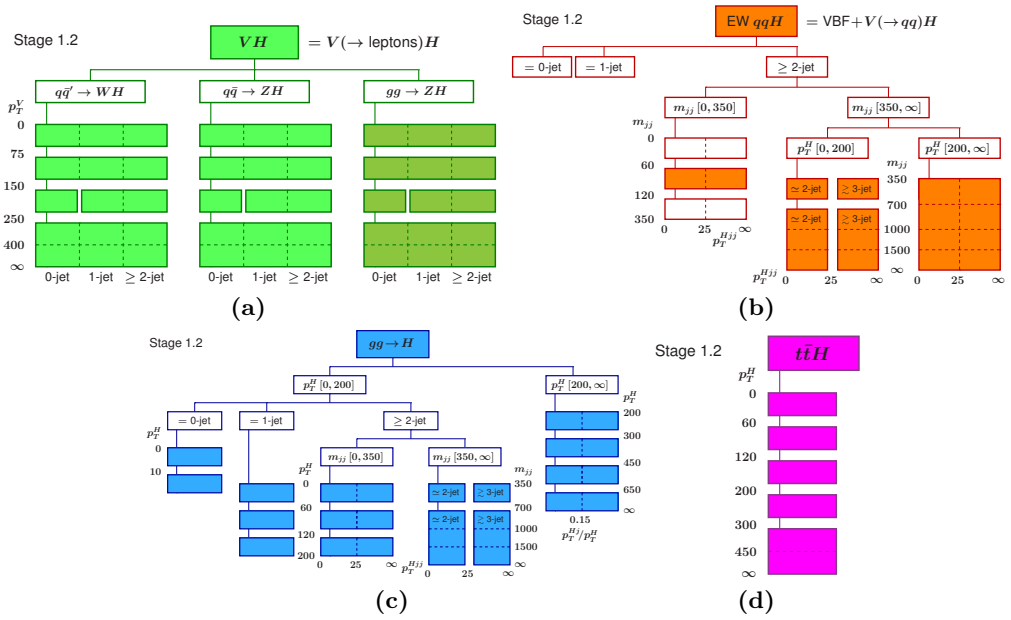


Figure 1.9: STXS stage 1.2 bin definitions for (a) $VH, V(\to \text{leptons})$, (b) vector boson fusion (VBF) and $VH, V(\to qq)$, (c) $gg \to H$ and (d) $t\bar{t}H$ processes (taken from Ref. [56]).

- minimizing the dependencies with respect to the theoretical uncertainties. In particular extrapolation of the measurements from a certain measured region in phase space to a much larger region of phase space are avoided and cases with a large variation in the experimental acceptance or sensitivity within a given bin are removed.
- Maximizing the experimental sensitivity,
- measuring cross-sections instead of signal strengths, measurement performed separately into production modes (or more generally into kinematic templates) while allowing the use of advanced analysis techniques such as event categorization, multivariate techniques,
- isolating possible BSM effects,
- finding an optimal trade-off between minimizing the number of bins without loss of experimental sensitivity.

An STXS measurement can moreover be easily used for theoretical interpretations as for instance in the context of an effective field theory interpretation which is detailed in Section 1.4.6.

In practice the STXS measurement is performed in the following way: the simulated signal samples are used to create one set of histograms for each STXS bin. Any simulated signal event passing the event selection is assigned to a specific STXS bin and used to fill a specific histogram depending on the reconstructed information but also the truth information *i.e.* the quantities obtained at event generation level. The STXS bins are defined to be as close as possible to the experimental kinematic selections. For instance for the $VH, H \to b\bar{b}$ STXS measurement the events are categorized as a function of the reconstructed p_{T}^{V}, the reconstructed number of jets in the event, the leptonic channel for which the event is passing the selection (0-,1- and 2-(charged)-lepton channels), and the truth p_{T}^{V} denoted $p_{\mathrm{T}}^{V,t}$ and true process mode which is either WH or ZH. Thanks to the response matrix (truth versus reconstructed categories) the results are unfolded which allows measuring cross-section as a function of the truth information as for instance the $p_{\mathrm{T}}^{V,t}$ for the $VH, H \to b\bar{b}$ analysis.

1.4.5 Coupling-strength modifiers

Experimental results can be interpreted with the κ-framework [53, 57] which parametrizes the $i \to H \to f$ decay as [58, 59]:

$$\sigma_{i\to H\to f} = \sigma_{i\to H} \cdot \mathcal{B}_{H\to f} = \sigma_i(\boldsymbol{\kappa}) \cdot \frac{\Gamma_{H\to f}(\boldsymbol{\kappa})}{\Gamma_H(\boldsymbol{\kappa}, \mathcal{B}_{\text{inv.}}, \mathcal{B}_{\text{u.}})} \tag{1.4.6}$$

$$\Leftrightarrow \sigma_{i\to H\to f} = \kappa_i^2 \cdot \kappa_f^2 \cdot \sigma_i^{\text{SM}} \cdot \frac{\Gamma_f^{\text{SM}}}{\Gamma_H(\boldsymbol{\kappa}, \mathcal{B}_{\text{inv.}}, \mathcal{B}_{\text{u.}})}, \tag{1.4.7}$$

with i the initial state, f a given final state resulting from a Higgs decay, $\mathcal{B}_{\text{inv.}}$ and $\mathcal{B}_{\text{u.}}$ are respectively the branching fractions of possible BSM invisible and undetected (due to large background) Higgs decays. In the second equation, the $i \to H$ and $H \to f$ notation are respectively replaced by i and f for conciseness. The κ_i and κ_f are called the coupling-strength modifiers defined as

$$\kappa_i^2 := \frac{\sigma_{i\to H}}{\sigma_{i\to H}^{\text{SM}}}, \qquad \kappa_f^2 := \frac{\Gamma_{H\to f}}{\Gamma_{H\to f}^{\text{SM}}}. \tag{1.4.8}$$

As a result, if the coupling-strength modifiers are found to be compatible with unity then the observed process is compatible with the SM predictions. The advantage of the coupling-strength modifier formalism is that it allows for a direct comparison of the Higgs couplings with different particles as for instance presented in Section 5.9.2.2 where the Higgs-to-charm coupling was experimentally measured to be weaker than the Higgs-to-bottom coupling by the ATLAS experiment.

The contributions to the total Higgs decay may also manifest finding values for $\mathcal{B}_{\text{inv.}}$ and $\mathcal{B}_{\text{u.}}$ differing from zero. The Higgs boson total width Γ_H is thus expressed as:

$$\Gamma_H(\boldsymbol{\kappa}, \mathcal{B}_{\text{inv.}}, \mathcal{B}_{\text{u.}}) = \kappa_H^2(\boldsymbol{\kappa}, \mathcal{B}_{\text{inv.}}, \mathcal{B}_{\text{u.}}) \cdot \Gamma_H^{\text{SM}} \tag{1.4.9}$$

$$\Rightarrow \kappa_H^2(\boldsymbol{\kappa}, \mathcal{B}_{\text{inv.}}, \mathcal{B}_{\text{u.}}) = \frac{\sum_j \mathcal{B}_{H\to j}^{\text{SM}} \cdot \kappa_j^2}{1 - (\mathcal{B}_{\text{inv.}} + \mathcal{B}_{\text{u.}})}. \tag{1.4.10}$$

Since the total Higgs width is predicted to be around $4\,\text{MeV}$ by the SM, and is hence much smaller than the experimental resolution of the Higgs boson mass measurements at the LHC, a direct measurement of the Higgs boson width is not feasible. As a result further assumptions are needed for the total Higgs width. Two typical (alternative) hypotheses are usually tested:

- ➢ either considering only SM particles hence fixing $\mathcal{B}_{\text{inv.}} = \mathcal{B}_{\text{u.}} = 0$, and loop-induced diagrams are resolved in terms of the particles running in the loops.
- ➢ Or $\mathcal{B}_{\text{inv.}}$ and $\mathcal{B}_{\text{u.}}$ are free parameters, and the coupling strength modifier of vector bosons (κ_V) is assumed to be $|\kappa_V| \leq 1$ since it is the case in a certain number of BSM scenarios such as two-Higgs-doublet model (2HDM).

1.4.6 Effective field theory

The effective field theory (EFT) [60–62] approach enables to probe BSM theories searching for hints of deviation from the SM due to new physics phenomena that would appear above a given energy scale (Λ). Usually, this energy scale is taken to be $\Lambda = 1\,\text{TeV}$ as such threshold is reachable at the LHC, but the constraints that are obtained can be extrapolated to different energy scales by means of a rescaling. The SM Lagrangian ($\mathcal{L}_{\text{SM}}$ from Equation (1.1.5)) is extended to a so-called SM effective field theory (SMEFT) Lagrangian ($\mathcal{L}_{\text{SMEFT}}$) defined as:

$$\mathcal{L}_{\text{SMEFT}} = \mathcal{L}_{\text{SM}} + \sum_{D\geq 5} \sum_{i\in\mathcal{W}_D} \frac{c_i^{(D)}}{\Lambda^{D-4}} \mathcal{O}_i^{(D)}, \tag{1.4.11}$$

where $D \geq 5$ is the dimension of the EFT operators ($\mathcal{O}_i^{(D)}$ expressed in the Warsaw basis [63]), $c_i^{(D)}$ are the Wilson coefficients associated to the operators and $\mathcal{W}_D$ is the ensemble of allowed type of operators for a given dimension. The Wilson parameters are free parameters of the EFT that can be constrained using experimental data. If all those coefficients are found to be vanishing then the observations are compatible with the SM. Since ultraviolet complete models beyond the SM can be mapped to an EFT Lagrangian, the constraints that are derived can then be interpreted as limits on the relevant parameters of such models. Typically, only dimension-6 operators are considered as odd operators are prohibited since they violate lepton and/or baryon number conservation while even operators of higher dimension $D \geq 8$ are suppressed by the energy scale and are thus assumed to be negligible.

For dimension-6 operators, there are in total 2499 additional operators which can be built using the SM fields but only a small subset will affect the Higgs boson physics. For instance only 17 operators modify the $qq \to V(\to \text{leptons})H(\to b\bar{b})$ process, of which only 4 operators affect the $H \to b\bar{b}$ decay. A list of Wilson coefficients associated to dimension-6 operators affecting some Higgs decays is provided in Table 1.1, in particular the Table 1.1b specifies some coefficients of interest to probe for the $VH, H \to b\bar{b}$ analysis.

The cross-section is proportional to the squared amplitude of the considered process ($\sigma \propto |A|^2$). In a SMEFT context, such amplitude can be decomposed in two terms $A_{\text{SMEFT}} = A_{\text{SM}} + A_{\text{BSM}}$ corresponding to SM and BSM phenomena. The cross-section is thus the sum of $|A_{\text{SM}}|^2$, the interference term which is linear in terms of $|A_{\text{BSM}}|$, and the BSM term which is quadratic ($|A_{\text{BSM}}|^2$). Hence the SMEFT cross-section can be split in three contributions:

$$\sigma^i_{\text{SMEFT}} = \sigma^i_{\text{SM}} + \sigma^i_{\text{int}} + \sigma^i_{\text{BSM}}, \tag{1.4.12}$$

where σ^i_{SM} is the cross-section computed with the SM predictions for the initial state i, σ^i_{int} accounts for potential interferences between the SM and the hypothetical BSM processes and σ^i_{BSM} is exclusively related to the hypothetical BSM processes. The possible cross-section deviations from the SM predictions can then be re-expressed thanks to the Wilson coefficients:

$$\frac{\sigma^i_{\text{SMEFT}}}{\sigma^i_{\text{SM}}} = 1 + \sum_j A_j^{\sigma_i} c_j + \sum_{j,k} B_{jk}^{\sigma_i} c_j c_k, \tag{1.4.13}$$

with $A_j^{\sigma_i}$ and $B_{j,k}^{\sigma_i}$ respectively referred to as the linear and the quadratic terms (associated to the initial state i) which are computed from the SMEFT operators and which define σ^i_{int} and σ^i_{BSM}:

$$\frac{\sigma^i_{\text{int}}}{\sigma^i_{\text{SM}}} := \sum_j A_j^{\sigma_i} c_j, \qquad \frac{\sigma^i_{\text{BSM}}}{\sigma^i_{\text{SM}}} := \sum_{j,k} B_{jk}^{\sigma_i} c_j c_k, \tag{1.4.14}$$

those linear and quadratic terms being respectively of the order of $1/\Lambda^2$ and $1/\Lambda^4$. In the same way, the partial Higgs width for a final state f should also be affected:

$$\frac{\Gamma^f_{\text{SMEFT}}}{\Gamma^f_{\text{SM}}} = 1 + \sum_j A_j^{\Gamma_f} c_j + \sum_{j,k} B_{j,k}^{\Gamma_f} c_j c_k = 1 + \frac{\Gamma^f_{\text{int}}}{\Gamma^f_{\text{SM}}} + \frac{\Gamma^f_{\text{BSM}}}{\Gamma^f_{\text{SM}}}, \tag{1.4.15}$$

with $A_i^{\Gamma_f}$ and $B_{ij}^{\Gamma_f}$ also obtained from the EFT predictions, and so is the total Higgs boson width which satisfies

$$\frac{\Gamma^H_{\text{SMEFT}}}{\Gamma^H_{\text{SM}}} = 1 + \sum_j A_j^{\Gamma_H} c_j + \sum_{j,k} B_{j,k}^{\Gamma_H} c_j c_k = 1 + \frac{\Gamma^H_{\text{int}}}{\Gamma^H_{\text{SM}}} + \frac{\Gamma^H_{\text{BSM}}}{\Gamma^H_{\text{SM}}}. \tag{1.4.16}$$

Combining the above parametrizations, the following relation between the SMEFT and SM quan-

Wilson coefficient	and operator	Affected process group		
		LEP/SLD EWPO	ATLAS Higgs	ATLAS electroweak
$c_{H\Box}$	$(H^\dagger H)\Box(H^\dagger H)$		✓	
c_G	$f^{abc}G^{a\nu}_\mu G^{b\rho}_\nu G^{c\mu}_\rho$		✓	✓
c_W	$\epsilon^{IJK}W^{I\nu}_\mu W^{J\rho}_\nu W^{K\mu}_\rho$		✓	✓
c_{HD}	$\left(H^\dagger D_\mu H\right)^*\left(H^\dagger D_\mu H\right)$		✓	✓
c_{HG}	$H^\dagger H\, G^A_{\mu\nu}G^{A\mu\nu}$		✓	
c_{HB}	$H^\dagger H\, B_{\mu\nu}B^{\mu\nu}$		✓	
c_{HW}	$H^\dagger H\, W^I_{\mu\nu}W^{I\mu\nu}$		✓	
c_{HWB}	$H^\dagger\tau^I H\, W^I_{\mu\nu}B^{\mu\nu}$	✓	✓	✓
c_{eH}	$(H^\dagger H)(\bar{l}_p e_r H)$		✓	
c_{uH}	$(H^\dagger H)(\bar{q}Y^\dagger_u u\tilde{H})$		✓	
c_{tH}	$(H^\dagger H)(\bar{Q}\tilde{H}t)$		✓	
c_{bH}	$(H^\dagger H)(\bar{Q}Hb)$		✓	
$c^{(1)}_{Hl}$	$(H^\dagger i\overleftrightarrow{D}_\mu H)(\bar{l}\gamma^\mu l)$	✓	✓	✓
$c^{(3)}_{Hl}$	$(H^\dagger i\overleftrightarrow{D}^I_\mu H)(\bar{l}\tau^I\gamma^\mu l)$	✓	✓	✓
c_{He}	$(H^\dagger i\overleftrightarrow{D}_\mu H)(\bar{e}\gamma^\mu e)$	✓	✓	✓
$c^{(1)}_{Hq}$	$(H^\dagger i\overleftrightarrow{D}_\mu H)(\bar{q}\gamma^\mu q)$	✓	✓	✓
$c^{(3)}_{Hq}$	$(H^\dagger i\overleftrightarrow{D}^I_\mu H)(\bar{q}\tau^I\gamma^\mu q)$	✓	✓	✓
c_{Hu}	$(H^\dagger i\overleftrightarrow{D}_\mu H)(\bar{u}\gamma^\mu u)$	✓	✓	✓
c_{Hd}	$(H^\dagger i\overleftrightarrow{D}_\mu H)(\bar{d}\gamma^\mu d)$	✓	✓	✓
$c^{(1)}_{HQ}$	$(H^\dagger i\overleftrightarrow{D}_\mu H)(\bar{Q}\gamma^\mu Q)$	✓	✓	
$c^{(3)}_{HQ}$	$(H^\dagger i\overleftrightarrow{D}^I_\mu H)(\bar{Q}\tau^I\gamma^\mu Q)$	✓	✓	
c_{Hb}	$(H^\dagger i\overleftrightarrow{D}_\mu H)(\bar{b}\gamma^\mu b)$	✓		
c_{Ht}	$(H^\dagger i\overleftrightarrow{D}_\mu H)(\bar{t}\gamma^\mu t)$	✓	✓	
c_{tG}	$(\bar{Q}\sigma^{\mu\nu}T^A t)\tilde{H}\, G^A_{\mu\nu}$		✓	
c_{tW}	$(\bar{Q}\sigma^{\mu\nu}t)\tau^I\tilde{H}\, W^I_{\mu\nu}$		✓	
c_{tB}	$(\bar{Q}\sigma^{\mu\nu}t)\tilde{H}\, B_{\mu\nu}$		✓	
c_{ll}	$(\bar{l}\gamma_\mu l)(\bar{l}\gamma^\mu l)$	✓		✓

(a)

Wilson coefficient	Operator	Impacted vertex	
		Production	Decay
c_{HWB}	$\mathcal{Q}_{HWB} = H^\dagger\tau^I HW^I_{\mu\nu}B^{\mu\nu}$	HZZ	
c_{HW}	$\mathcal{Q}_{HW} = H^\dagger HW^I_{\mu\nu}W^{\mu\nu}_I$	HZZ, HWW	
$c^{(3)}_{Hq}$	$\mathcal{Q}^{(3)}_{Hq} = (H^\dagger i\overleftrightarrow{D^I_\mu}H)(\bar{q}_p\tau^I\gamma^\mu q_r)$	$qqZH, qq'WH$	
$c^{(1)}_{Hq}$	$\mathcal{Q}^{(1)}_{Hq} = (H^\dagger i\overleftrightarrow{D_\mu}H)(\bar{q}_p\gamma^\mu q_r)$	$qqZH$	
c_{Hu}	$\mathcal{Q}_{Hu} = (H^\dagger i\overleftrightarrow{D_\mu}H)(\bar{u}_p\gamma^\mu u_r)$	$qqZH$	
c_{Hd}	$\mathcal{Q}_{Hd} = (H^\dagger i\overleftrightarrow{D_\mu}H)(\bar{d}_p\gamma^\mu d_r)$	$qqZH$	
c_{dH}	$\mathcal{Q}_{dH} = (H^\dagger H)(\bar{q}dH)$		Hbb

(b)

Table 1.1: List of some Wilson coefficients and their corresponding dimension-6 operators in the Warsaw formulation affecting (a) Higgs decays and (b) $qq \to V(\to \text{leptons})H(\to b\bar{b})$ processes (taken from Refs. [64, 65]).

tities can be derived

$$(\sigma \times \mathcal{B})^{i\to H\to f}_{\text{SMEFT}} = (\sigma \times \mathcal{B})^{i\to H\to f}_{\text{SM}} \times \left(\frac{\sigma^{i}_{\text{SMEFT}}}{\sigma^{i}_{\text{SM}}}\right) \times \left(\frac{\frac{\Gamma^{f}_{\text{SMEFT}}}{\Gamma^{f}_{\text{SM}}}}{\frac{\Gamma^{H}_{\text{SMEFT}}}{\Gamma^{H}_{\text{SM}}}}\right). \tag{1.4.17}$$

The maximum likelihood fits are then performed across the different cross-section measurements (*e.g.* over all STXS bins in case of an STXS measurement) to determine and constrain the Wilson coefficients. Since there are not enough measurements to constrain all the Wilson coefficients simultaneously, typically limits on each individual Wilson coefficient are set by assuming all the other ones to be vanishing leading to a one-dimensional confidence level (CL) intervals inferred for the coefficient under study both with and without the quadratic term configuration. Those two configurations are referred to as linear and linear-plus-quadratic parametrizations and are usually both tested as for the linear case the BSM terms are neglected while the linear-plus-quadratic parametrization provides an insight on the impact of the BSM terms and how it affects the derived constraints on the Wilson coefficients. In principle the interferences with dimension-8 operators should be included as they could be of same size as the BSM dimension-6 terms so the comparison of results, to some extent, allows assessing if limits obtained can be trusted or if it would require including as well the dimension-8 operators. Sometimes, simultaneous fits with two parameters can be performed with the linear or linear-plus-quadratic parametrizations which allows extracting two-dimensional confidence levels on a pair of Wilson coefficients.

1.5 Event simulation

Thanks to the SM Lagrangian from Equation (1.1.5), interactions between particles can be simulated to predict the outcome of real data collisions occurring at the LHC. Such comparison allows assessing if deviations from SM prediction are observed which would indicate either new physics phenomena or an experimental or theoretical bias (*e.g.* non-valid approximation in matrix element computation, a need for higher order corrections or so). An example of an event simulation is presented in Figure 1.10. Simulations of events is a rather complex task which is divided in several steps as highlighted in that figure and which are detailed below [66]. Simulations are relying on several available Monte Carlo (MC) programs which are also detailed at the end of this section.

- **Parton distribution functions (PDFs)**: as already discussed in Section 1.4.1, a PDF corresponds to the probability of parton to carry a fraction x of the total momentum. PDFs are obtained from fits of experimental results over a large range of energies. However, each set of PDFs depends on the choices of input datasets used to derive them, the order in perturbation theory or also the input parameterization. Therefore, this can result in discrepancies between PDF sets thus PDFs set uncertainties are important to account for such possible bias. There exist several PDFs sets such as PDF4LHC, NNPDF3.0, CT14 and so on which are regularly updated to include new measurements and thus increase precision of their prediction.
- **Hard scattering and underlying event (UE)**: the hard scattering designates collisions between partons where a high momentum (Q^2) is transferred. Those type of collisions are thus of interest since new potentially massive particles such as the Higgs are created in those processes as opposed to low energy inelastic scattering or elastic scattering where such energy transfer is much lower and are thus pile-up collisions. The underlying event contains all events not originating from the primary hard scattering. The UE involves contributions from initial and final state (ISR and FSR) radiations (*i.e.* particles that are emitted by initial state or final state particles), beam-beam remnants or multiple parton interactions. The underlying event is split in two categories with an energy cutoff: interactions whose energy is above this threshold will be

included in the parton showering while others are included in the UE. The choice of the cutoff energy depends on the PDF set, the collision energy and the impact parameter of the protons. The UE activity is simulated by tunable parametric tools based on minimum bias data collected with very loose triggers.

- **Matrix element (ME)**: concerns the hard scattering part of the collision process. Matrix element amplitudes are computed from the Feynman rules taking into account also higher order corrections to predict the final state particles. For instance at ME stage is simulated $qq \to VH, H \to b\bar{b}$ or $gg \to ZH, H \to b\bar{b}$ processes and decays of short-lived particles as $W \to l\nu$ or $H \to b\bar{b}$ are included in the computation. The matrix elements are computed for a fixed renormalization scale denoted μ: changing the choice of that energy scale can change the outcome of the simulation.
- **Parton shower (PS)**: radiation of particles by the final state particles must be simulated, final state particles that are predicted by the ME computation. Indeed, the final state partons produced by the ME calculation are not spontaneously forming hadrons because they are still in the asymptotic QCD freedom regime. They therefore lose energy through radiations until their energy is low enough to combine to form hadrons (hadronization step). If the ME and PS algorithms used are different, the combination of the two generators results has to be performed carefully to prevent overlapping Feynman diagrams as for instance at NLO, the ME generators already simulate radiations of one extra parton which should hence be prohibited at PS level to avoid double counting such emission as it would bias the simulation of the process. The parton showering is simulated thanks to the Sudakov factor [67] which basically represents the probability for a parton to radiate particles over a certain time interval. To control the radiations at each stage, the DGLAP equations are required to be satisfied (standing for Dokshitzer-Gribov-Lipatov-Altarelli-Parisi). Such equations are describing the variation of parton distribution functions with varying energy scales and radiations.
- **Hadronization**: after parton showering, partons cannot be treated anymore with perturbative QCD as their energy has decreased and is of the order of $1\ \mathrm{GeV}$. At this stage, partons are bounding together to form hadrons (due to the confinement principle). Such step is usually simulated by the parton showering algorithm and is complicated as it corresponds to the transition between two QCD regimes, transition which is not well formalized. Phenomenological models based on parameters tunes are thus used for this step. There are basically two main ways to simulate the formation of hadrons either the Lund string model [68] (used by PYTHIA) or the cluster model [69] (used by HERWIG and SHERPA, also referred to as cluster fragmentation) which are presented in Figure 1.11. In the Lund string model, quarks are treated as strings that splits to form new hadrons as shown in Figures 1.3 and 1.11a while the cluster model is based on the colour preconfinement where close-by quarks are grouped such that they form colour neutral groups (referred to as colour connection) as illustrated in Figure 1.11b. Since most of the hadrons produced are unstable particles, the hadronization algorithm simulate their further decay to lighter hadrons based on measurements of particle decays branching ratios.
- **Pile-up**: at the LHC several hard scattering proton-proton collisions occur per bunch crossing such phenomena is called pile-up and needs to be simulated to reproduce data taking conditions. Hard primary scattering simulated processes are thus overlaid with simulation of the pile-up conditions expected at the LHC.

To fully simulate an event from matrix elements computation to parton showering and underlying events, one or several event generators can be used. Indeed, some event generators are only specialized for some steps. Below is the list of event generators commonly used by ATLAS.

- **PYTHIA** [73, 74]: is a multipurpose generator that can be used for ME, PS and UE simulation stages. The ME calculations are performed at LO and the PS and UE models are tuned thanks to reference measurements. The hadronization step relies on the Lund model. Several (PDFs) set of tunes are provided, the most common one being the AZNLO and A14 tunes. This generator

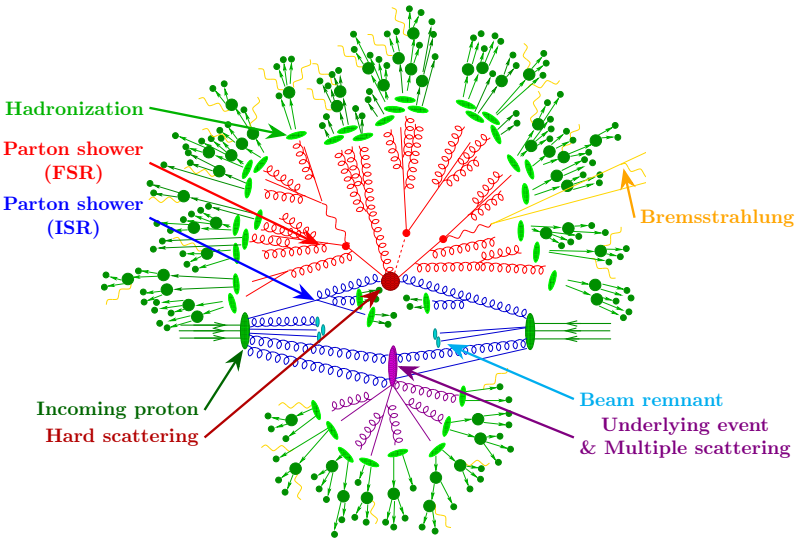


Figure 1.10: Sketch of a proton-proton collision simulated with the Sherpa event-generator (modified figure taken from Ref. [70]).

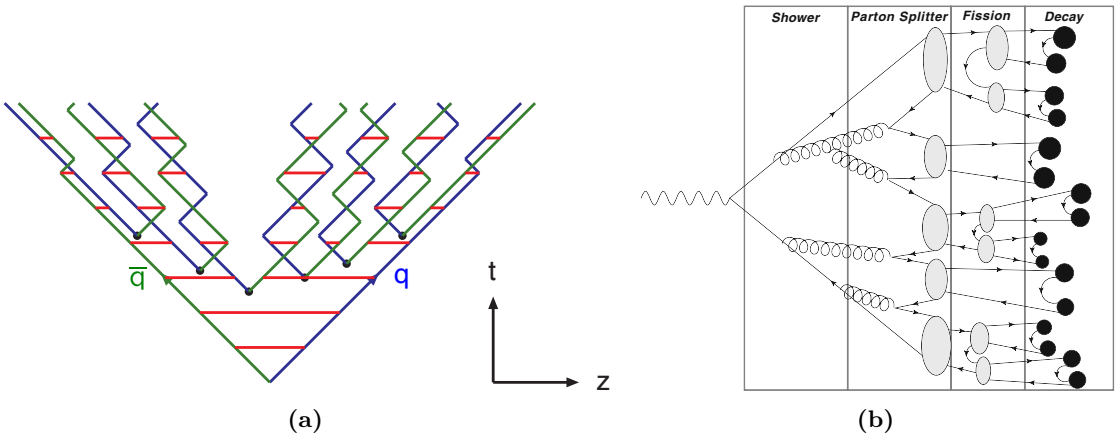


Figure 1.11: Schematization of the hadronization for the (a) Lund string model versus (b) the cluster model (taken from Refs. [71, 72]).

can also be used to only simulate some steps of the event generation which is useful as it allows for combination with other generators. For instance since the ME computation are only computed a LO precision for PYTHIA, this generator is often replaced by POWHEG or MADGRAPH for that step.

- **SHERPA** [75, 76]: is a multipurpose generator that performs at once ME and PS simulations. This generator is the preferred one for processes with additional radiated jets that are directly included in the ME computations. The hadronization model of SHERPA is based on the cluster fragmentation model. Unlike PYTHIA, SHERPA cannot be used to partially simulate an event.
- **HERWIG** [77, 78]: is also a multipurpose generator. The hadronization model is based on the cluster fragmentation model. When combined with POWHEG, this generator is useful to assess uncertainties associated to fragmentation and hadronization modellings.
- **POWHEG** [79, 80]: is a ME NLO event generator able to simulate a large variety of processes. POWHEG needs to be interfaced to other generators for the PS steps and require a ME-PS matching to avoid double counting issues. It is usually coupled with PYTHIA or HERWIG.
- **MADGRAPH** [81]: is similar to POWHEG, it is only meant for ME computation which are provided at NLO accuracy. MADGRAPH can be used to study modelling uncertainties when interfaced with PS generators PYTHIA or HERWIG.

After events have been fully simulated, they are then passed through a detailed simulation of the ATLAS detector [82] which is based on GEANT4 [83] to simulate the response of the detector and be compared to experimental results.

1.6 Higgs boson measurements at the LHC

Almost 50 years after its theoretical prediction [41–44, 84, 85], the Higgs boson was both discovered by the ATLAS and CMS collaborations [1, 2] in 2012, using about $11\ \text{fb}^{-1}$ of data from proton-proton collisions at a center-of-mass-energy $\sqrt{s} = 7$ and $8\ \text{TeV}$ recorded at the LHC, analyzing events presenting the characteristics of the so-called *golden channels*: $H \to \gamma\gamma$ and $H \to ZZ^* \to 4l$ ($l = e, \mu$) decays. Such analyses, despite their very low cross-sections, are highly pure in signal events thanks to the strong and clean signature of those Higgs decays as presented in Figure 1.12. Moreover, the ATLAS and CMS detectors have a precise energy resolution for photons and leptons which allows for high precision mass measurements. As a result, the most precise measurement to date, $m_H = 125.38 \pm 0.14\ \text{GeV}$ (uncertainty of 0.1% on the mass), of the Higgs mass has been achieved by CMS [86] combining the two golden channel results from Run 1 and 2016 (partial Run 2) data taking periods[2]. On the other hand, the best precision from ATLAS was obtained thanks to $H \to ZZ^* \to 4l$ events [87] with a mass measured to be $m_H = 124.94 \pm 0.17\,\text{GeV}$ with the full Run 2 data-taking. Combining the ATLAS and CMS full Run 1 results for the golden channels, the Higgs mass was measured to be equal to $m_H = 125.09 \pm 0.24\ \text{GeV}$ [88]. A combination of the ATLAS and CMS full Run 2 results is not yet available, but it should further improve the precision on the Higgs mass measurement thanks to the increase of statistics as shown in Figure 1.12.

This year (2022), for the 10th birthday of its discovery, ATLAS and CMS both published [58, 89] a summary of the main achievement concerning the Higgs boson measurements combining results from diverse Higgs production and decay mode so far measurable at the LHC. The ATLAS results are presented in this section. For the Run 2, since the collision center-of-mass energy is $\sqrt{s} = 13\ \text{TeV}$ and assuming a Higgs mass of $m_H = 125\ \text{GeV}$, then the main Higgs production modes and their expected rates [90] are the gluon-gluon fusion (ggF, 87%), the vector boson fu-

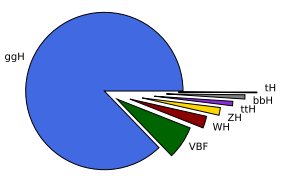
2. For the Run 1 (respectively the Run 2) of the Large Hadron Collider, proton-proton collisions occured at a center-of-mass-energy of $\sqrt{s} = 7$ and $8\ \text{TeV}$ (respectively $\sqrt{s} = 13\ \text{TeV}$).

sion (VBF, 7%), the Higgs-strahlung (VH with $V = W$ or Z boson, 4%), the $t\bar{t}H$ (1%) and the $b\bar{b}H$ (1%) mechanisms whose cross-sections are reported in Table 1.2 with as well the different Higgs decay branching ratios (which only depend on the Higgs mass). Indeed, the Higgs boson is an unstable particle with a lifetime of 1.6×10^{-22} s implying that it can only be observed indirectly by detecting its decay products. The associated leading order Feynman diagrams for Higgs production and Higgs decay modes are respectively provided in Figure 1.13 and 1.14. As shown in particular in Figure 1.15c, the branching fraction of the Higgs decays is strongly dependent on the mass of the Higgs boson hence the importance of precisely measuring its mass since m_H is a free parameter and is not predicted nor constrained by the SM. Since the coupling of the Higgs to fermions is proportional to their mass, the Higgs has a higher probability to decay into heavy fermions compared to light-fermions. The top mass is too large for a direct Higgs decay in a $t\bar{t}$ pair explaining why $H \to b\bar{b}$ and $H \to \tau^+\tau^-$ are its main fermionic decays.

Since its discovery, many Higgs production modes (ggF, VBF, VH, $t\bar{t}H$) and Higgs boson decays ($H \to \gamma\gamma, ZZ, W^+W^-, b\bar{b}, \tau^+\tau^-$), predicted by the Standard Model have been observed [3, 4, 88, 91–101] by ATLAS and CMS. As a consequence, many analyses have now entered into a precision measurement phase, performing differential and STXS measurements to probe extensively the SM predictions exploiting the increase of the amount of data collected over time, increase which is also crucial to be able to observe Higgs processes that are either rare and/or predominantly contaminated by background processes. ATLAS and CMS are in particular currently focusing on studying the Higgs boson coupling to the second generation of fermions: $H \to \mu^+\mu^-$ and $H \to c\bar{c}$ which are challenging measurements due to the very low branching ratio (0.02%) for $H \to \mu^+\mu^-$ decays, the large irreducible $Z \to \mu^+\mu^-$ background, and the low branching ratio (3%) for $H \to c\bar{c}$ decays which can only be studied for $VH, V \to$ leptons processes, which has a low cross-section, since other production modes suffer from a large contamination of QCD-multijet background.

Thanks to the observation of the $H \to \gamma\gamma$ decay, the Higgs boson cannot be a spin-1 particle due to the Landau-Yang theorem [102, 103]. However, it could either be a spin-0 (SM prediction) or a spin-2 particle (non-SM scenario). From studies of the angular distributions in $H \to ZZ^* \to 4l$, $H \to WW^* \to e\nu\mu\nu$ and $H \to \gamma\gamma$ decays [104] the non-SM spin hypotheses are excluded at more than 99.9% confidence level in favour of the SM spin-0 hypothesis. The CP-properties of the Higgs are however still currently under study in particular for $H \to \tau^+\tau^-$ and $H \to \gamma\gamma$ decays, and $t\bar{t}H$ production mode, as the Higgs could perhaps be a CP mixture of even and odd eigenstates [105].

Production mode	Cross section (pb)	Decay channel	Branching fraction (%)
ggH	48.31 ± 2.44	bb	57.63 ± 0.70
VBF	3.771 ± 0.807	WW	22.00 ± 0.33
WH	1.359 ± 0.028	gg	8.15 ± 0.42
ZH	0.877 ± 0.036	$\tau\tau$	6.21 ± 0.09
ttH	0.503 ± 0.035	cc	2.86 ± 0.09
bbH	0.482 ± 0.097	ZZ	2.71 ± 0.04
tH	0.092 ± 0.008	$\gamma\gamma$	0.227 ± 0.005
		Zγ	0.157 ± 0.009
		ss	0.025 ± 0.001
		$\mu\mu$	0.0216 ± 0.0004



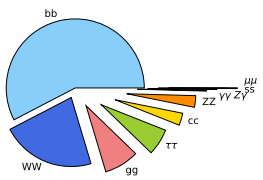


Table 1.2: Standard model Higgs theoretical production cross-sections and branching fractions for $\sqrt{s} = 13$ TeV and for the Higgs mass $m_H = 125.38$ GeV which is the value measured by CMS (taken from Ref. [89]).

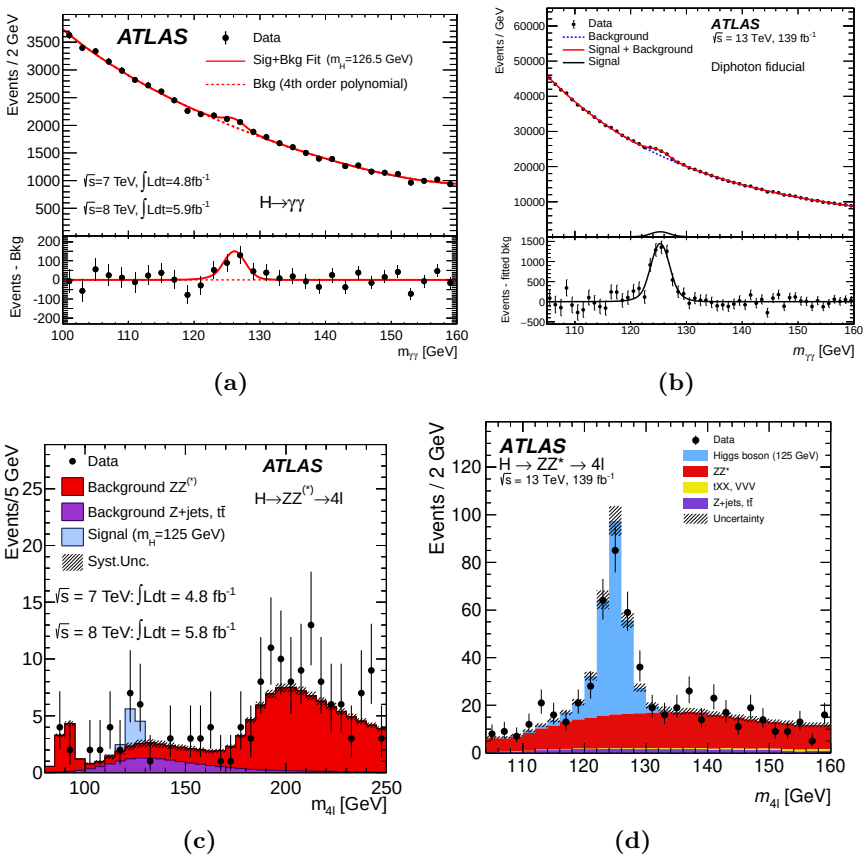


Figure 1.12: Invariant mass of the Higgs boson candidates in the (a) and (b) $H \to \gamma\gamma$, and (c) and (d) $H \to ZZ^* \to 4l$ channels as measured by ATLAS (a) and (c) for the observation of the Higgs boson [1] (partial Run 1 dataset), and (b) and (d) for the full Run 2. The corresponding integrated luminosities used for those measurements are reported in the figures. The m_{4l} distributions are pre-fit distributions while $m_{\gamma\gamma}$ are post-fit parametrized distributions (taken from Refs. [1, 87, 106]).

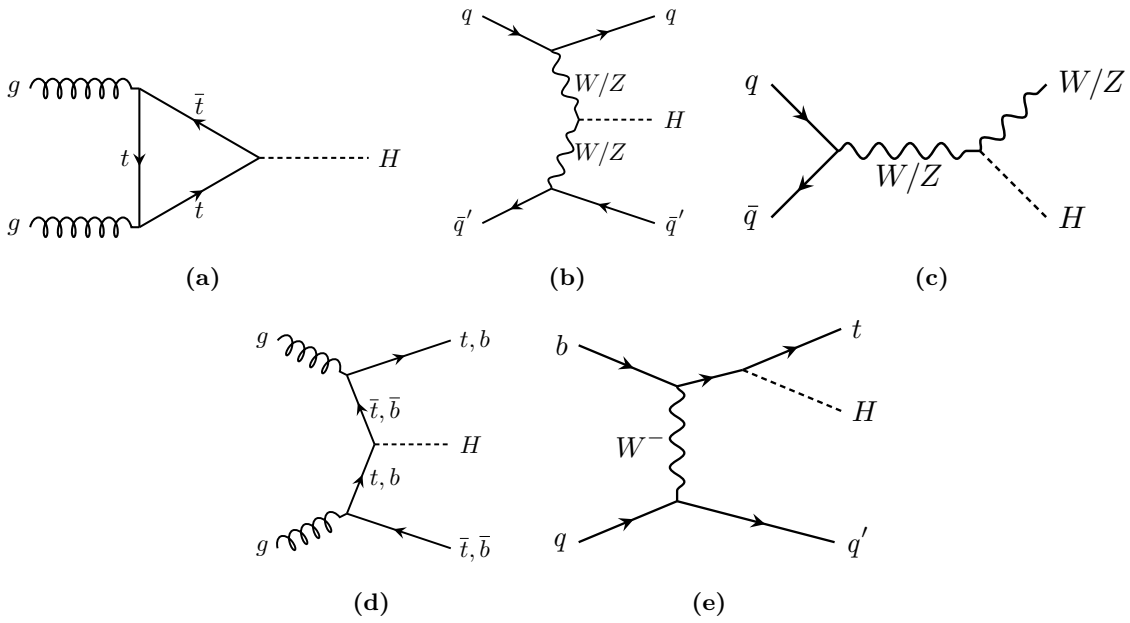
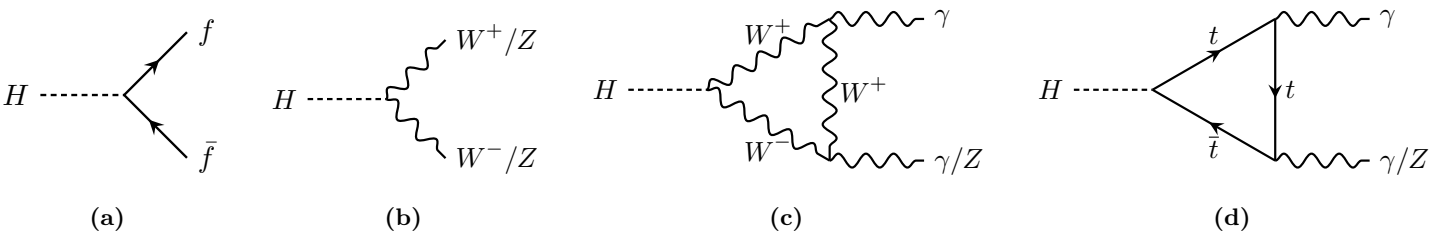
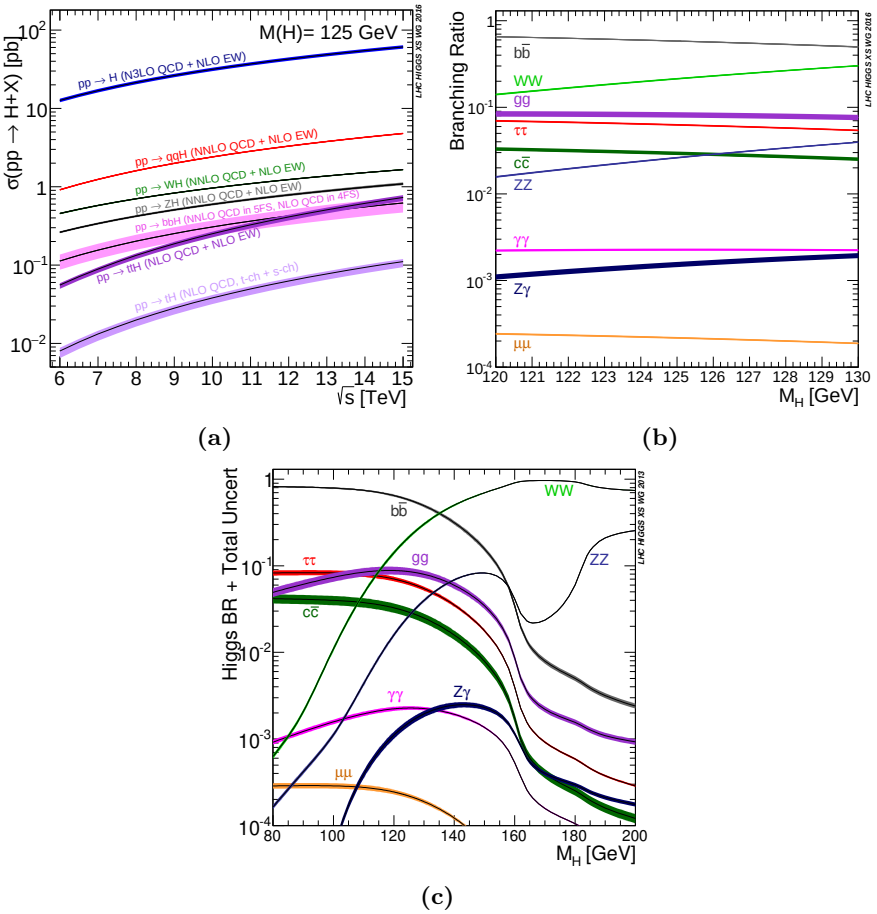


Figure 1.13: Leading order Feynman diagrams of the main Higgs production mechanisms: (a) gluon-gluon fusion (ggF), (b) vector boson fusion (VBF), (c) VH, $V = W, Z$ (Higgs-strahlung or VH mechanism), (d) $t\bar{t}H$ and $b\bar{b}H$, and (e) tH processes.



(a) **(b)** **(c)** **(d)**

Figure 1.14: Feynman diagrams of some Higgs decay channels: Higgs decays to (a) fermions, (b) vector bosons ($V = W, Z$), and (c) and (d) two photons ($\gamma\gamma$) or a Z boson and a photon ($Z\gamma$).



(a) **(b)**

(c)

Figure 1.15: (a) Higgs boson production cross-sections as a function of the centre-of-mass-energy ($\sqrt{s}$) for a Higgs mass $m_H = 125$ GeV. The tH production cross-section accounts for t-channel and s-channel only (no tWH production). (b) and (c) Higgs boson branching ratios and their uncertainties for different mass ranges around 125 GeV (taken from Refs. [53, 57]).

The Figure 1.16 presents measurement of cross-sections, branching fractions and signal strengths for different Higgs production and decay modes. An overall p-value of 72% is found for the signal strength measurements [58] proving an overall good compatibility with the SM predictions. Under the assumption that all the production decay processes are depending on the same global signal strength $\mu = \mu_{if}$, the inclusive Higgs boson signal strength found is

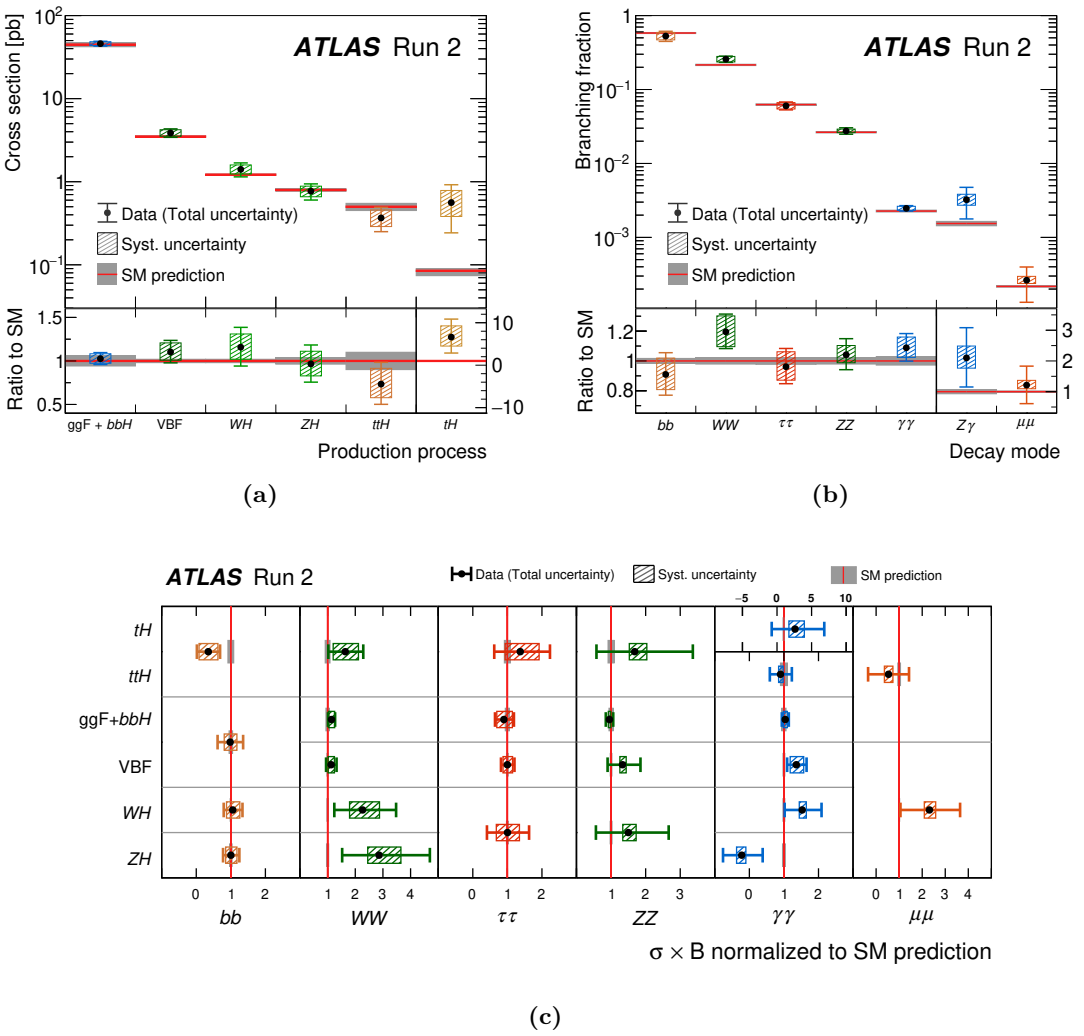
$$\mu = 1.05 \pm 0.06 = 1.05 \pm 0.03\,(\text{stat.}) \pm 0.03\,(\text{exp.}) \pm 0.04\,(\text{sig. th.}) \pm 0.02\,(\text{bkg. th.}), \quad (1.6.1)$$

which is in agreement with the SM predictions, the uncertainties mentioned corresponding respectively to the statistical, the experimental, the systematic, and the signal and background theoretical uncertainties. The results obtained can be interpreted with the kappa-framework leading to coupling-strength modifier measurements shown in Figures 1.17, 1.18 and 1.19 corresponding to three classes of models with more or less stringent hypotheses: either assuming a coupling strength is shared for some particles or having one coupling strength per particle. The reduced coupling strength modifiers for fermions ($\kappa_F g_F$) of mass m_F and coupling constant g_F, and weak vector bosons ($\kappa_V g_V$) of mass m_V and coupling constant g_V can be expressed as:

$$\kappa_F g_F = \kappa_F \cdot \frac{m_F}{v}, \qquad \sqrt{\frac{\kappa_V g_V}{2v}} = \sqrt{\kappa_V} \cdot \frac{m_V}{v}, \quad (1.6.2)$$

with v the vacuum expectation value of the Higgs field defined in Equation (1.2.4). For each model the results obtained are in good agreement with the SM expected values ($\kappa_F = \kappa_V = 1$). Moreover, such coupling constant measurement spans over 3 orders of magnitude in terms of the mass of particles considered between the lightest and heaviest particles as shown in Figure 1.18.

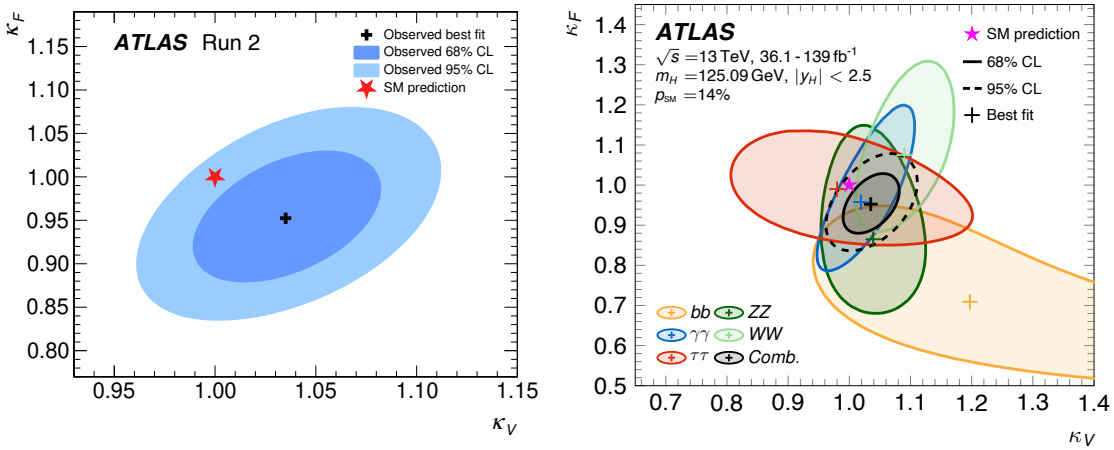
Finally, STXS measurements over several production and decay modes can be reinterpreted with an effective field theory, as performed in Reference [64], using dimension-6 EFT operators and the Wilson coefficients presented in Table 1.1. The results are reported in Figure 1.20: no deviations from the SM are observed since all Wilson coefficients are found to be compatible with zero.



(a) **(b)**

(c)

Figure 1.16: Observed and predicted Higgs boson production cross-sections, branching fractions and signal strengths. (a) The cross-sections for different Higgs boson production processes are measured assuming SM values for the decay branching fractions. (b) The branching fractions for different Higgs boson decay modes are measured assuming SM values for the production cross-sections. The lower panels show the ratios of the measured values to their SM predictions. The vertical bar on each point denotes the 68% confidence interval. (c) The signal strength for different combinations of Higgs boson production and decay processes. The horizontal bar on each point denotes the 68% confidence interval. On each figure, the narrow grey bands indicate the theory uncertainties (taken from Ref. [58]).



(a) (b)

Figure 1.17: (a) Negative log-likelihood contours corresponding to the 68% and 95% confidence level in the (κ_V, κ_F) plane. They are obtained from a combined fit assuming no contributions from invisible or undetected non-SM Higgs boson decays. A p-value of 14% is found for the compatibility of the combined measurement and the SM prediction. (b) Same figure showing the individual decay modes (colour) 68% confidence level contour in the (κ_V^k, κ_F^k) plane and their combination (black). The 95% CL contour of the combined measurement is also shown (black dashes) (taken from Ref. [58]).

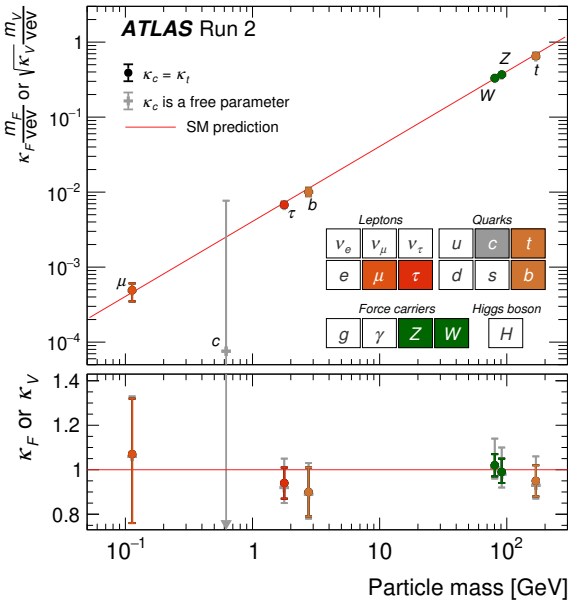


Figure 1.18: Reduced Higgs boson coupling strength modifiers and their uncertainties. Two fit scenarios with $\kappa_c = \kappa_t$ (coloured circle markers), or κ_c left free-floating in the fit (grey cross markers) are shown. Higgs boson decays to non-SM particles are not allowed. The vertical bar on each point denotes the 68% confidence interval. The p-values for compatibility of the combined measurement and the SM prediction are 56% and 65% for the respective scenarios. The lower panel shows the values of the coupling strength modifiers (taken from Ref. [58]).

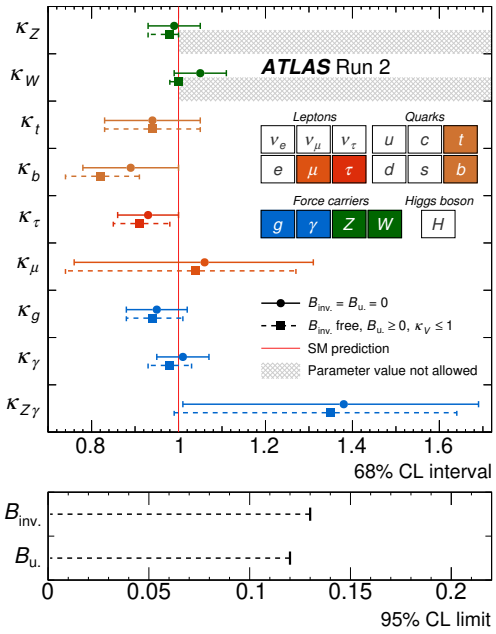


Figure 1.19: Reduced coupling strength modifiers and their uncertainties per particle type with effective photon, $Z\gamma$ and gluon couplings. The horizontal bars on each point denote the 68% confidence interval. Two different scenarios are tested: either $\mathcal{B}_{\text{inv.}} = \mathcal{B}_{\text{u.}} = 0$ (solid lines with circle markers) or $\mathcal{B}_{\text{inv.}}$ and $\mathcal{B}_{\text{u.}}$ are allowed to contribute to the total Higgs boson decay width while assuming that $\kappa_V \leq 1$ and $\mathcal{B}_{\text{u.}} \geq 0$ (dashed lines with square markers). The lower panel shows the 95% confidence upper limits on $\mathcal{B}_{\text{inv.}}$ and $\mathcal{B}_{\text{u.}}$ (taken from Ref. [58]).

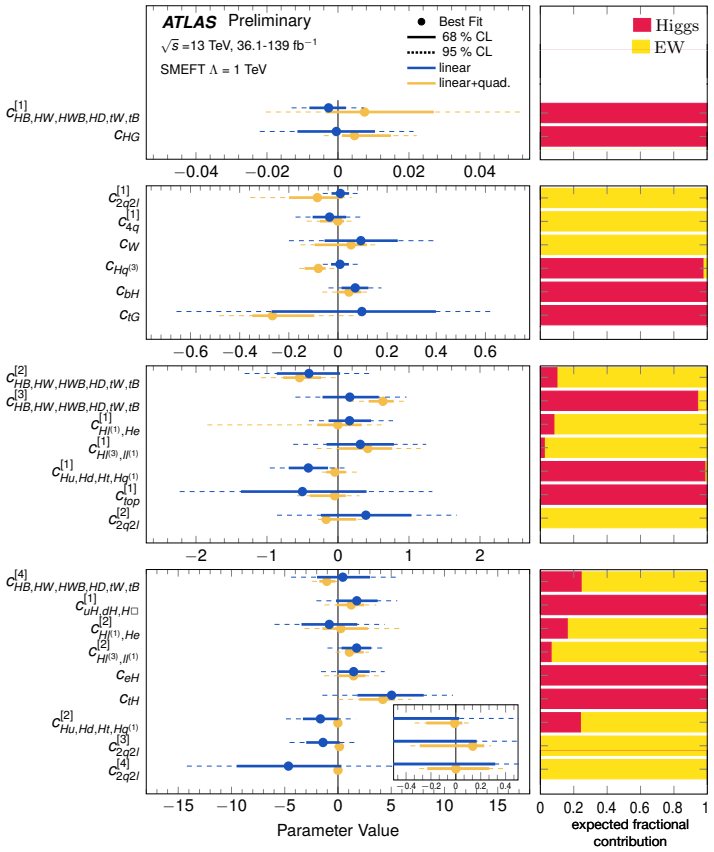


Figure 1.20: Constraints on Wilson coefficients presented in four blocks with different x-axis ranges. The right-hand side panel shows the contribution of each input measurement group to the eigenvector constraint in the Gaussian approximation of the linear model (taken from Ref. [64]).

Another important milestone, aimed by ATLAS and CMS, is the measurement of the Higgs self coupling (λ, as sketched in Figure 1.5a) which for the moment can only be constrained via upper confidence level signal strength limits due to the low cross-section of the processes that provide sensitivity to it. Such Higgs self interaction is in particular accessible studying Higgs boson pair production also referred to as double-Higgs production, production modes that are shown in Figure 1.21. Measuring λ is crucial since it is the other parameter of the Higgs potential (with μ) and therefore it affects its shape which can be related, for instance, to the possible role played by the Higgs field for the baryogenesis[3] [107]. The coupling strength modifier (κ_λ) for that trilinear coupling is defined as $\kappa_\lambda = \lambda_{HHH}/\lambda^{\text{SM}}_{HHH}$.

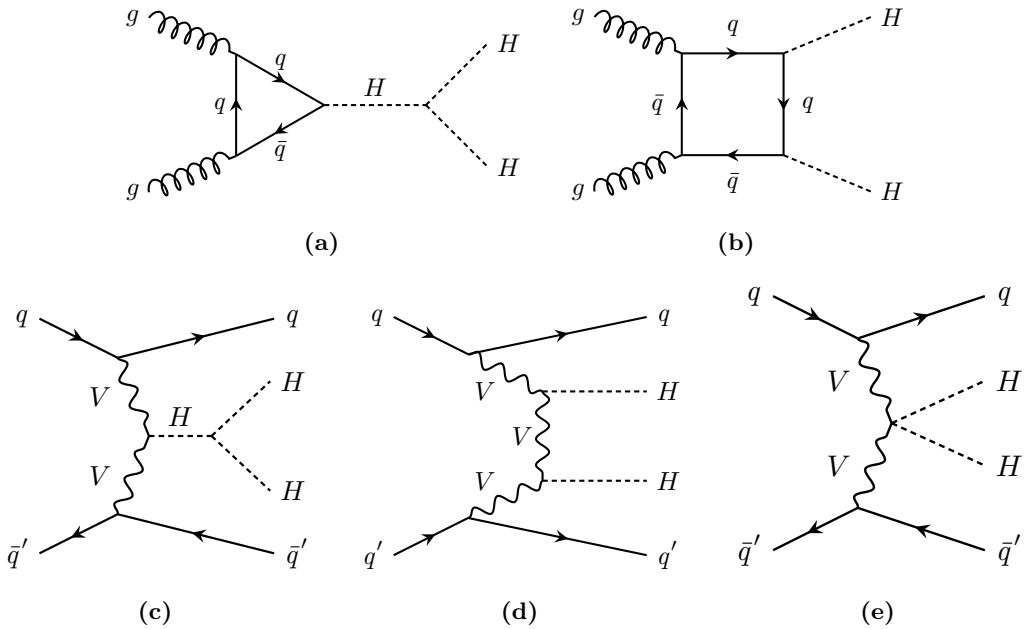


Figure 1.21: Feynman diagrams of the di-Higgs (a) and (b) gluon-gluon fusion, and (c), (d) and (e) vector boson fusion production modes.

The three most sensitive decay channels for di-Higgs production modes are $HH \to b\bar{b}\gamma\gamma, b\bar{b}\tau^+\tau^-$ and $b\bar{b}b\bar{b}$ decays which have been combined by ATLAS using ggF HH and VBF HH production modes [108]. The ATLAS Run 2 dataset allows to constrain the signal strength μ_{HH} with an observed (expected) upper limit of 2.4 (2.9) at a 95% confidence level as shown in Figure 1.22. The best combined fit value obtained $\mu_{HH} = -0.7 \pm 1.3$ is compatible with the SM and leads to a p-value of 0.2. That measurement also allows constraining κ_λ with an observed (expected) 95% CL limits: $-0.6 < \kappa_\lambda < 6.6$ ($-2.1 < \kappa_\lambda < 7.8$). Moreover, the VBF HH production mode offers sensitivity to the HHVV quartic interaction shown in Figure 1.5d. The associated coupling modifier κ_{2V} is observed (expected) to be within $0.1 < \kappa_{2V} < 2.0$ ($0.0 < \kappa_{2V} < 2.1$) with a 95% CL fixing all modifiers to their SM predictions.

Combining single Higgs STXS measurements ($H \to \gamma\gamma, ZZ^*, WW^*, \tau^+\tau^-$ and $b\bar{b}$) with di-Higgs ($HH \to b\bar{b}\gamma\gamma, b\bar{b}\tau^+\tau^-$ and $b\bar{b}b\bar{b}$) measurements, more stringent constraints on κ_λ can be reached as presented in Figure 1.23 with κ_t the coupling modifier of the top quark. κ_λ is observed (expected) to be in the interval $-0.4 < \kappa_\lambda < 6.3$ ($-1.9 < \kappa_\lambda < 7.6$).

Finally, invisible decays of the Higgs boson are also an important domain of research as they could solve the mystery about the nature of dark matter for instance. Dark matter candidates could be weakly interacting massive particles (WIMPs) and in that case should strongly couple with the Higgs boson leading to potential deviations of the SM predictions that could be observed in direct

3. In the scenario of the Big Bang, the baryogenesis is the period of time in the early Universe where baryons were created. This hypothetical physical process is assumed to be responsible for the baryonic asymmetry observed currently in our Universe.

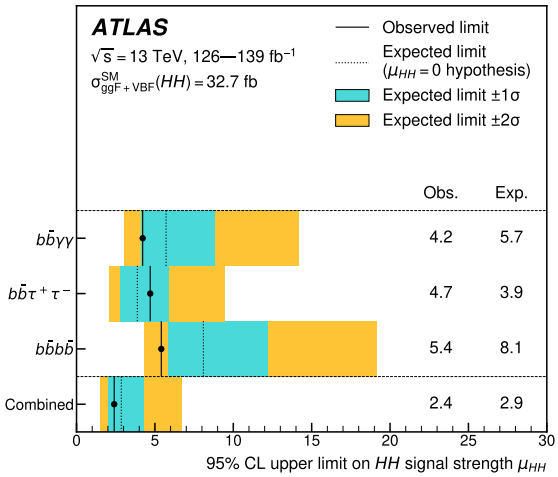


Figure 1.22: Observed 95% CL upper limits on the signal strength for double-Higgs production from the $HH \to b\bar{b}\gamma\gamma, b\bar{b}\tau^{+}\tau^{-}$ and $b\bar{b}b\bar{b}$ decay channels, and their statistical combination. The Higgs mass $m_H = 125.09\,\text{GeV}$ is assumed when deriving the predicted SM cross-section. The expected limit and the corresponding error bands are derived assuming the absence of the HH process (taken from Ref. [108]).

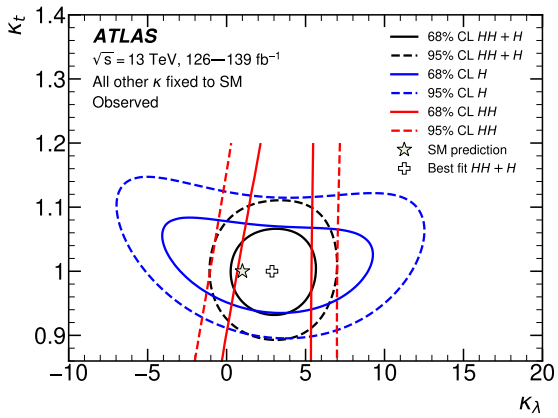


Figure 1.23: Observed constraints in the κ_λ-κ_t plane from single-Higgs (blue) and double-Higgs (red) analyses, and their combination (black). The solid (dashed) lines show the 68% (95%) CL contours. The double-Higgs contours are shown for values of κ_t smaller than 1.2 (taken from Ref. [108]).

or indirect Higgs measurements. The SM expected branching fraction for invisible Higgs decays $\mathcal{B}_{\text{inv.}}$ is 0.12% [109] due to $H \to ZZ^* \to \nu\bar{\nu}\nu\bar{\nu}$ decays. Any significant deviation from this value would imply new physics phenomena. The VBF production mode is expected to provide the best sensitivity for $H \to \text{inv.}$ decay studies. In its latest results, ATLAS performed such VBF study [109] and set an upper limit of 0.145 on the branching fraction for $\mathcal{B}_{\text{inv.}}$ at 95% CL with an expected limit of $0.103^{+0.041}_{-0.028}$. In Figure 1.24 are reported upper limits on the spin-independent WIMP-nucleon cross-section as a function of the mass of the WIMP candidates and is compared to constraints set by direct detection experiments, for such comparison the observed 90% CL level limit of 12.7% on $\mathcal{B}_{\text{inv.}}$ is used (the expected limit being 8.7% in that case).

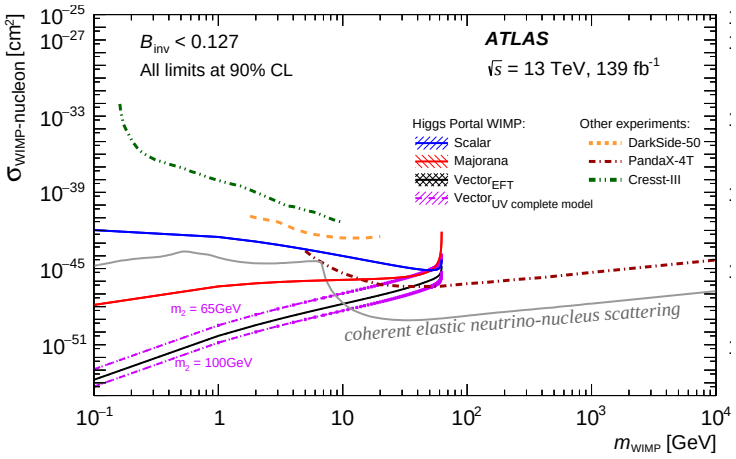


Figure 1.24: Upper limits on the spin-independent WIMP-nucleon cross-section using Higgs portal interpretations of $\mathcal{B}_{\text{inv.}}$ at 90% CL as a function of the mass of the WIMP candidates (m_{WIMP}). A selection of the most sensitive direct detection experiment is shown in this figure (taken from Ref. [109]

As a consequence, for all the latest measurements so far obtained by ATLAS and CMS, the combinations of several Higgs production and decay modes are all in good agreement with the SM predictions within uncertainties. Precision measurements are hence important to constrain even more BSM theories and search for hints of new physics. Reduction of uncertainties from an experimental and theoretical point of view are also a key feature. The development and use of new techniques to enhance sensitivity to signal and improve background rejection via for instance machine learning algorithms to classify events, tagging particles are also an active domain of those collaborations. Increasing the collected data will allow for even more precise measurement and to probe channels that are currently experimentally not accessible such as the diboson events which would allow measuring the Higgs self coupling.

1.7 Limitations of the Standard Model

Despite its impressive predictive power which has been confirmed in many occasions by experiments, some important questions and issues summarized below are raised by the Standard Model which could hint for a more fundamental and broader theory than our current understanding of Nature. Some problems listed below can be probed with Higgs measurements or are a motivation to test BSM theories through the Higgs property studies.

➢ **The dark matter and dark energy problem**: for instance, the rotational curve of galaxies [110, 111] cannot be explained based on the observable content of galaxies. However, their rotation can be described assuming the existence of dark matter which is not predicted by the SM and which is composed of non-electromagnetic interacting or weakly interacting particle(s) (such as hypothetical WIMPs) thus difficult to detect. Despite several congruent indirect observations

in astrophysics and cosmology, such particle(s) have never been directly observed. In case of heavy dark matter candidates, the Higgs could be a portal for dark matter candidates discovery thus the importance of Higgs invisible decay searches.

Moreover, the so-called inflation of the universe corresponding of the expansion of the universe [112] and the acceleration of this process with time measured with the redshift of galaxies is interpreted as the presence of dark energy in the Universe which again is not a prevision of the SM. As a result, currently around only 5% of the content of the Universe can be well understood thanks to the SM (or at least follows closely its prediction) while the nature and origin of dark matter and dark energy, representing respectively 26% and 69% of the content of the Universe, remain unsolved[4].

- ➢ **The grand unification**: the SM coupling constants of the weak, electromagnetic and strong interactions ($\alpha_{\text{weak}}, \alpha_{EM}, \alpha_s$) are running coupling constant that depends on the energy of the process considered. When looking at the trend of those coupling constants over a large range of energies it seems that in for high energy they would converge. The energies, for which two constants are equal, are not coinciding however they all seem to be between 10^{12} to 10^{16} GeV as shown in Figure 1.7. This can be interpreted as a unification into a single force of those interactions at high energies exactly as for the electroweak case and are decoupled in the low energy limit due to a spontaneous symmetry breaking. The Super Symmetry (SUSY) theories allow for such unification however, among other predictions, they imply extra Higgs bosons or enhancements in the 125 GeV Higgs couplings, which to date has never been observed nor other supersymmetric predicted particles.
- ➢ **The unification of the SM with the General Relativity theory**: the gravity, which is one of the four fundamental interactions, is described by the General Relativity (GR) theory but is not included in the SM theory. The main problem comes from large conceptual differences between those two theories. For the GR, spacetime is considered as curved over large scales (more precisely is treated as Riemanian or Minkowskian respectively for the General Relativity and the Special Relativity) but the matter in the Universe is treated in the classical approach *i.e.* it is considered as a continuum, while the SM is a quantum theory *i.e.* energy is quantized, matter is treated as non-continuous but spacetime is essentially considered as globally flat (Euclidian to be more precise). A search for a *theory of everything* describing those four fundamental interactions and solving those differences in concepts is a longstanding issue and led to the development of the Super String or the Quantum Loop Gravity theories which aim to find a quantized approach of the gravity theory. However, verifications of those theories are experimentally challenging if not unfeasible to date.
- ➢ **The CP violation and matter versus antimatter asymmetry**: the most widespread theory is that 13 billion years in the past from now our Universe was created by the so-called *Big Bang* which should have produced the same amount of matter and antimatter due to the almost exact observed and verified experimentally CP symmetry. If the CP symmetry was perfectly true then the processes that create matter would have to create equal amounts of matter and antimatter, and these would annihilate to leave a universe containing just photons. Such CP violation is allowed by the SM thanks to the additional phase terms contained in the CKM and PMNS matrices corresponding to weak interaction processes. However, currently the Universe is completely dominated by matter as confirmed by observations as if for instance some stars or galaxies would be currently composed of antimatter, the antiparticles annihilation with matter would emit a constant flux of photons coming from some direction in space, phenomena which is not observed. The issue is the CP violation from the SM is too small to explain such current matter-antimatter discrepancy. However, CP violation in the Higgs coupling could result in a larger CP violation than predicted by the SM which could potentially be the reason for the

4. The existence of dark matter and dark energy and their proportion in the Universe are determined assuming that the theory of the General Relativity and the measurements performed are correct. However, such observations could perhaps also be explained with, for instance, modified gravity theories or in case of unknown biases affecting astrophysical and cosmological measurements.

matter-antimatter asymmetry.

➢ **The Higgs mass problem**: as the Higgs boson is a scalar particle it is expected to receive large radiative loop corrections when including higher orders corrections from contributions loops of self-interactions, fermions and bosons. The Higgs mass ($m_H = 125$ GeV) measured experimentally can be expressed as a function of the bare Higgs mass m_{bare} obtained from the unrenormalized Lagrangian *i.e.* the uncorrected mass as [113]:

$$m_H^2 = m_{\text{bare}}^2 + \frac{y_t^2}{16\pi^2}\Lambda^2 + \mathcal{O}(m_{\text{weak}}^2), \tag{1.7.1}$$

y_t being the top quark Yukawa coupling ($y_t \approx 1$), Λ is the scale until which the SM is valid and $\mathcal{O}(m_{\text{weak}}^2)$ are all the other corrections. If no new phenomena are present up to the next known energy scale, which is the Planck scale ($\approx 10^19\,\text{GeV}$) at which quantum gravitational effects are expected to dominate with respect to other SM forces and if the SM is valid up to that energy, then the bare mass of the Higgs should be very large in order to cancel the large contribution from the Λ^2 term. It implies a very fine-tuning of the two terms on the right-hand side of the equation to reach the measured mass m_H of the Higgs which is extremely smaller than those energy scales. This fine-tuning is known as the *naturalness problem* as such fine-tuning seems unnatural and unlikely. This is one of the main reason that hints for new physics beyond the SM: the SM could an approximation in the low energy limit of a much broader theory. Hence, the importance of EFT reinterpretations in the Higgs sector to test if any deviations is observed above a certain energy scale.

➢ **The mass hierarchy and the hierarchy problem**: the mass of particles ranges from less than 1 electronvolt for neutrinos to hundreds of giga-electronvolts for the top quark, W and Z bosons ($m_t = 172.5$ GeV, $m_Z = 91.2$ GeV and $m_W = 80.4$ GeV) as presented in Figure 1.1a hence spanning over more than 11 order of magnitudes. Such impressive mass difference, with such ordering also observed between the fermionic families is not understood and cannot be predicted by the SM model as the Higgs mechanism and Yukawa coupling constants are free parameters of the Standard Model. Some BSM theories are trying to predict such hierarchy. Moreover, there are 19 orders of magnitude between the strong interaction to the Planck scale *i.e.* the energy scale for which gravity starts to play a role in quantum field theory. In addition, the weak interaction is approximately a factor 10^{24} stronger than the gravity. On the other hand, the SM relies on precise cancelation terms and an extreme fine-tuning (as explained in for Higgs mass problem) of its parameters which contrast with the large gap observed between interaction forces.

➢ **The neutrino masses**: the observation of the oscillation of neutrinos implies they have a mass. A corresponding mass term can be introduced in the Lagrangian without breaking the gauge invariance. Such term is different depending on the nature of the neutrinos as they can either be Dirac fermions *i.e.* their antiparticle is different from the particle or Majorana fermions the neutrino is its own antiparticle. Such question is unanswered to date but can be tested experimentally as if neutrino are of Majorana type then it would lead to the observation of neutrinoless double beta decay. Several experiments are searching for such decays but not evidence so far was found.

Chapter 2

The LHC and the ATLAS detector

2.1 The Large Hadron Collider . . . 57
2.1.1 Luminosity and number of interaction per bunch crossing . . . 58
2.1.2 Production and acceleration of particles . . . 60
2.1.3 Experiments of the LHC . . . 61
2.2 The ATLAS detector . . . 62
2.2.1 System of coordinates and physics related definitions . . . 63
2.2.2 Particle energy loss and shower development . . . 65
2.2.3 Magnets system . . . 67
2.2.4 Inner detector and tracking system . . . 69
2.2.4.1 The pixel detector . . . 71
2.2.4.2 SemiConductor Tracker . . . 72
2.2.4.3 Transition Radiation Tracker . . . 72
2.2.5 Electromagnetic and hadronic calorimeters . . . 73
2.2.5.1 Electromagnetic calorimeter . . . 73
2.2.5.2 Hadronic calorimeter . . . 75
2.2.6 Muon spectrometer . . . 76
2.2.7 Trigger and data acquisition . . . 77

In this chapter, an overview of the Large Hadron Collider is presented in Section 2.1 and more details about the ATLAS detector are provided in Section 2.2.

2.1 The Large Hadron Collider

The Large Hadron Collider [114–117] is the largest and most powerful accelerator of particles ever built to date. This circular hadron accelerator of 27 kilometers of circumference is located underground, at a mean depth of 100 meters, at the border between Switzerland and France. As illustrated in Figure 2.1, bunches of protons, forming two opposite direction beams, are accelerated up to almost the speed of light ($99.9999991\% \times c$) travelling round the ring more than 11 000 times per second. The LHC was designed to reach a maximal center of mass energy $\sqrt{s} = 14\,\text{TeV}$ *i.e.* an energy of 7 TeV per beam. Protons are eventually collided in 4 collision points where the ATLAS, CMS, LHCb and ALICE detectors are located: collisions are occurring every 25 ns. The LHC can also accelerate lead-ions (Pb^+) at a center of mass energy $\sqrt{s_{NN}} = 5.02$ TeV corresponding to an energy of 2.56 TeV per nucleon.

The LHC was built between 1998 and 2008 by the European Organization for Nuclear Research (CERN) as a replacement for the previous CERN's collider, the Large Electron-Positron (LEP) and used the same underground tunnel that was created for the LEP. The LHC started operating on the 10th of September 2008 but unfortunately, due to a "quench" problem that occurred on the 18th of September, some of its magnets were severely damaged and the Run 1 of the LHC had to

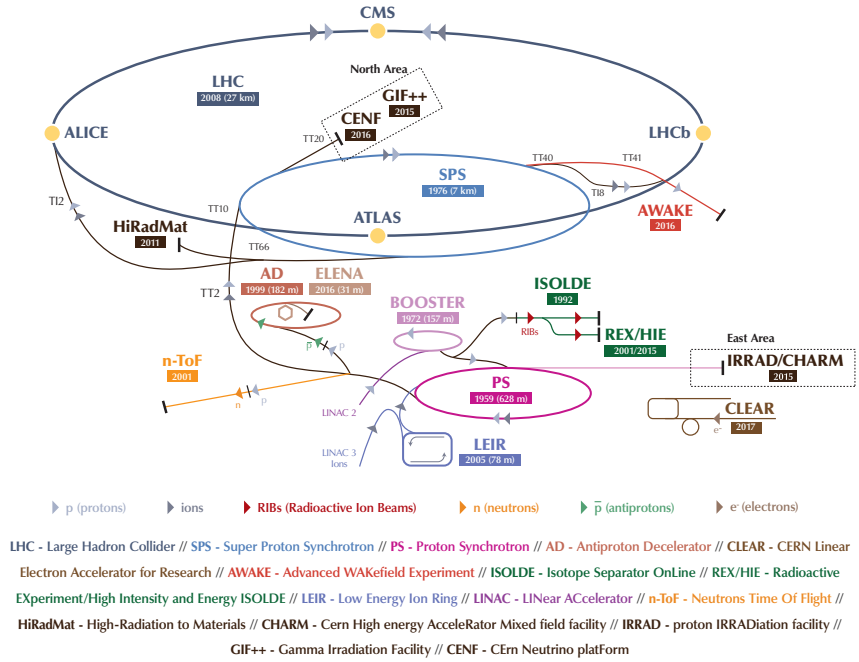


Figure 2.1: Scheme of the LHC. To be noted that the ALICE, ATLAS and LHCb detectors are in reality much closer to each other than represented in that scheme (taken from Ref. [118]).

be delayed. The Run 1 started in 2009 and lasted until the end of 2012 were the LHC operated at $\sqrt{s} = 7$ and 8 TeV. It was followed by a first period of two years of long shutdown (LS1) to repair and upgrade the different detectors whose performances can be altered by radiation. The Run 2 data taking took place from 2015 to 2018 at $\sqrt{s} = 13$ TeV before entering again in a long shut down phase (LS2). Finally, exactly ten years and one day after the Higgs discovery announcement (the 4$^{\text{th}}$ of July 2012), the Run 3 started on the 5$^{\text{th}}$ of July 2022 reaching an unprecedented energy of $\sqrt{s} = 13.6$ TeV.

2.1.1 Luminosity and number of interaction per bunch crossing

The instantaneous luminosity ($\mathcal{L}$) of an accelerator of particles can be defined as the number of particle crossings per unit of area and time, and it is related to the cross-section (σ) of a certain interaction process and the rate of events of such process (dN/dt) by:

$$\frac{dN}{dt} := \mathcal{L} \times \sigma \tag{2.1.1}$$

The cross-section of a process (expressed in cm^2) is proportional to the probability of a process to occur. The cross-section depends on the physics process considered and on the energy of the collisions as some resonances can appear or be suppressed: it is usually measured in barns (b) with $1\,\text{b} = 10^{-24}\,\text{cm}^2$. The luminosity $\mathcal{L}$ is expressed in $\text{cm}^{-2}\text{s}^{-1}$ or commonly in $\text{fb}^{-1}\,\text{s}^{-1}$.

The instantaneous luminosity is intrinsic to the accelerator and its settings. Assuming round and identical beams, for a circular collider it can be expressed as [119]:

$$\mathcal{L} = \frac{N_b^2 n_b f_r \gamma_r}{4\pi\varepsilon_n \beta^*} F, \tag{2.1.2}$$

where:

- N_b is the number of particles per bunch. When injected in the LHC, a bunch typically contains $N_b = 1.15 \times 10^{11}$ protons.

- n_b is the number of bunches per beam, $n_b = 2808$ for the LHC[1].
- f_r is the revolution frequency: $f_r = 11245$ Hz.
- γ_r is the relativistic gamma factor: $\gamma_r \approx 7000$.
- ε_n is the normalized transverse beam emittance: $\varepsilon_n = 3.75\mu\text{m}$. The emittance is defined through the invariance of the ellipse area enclosed by a single particle in the position-angle phase space in the transverse plane (Liouville's theorem) and can be generalized to a bunch of particles through the standard deviation (RMS).
- β^* is the so-called *beta function* at the collision point it is related to the focusing of the bunches: $\beta^* = 0.55$ m.
- F is the geometric luminosity reduction factor which accounts for non heads-on collision of the bunches. This parameter is equal to $F = 0.84$ for the LHC as bunches are collided with a crossing angle $\theta_c = 285$ µrad to prevent collisions outside the nominal interaction points. This factor can be computed with the RMS bunch length ($\sigma_z = 7.55$ cm) and transverse bunch width ($\sigma^* = 16.7$ µm) [119, 122]:

$$F = \left[1 + \left(\frac{\sigma_z}{\sigma^*}\tan\left(\frac{\theta_c}{2}\right)\right)^2\right]^{-\frac{1}{2}} \approx \left[1 + \left(\frac{\sigma_z\theta_c}{2\sigma^*}\right)^2\right]^{-\frac{1}{2}}. \tag{2.1.3}$$

For instance, to reduce the luminosity, the crossing angle can be increased or alternatively the transverse distance between the beams can be increased to limit the number of interaction per bunch crossing. That is how a reduction of the luminosity is achieved for the LHCb experiment as it would not be able to cope with the same luminosity obtained in ATLAS and CMS: the peak luminosity of LHCb is $\mathcal{L}(\text{LHCb}) = 10^{32}\ \text{cm}^{-2}\text{s}^{-1}$ to be compared with $\mathcal{L}(\text{ATLAS or CMS}) = 10^{34}\ \text{cm}^{-2}\text{s}^{-1}$. On the other hand in the ATLAS and CMS interaction points, the bunches are squeezed in order to increase the probability of collisions and increase the luminosity [116].

For a given period of time (T), the integrated luminosity ($\mathcal{L}_{\text{int}}$) is also commonly used and corresponds to the integration of the instantaneous luminosity:

$$\mathcal{L}_{\text{int}} := \int_0^T \mathcal{L}(t)dt. \tag{2.1.4}$$

During the Run 1, combining data collected at $\sqrt{s} = 7$ and 8 GeV, ATLAS and CMS collected respectively $25\,\text{fb}^{-1}$ of good quality data *i.e.* that can be used for physics analysis as all the relevant detector components are known to have been in good operating conditions [123]. For the Run 2, those two experiments collected $139\ \text{fb}^{-1}$ each as shown for ATLAS in Figure 2.2, while for Run 3, they are expected to collect around $300\ \text{fb}^{-1}$ of data each.

For $\sqrt{s} = 13$ TeV the pp inelastic collision cross-section is approximately equal to $\sigma_{\text{inel}} \approx 70$ µb [124] implying there are usually several pp collisions per bunch crossing. Those additional interactions are called *in-time pile-up*. The mean number of interaction (μ) per bunch crossing for the Run 2 of the LHC was varied between 13 and 38 over the years with a mean value of 33.3 interactions per bunch crossing summing all Run-2 data taking periods as shown for ATLAS in Figure 2.3. Thanks to the high granularity of detectors, it is possible to distinguish the different interactions vertices and associated produced particles. Furthermore, the value μ tends to decrease with time as bunches are losing some protons at each turn due to collisions. The instantaneous luminosity is hence decreasing over time mainly due to collisions until the beams are dumped (after several hours) and replaced. This decrease in luminosity can be modelled at first order as [119]:

$$\mathcal{L}(t) = \frac{\mathcal{L}_0}{(1 + t/\tau_{\text{nuclear}})^2}, \tag{2.1.5}$$

1. Since 2016, a new system called the Batch Compression Merging and Splitting (BCMS) [120, 121] enabling to reduce the beam size while keeping the same number of protons per bunch has been introduced. It has increased by 20% the instantaneous luminosity delivered by the LHC but as a result, the number of bunches for the LHC is now limited to 2556 compared to 2808 before.

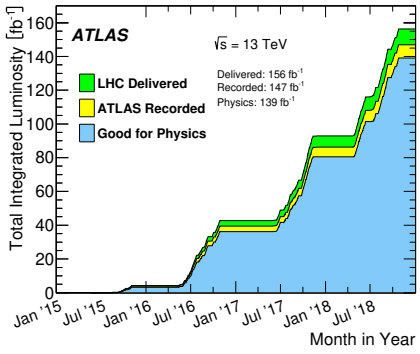


Figure 2.2: The LHC and ATLAS integrated luminosity evolution over time for the full Run 2 (taken from Ref. [123]).

with $\mathcal{L}_0 = 10^{34}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$ the peak luminosity and $\tau_{\mathrm{nuclear}} \approx 50\,\mathrm{h}$ being the decay time of the bunch intensity due to proton collision losses. Other smaller loss effects need to be taken into account such as emittance blow up, transverse growth of the beam size, rare collisions with remaining gas atom and so on implying that in practice the lifetime of the beam is around 15 hours for the LHC.

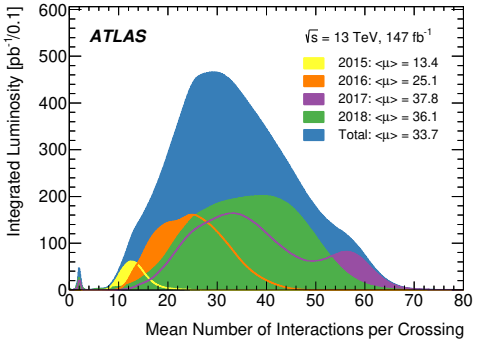


Figure 2.3: ATLAS luminosity-weighted distribution of the mean number (μ) of interactions per bunch crossing for the full Run 2 (taken from Ref. [123]).

2.1.2 Production and acceleration of particles

Protons for the LHC are produced by stripping electrons from hydrogen atoms by means of a strong electric field [116]. They are accelerated to 50 MeV in a linear accelerator called the Linac 2 to be afterwards injected successively in several circular accelerators of increasing size and increasing accelerating power. First they reach a kinetic energy of 1.4 GeV inside the Proton Synchrotron Booster (denoted Booster in Figure 2.1, abbreviated PSB) and then 25 GeV inside the Proton Synchrotron (PS). They are brought to an energy of 450 GeV by the Super Proton Synchrotron (SPS). Finally, they are injected in the LHC both in the clockwise and anticlockwise directions where it takes 20 minutes to accelerate them to their nominal energy of 6.5 TeV for Run 2. The acceleration chain for Run 3 is slightly different, the Linac 2 was replaced by a new linear accelerator (Linac 4) and acceleration energies are a bit different: 160 MeV for the Linac 4, 2 GeV for the PSB, 25 GeV for PS, 450 GeV for the SPS and 6.8 TeV for the LHC.

Protons are circulating in beam pipes which are kept at ultrahigh vacuum in order to avoid loss of the beam due to beam-gas interactions: the local pressure is required to be below 10^{-5} Pa. At the interaction points, the experiments are requiring even lower pressures, down to 2×10^{-9} Pa in order to minimize background coming from beam-gas interactions [125].

The LHC is technically composed of straight (respectively curved) portions for accelerating (re-

spectively bending trajectories and for the focusing) of charged particles (proton or ions) thanks to electric (respectively magnetic) fields:

- the acceleration is performed by radio-frequency (RF) cavities which at the same time are forming the bunch structure of the beam as particles are only accelerated when the electric field is oriented in the correct direction during the RF cycle. Per beam in the LHC, height superconducting RF cavities of 400 MHz increase the proton's energy by $485\ \mathrm{keV}$ at each turn until they reach their nominal energy. Afterwards the RFs simply compensate proton energy loss per turn.
- In order to bend trajectories of the bunches, 1232 superconducting dipole magnets of 15 meters long each operating at $1.9\ \mathrm{K}$ producing a magnetic field of $8.3\ \mathrm{T}$ are used per beam. They are made of Nb-Ti and have required extensive development to reach such powerful magnetic field. They are also equipped with sextupole, octupole and decapole magnets, to correct for small imperfections in the magnetic field at their extremities.
- In order to focus beams, 78 superconducting quadrupoles also made of Nb-Ti are used. They allow to keep the particles in a tight beam by squeezing it either vertically or horizontally depending on the orientation of the 4 poles of the magnets.
- The two previous set of magnets mentioned are called *lattice magnets*. LHC is also composed of *insertion magnets* which are used when particles are entering the detectors. Those magnets are strongly squeezing the bunches in order to increase the luminosity.
- Finally, there are approximately 3800 single aperture and 1000 twin aperture corrector magnets that are used for orbital corrections.

The high center of mass energies achieved at the LHC can only be reached accelerating relatively heavy particles, such as protons or heavy ions (lead ions), for which the energy loss in curved trajectories due to synchrotron radiation is much lower with respect to lighter particles such as electrons. For an ultra-relativistic particle of charge q, mass m whose energy E is kept constant inside a circular accelerator, the instantaneous synchrotron radiation power loss can be expressed as:

$$P = \frac{c}{6\pi\varepsilon_0} \cdot \frac{E^4}{\rho^2} \cdot \frac{q^2}{(mc^2)^4}, \tag{2.1.6}$$

with ρ being the curvature radius and ε_0 is the vacuum permittivity. The equation (2.1.6) is a particular case of the relativistic Larmor equation [126, 127]. From this equation follows that the energy loss (U_0) for a particle per turn is:

$$U_0 = \frac{1}{3\varepsilon_0} \cdot \frac{E^4}{\rho} \cdot \frac{q^2}{(mc^2)^4}. \tag{2.1.7}$$

Since a proton is roughly 2000 times heavier than an electron, the energy lost due to synchrotron radiation is considerably reduced for protons.

2.1.3 Experiments of the LHC

The LHC hosts four main detectors, as shown in Figure 2.1.

- **ATLAS (A Toroidal LHC ApparatuS)** [128–130]: ATLAS is one of the two general-purpose detectors of the LHC. It was designed to extensively probe the Standard Model and perform precision measurements going from Higgs boson properties studies to searches for new physics beyond the Standard Model (BSM) as for instance super-symmetry (SUSY) or searching for dark-matter candidates. This cylindrical detector is the largest detector ever constructed for a particle collider (46 meters long, 25 meters in diameter and weighting 7000 tonnes). As its name

indicates, the magnetic field in the outer part of the detector, for muon tracking, is produced by large toroidal magnets. The tracking part of the detector is contained in a solenoid magnet. More than 5500 persons from 245 institutions in 42 countries work for the ATLAS experiment.

- **CMS (Compact Muon Solenoid)** [131–133]: CMS is the other general-purpose detector of the LHC with very similar physics goals as ATLAS. The CMS detector is a cylindrical "compact" detector (21 meters long, 15 meters in diameter and weighting 12 500 tonnes) with a strong solenoidal magnetic field of 3.8 T. The difference in technologies and design in ATLAS and CMS is crucial to validate any measurement performed by one or the other experiment and reduce potential bias. The combination of measurements of these two experiments increases the precision and allows setting tighter constraints on the SM and BSM theories. CMS is involving over 5500 people from around 241 universities and institutes in 54 countries.
- **LHCb (Large Hadron Collider beauty)** [134–136]: unlike the other LHC detectors, LHCb is an asymmetric one-sided forward detector (21 meters long, 10 meters high, 13 meters wide and weighting 5600 tonnes). It has been purposely designed this way to study the asymmetry between matter and antimatter targeting bottom (also called beauty) and charm hadrons decays which are mainly produced in forward and backward regions *i.e.* close to the beam axis. LHCb is meant for very precise flavour physics measurement and CP violation studies. About 1600 persons from 20 countries are part of the LHCb collaboration
- **ALICE (A Large Ion Collider Experiment)** [137–139]: ALICE is a detector optimized for studying heavy-ion (lead-ion) collisions, and in particular properties of the quark-gluon plasma produced in those collisions which is a state of matter where quark and gluons are no longer confined inside hadrons. Those measurements reproduce what possibly existed right after the Big Bang before the quark-gluon recombination into usual hadrons such as protons and neutrons. ALICE is hence probing the quantum chromodynamics (QCD) theory and quark confinement. As ATLAS and CMS, ALICE is a cylindrical detector (26 meters long, 16 meters in diameter and weighting 10 000 tonnes). The ALICE collaboration is composed by more than 2000 members from 174 institutes in 40 countries.

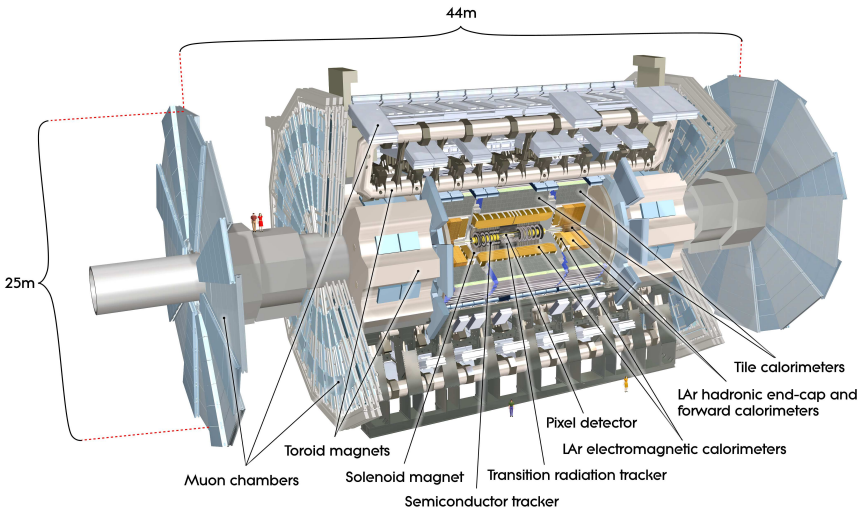
2.2 The ATLAS detector

The ATLAS detector [128–130] is shown in Figure 2.4. It has a cylindrical shape and is composed of several layers of sub-detectors designed to measure and reconstruct precisely energy, momentum and trajectories of particles as illustrated in the Figure 2.4b. To provide the most complete coverage in detecting particles the ATLAS detector is divided in two parts: the *barrel* which is the central part of the cylinder from the *endcaps* at each end (left and right) of the barrel-shaped detector.

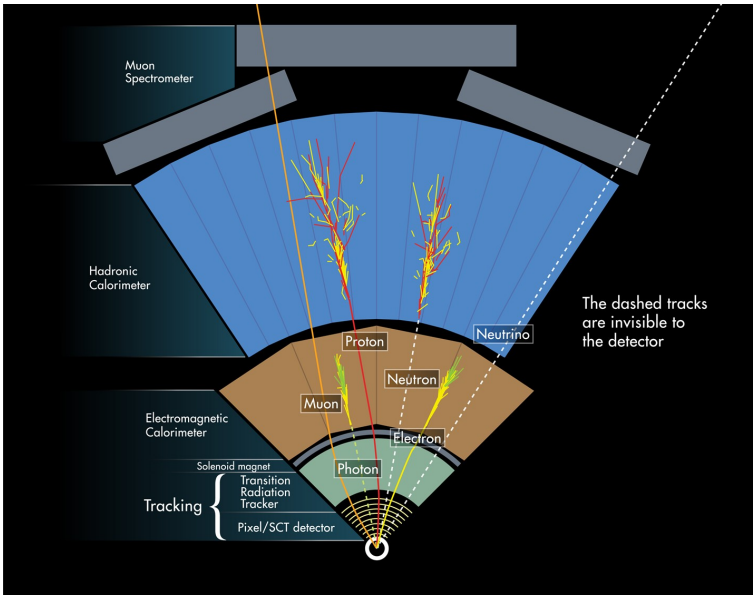
ATLAS is composed of the following nested sub-detectors, detailed in the incoming sections, which are listed from the closest to the interaction point to the furthest:

- **The inner detector (ID)**: is the closest detector to the beam-pipe. It is an essential sub-detector to reconstruct trajectories of charged particles (electron, protons, charged pions and so on), to precisely measure their momentum and impact parameters thanks to the bending of their trajectories (and curvature radius) due to the magnetic field created by the solenoid magnet surrounding the ID. It enables to discriminate positive from negative charged particles as their trajectories are bent in one or the other direction depending on their charge. The ID is also crucial for primary and secondary vertex reconstructions which are for instance needed to identify pp collisions of interest, for rejection of pile-up and for flavour-tagging algorithms.
- **The electromagnetic (ECAL) and hadronic (HCAL) calorimeters**: aims to respectively measure energies and directions of electromagnetic particles (electrons, positrons, photons…) and hadrons (protons, neutrons, pions…).

➢ **The muon spectrometer (MS)**: allows the identification and the measurement of the momenta of muons as such particles pass through all the other sub-detectors.



(a)



(b)

Figure 2.4: (a) ATLAS detector scheme and (b) transverse view of the detector. The bending of the trajectories for charged particles is due to the magnetic field inside ATLAS (taken from Refs. [140, 141]).

2.2.1 System of coordinates and physics related definitions

ATLAS uses a right-handed coordinate system with its origin at the nominal interaction point (IP) in the center of the detector as shown in Figure 2.5. The z-axis is along the beam pipe while the x-axis points from the IP to the center of the LHC ring, and the y-axis points upwards. The x-y plane is commonly called the transverse plane while a longitudinal plane refers to any plane containing the z-axis. Cylindrical coordinates (r, ϕ) are used in the transverse plane, $\phi \in [0, 2\pi[$ being the azimuthal angle around the z-axis. Finally, the angle formed with the beam axis ($\theta \in [0, \pi]$) is called the polar angle.

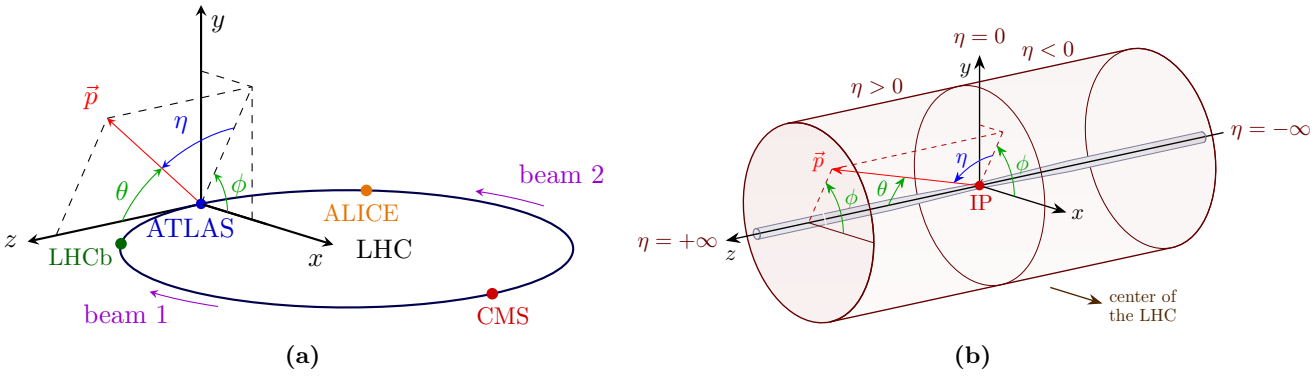


Figure 2.5: LHC scheme and system of coordinates for the ATLAS detector (scheme highly inspired by Ref. [45]).

The pseudorapidity, more used in ATLAS than the polar angle θ, is defined as:

$$\eta := -\ln\left[\tan\left(\frac{\theta}{2}\right)\right]. \tag{2.2.1}$$

The Figure 2.6 shows values of the pseudorapidity depending on the polar angle. Basically $\eta = 0$ in the transverse plane and $\eta = \pm\infty$ along the beam axis.

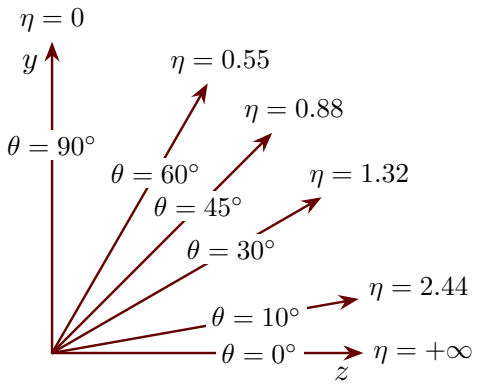


Figure 2.6: Correspondence between the pseudorapidity (η) and some polar angle (θ) values (taken from Ref. [45]).

Let $(E, \boldsymbol{p})$ be the four-momentum of a particle with E its energy and $\boldsymbol{p} = (p_x, p_y, p_z)$ the momentum of the particle p_x, p_y and p_z being the x, y and z components of the momentum ($\boldsymbol{p}$). Then the transverse momentum ($\boldsymbol{p_T}$) of that particle is defined as the projection of the momentum on the transverse plane: $\boldsymbol{p_T} = p_x\boldsymbol{e_x} + p_y\boldsymbol{e_y}$ with $\boldsymbol{e_x}$ and $\boldsymbol{e_y}$ being the normed vectors of the x and y axis. The transverse momentum norm (p_T) and $\boldsymbol{p}$ components satisfy the following equations:

$$p_T = \sqrt{p_x^2 + p_y^2} = |\boldsymbol{p}|\sin(\theta) \tag{2.2.2}$$
$$p_x = p_T\cos(\phi) = |\boldsymbol{p}|\sin(\theta)\cos(\phi) \tag{2.2.3}$$
$$p_y = p_T\sin(\phi) = |\boldsymbol{p}|\sin(\theta)\sin(\phi) \tag{2.2.4}$$
$$p_z = \frac{p_T}{\tan(\theta)} = |\boldsymbol{p}|\cos(\theta). \tag{2.2.5}$$

For an ultra-relativistic particle *i.e.* in the massless particle limit ($E \gg m$), its pseudorapidity is equivalent to its rapidity (y):

$$y := \frac{1}{2}\ln\left(\frac{E - p_z}{E + p_z}\right) = \tanh^{-1}\left(\frac{p_z}{E}\right). \tag{2.2.6}$$

The angular distance between two particles, which is a Lorentz invariant quantity, is defined as:

$$\Delta R := \sqrt{(\Delta y)^2 + (\Delta\phi)^2} \underset{\substack{\text{massless}\\ \text{particle}}}{\approx} \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}. \tag{2.2.7}$$

2.2.2 Particle energy loss and shower development

The stopping power of a material is one of the key feature for detectors to be able to measure energy of particles. This section briefly summarizes the different mechanisms that are involved in the energy loss of particles when they go through materials.

For ultra-relativistic electrons (or other low mass charged particles such as muons or charged pions), their main loss of energy in matter is due to bremsstrahlung[2]. High-energy photons lose their energy due to e^+e^- pair production[3]. The characteristic amount of matter traversed for these related interactions is called the radiation length (X_0), usually measured in $\mathrm{g\,cm^{-2}}$. The radiation length can also be expressed in centimeters: $X_0[\mathrm{cm}] = X_0[\mathrm{g\,cm^{-2}}]/\rho$ with ρ the density of the material. When expressed in centimeters, the radiation length is both the mean distance over which a high-energy electron loses $(1 - 1/e) \approx 63\%$ of its energy by bremsstrahlung, and 7/9 of the mean free path for pair production by a high-energy photon [18].

For a pure material, with atomic number $Z > 4$ and atomic mass A (assigned unit $\mathrm{g\,mol^{-1}}$), the radiation length can be estimated thanks to the Dahl's formula which is accurate to few percent [18, 142]:

$$X_0 = \frac{716.4A}{Z(Z+1)\ln\left(\frac{287}{\sqrt{Z}}\right)}\ \mathrm{g\,cm^{-2}}. \tag{2.2.8}$$

For moderately relativistic charged heavy particles of mass M, charge z with relativistic factor γ, the mean rate of energy loss ($\langle -dE/dx \rangle$) also called the mass stopping power, expressed in $\mathrm{MeV\,g^{-1}\,cm^2}$, of a material with atomic number Z, atomic mass A is well described by the Bethe-Bloch formula:

$$\left\langle -\frac{dE}{dx} \right\rangle = Kz^2 \frac{Z}{A}\frac{1}{\beta^2}\left[\frac{1}{2}\ln\left(\frac{2m_e c^2\beta^2\gamma^2 W_{\max}}{I^2}\right) - \beta^2 - \frac{\delta(\beta\gamma)}{2}\right], \tag{2.2.9}$$

with $K = 4\pi N_A r_e^2 m_e c^2$ a constant, $r_e = e^2/4\pi\varepsilon_0 m_e c^2 = 2.82$ fm the classical radius of the electron, $\delta(\beta\gamma)$ being the density correction due to the transverse extension of electric field of the particle for high $\beta\gamma$ values, $W_{\max} = 2m_e c^2\beta^2\gamma^2/(1+2\gamma m_e/M+(m_e/M)^2)$ is the maximal energy transfer to an electron in a single collision and I is the mean excitation energy of the material which can be approximated by $I = (10\ \mathrm{eV})Z$.

The Bethe-Bloch formula (Equation (2.2.9)) has an accuracy of few percent in the range $0.1 \lesssim \beta\gamma \lesssim 1000$. Basically moderately relativistic charged particles lose energy in a material by ionizing and exciting it. To be noted that for low energy particles the energy loss increases as β^{-2}: slow particles will be more ionizing than fast particles. On the other hand, the faster a particle is, the stronger the electric field it generates becomes and therefore the particle can ionize atoms at larger distances and loses more energy: that rise of the energy loss is logarithmic. Between those two regimes the energy loss of a particle reaches a minimum for $\beta\gamma \approx 3 - 4$ commonly referred to as minimum ionizing particle (MIP).

2. The bremsstrahlung is the radiation of photons by a charged particle inducing its deceleration when deflected by another charged particle as for instance an electron deflected by an atomic nucleus.
3. For a pair production process, an incident photon of sufficiently high energy is annihilated in the Coulomb field of a nearby charged particle, resulting in the creation of an electron-positron pair. Conservation of momentum requires the presence of a third body which takes up the balance of the momentum in the form of a recoil. For a recoil nucleus of mass much greater than the mass of the electron, the threshold energy for pair production is $E_{\mathrm{thresh}} = 2m_e c^2 = 1.022$ MeV.

The Figure 2.7 summarizes the two effects discussed above: at very high energy a charged particle predominantly looses its energy due to radiative bremsstrahlung effects while at lower energies the energy is transferred by ionization, excitation and collision with the material.

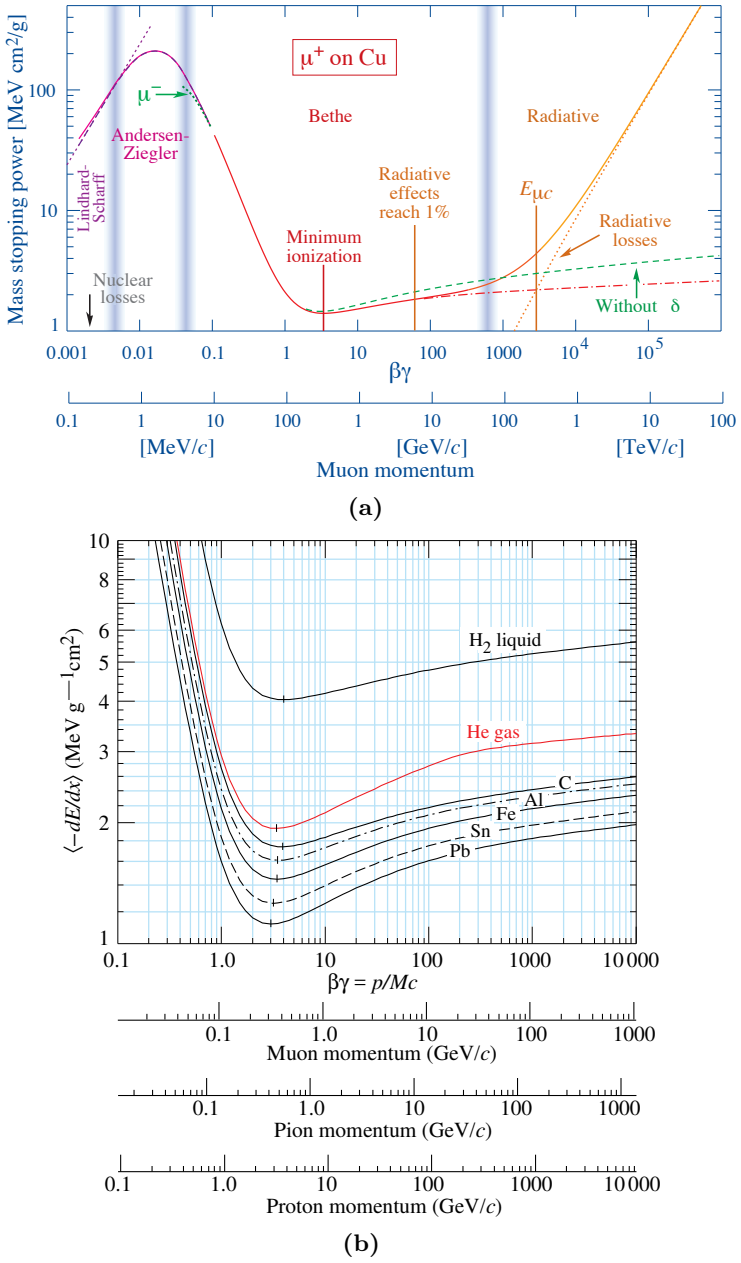


Figure 2.7: (a) Stopping power ($\langle -dE/dx \rangle$) for positive muons in copper as a function of $\beta\gamma$. The solid line indicates the total stopping power. Vertical bands indicate boundaries between different approximations. (b) Energy loss for muons, pions and protons in various types of materials (taken from Ref. [18]).

As a result when an energetic particle go through a thick enough material, it is going to ionize the medium creating several other particles of lesser energy. Each of these produced particles then interacts in the same way creating a cascade phenomenon called a particle shower. The cascade stops when the particles created do not have enough energy to ionize the material and are then stopped by the material.

The shower development in particular its longitudinal and transverse profile as well as energy loss profile can help identify the characteristics of the incident particle (nature, initial energy and so on) that created the shower.

For instance, for electromagnetic cascade (*i.e.* initiated by electron, positron or photon), the transverse radius of the shower R_M called the Molière radius, containing 90% of the shower's energy deposition, can be expressed as [18, 143, 144]:

$$R_M = X_0 E_s / E_c, \tag{2.2.10}$$

where $E_s \approx 21\text{MeV}$ and E_c is the critical energy for which the particle radiative energy loss is equal to its ionization energy loss. The $2R_M$ radius contains 95% of the shower's energy deposition.

Moreover, the longitudinal profile of the energy deposition in an electromagnetic shower, at energies between 1 GeV and 100 GeV can be described by:

$$\frac{dE}{dt} = E_0 b \frac{(bt)^{a-1} e^{-bt}}{\Gamma(a)}, \tag{2.2.11}$$

with $t = x/X_0$ the "normalized" depth of the shower, E_0 the energy of the incident particle, Γ is the gamma function[4], a and b are free parameters ($b \approx 0.5$ is a good approximation). It implyies that a shower reaches its maximal energy deposition at a depth ($t_{\max} = x_{\max}/X_0$) of:

$$t_{\max} = \frac{a-1}{b} = \ln\left(\frac{E_0}{E_c}\right) + C_{e\gamma}, \tag{2.2.12}$$

with $C_{e\gamma} = +0.5$ for photon-induced showers or $C_{e\gamma} = -0.5$ for electron-induced (or positron-induced) showers. Finally, the depth (L) containing 95% of the shower energy can be approximated by:

$$L = (t_{\max} + 0.08Z + 9.6)X_0. \tag{2.2.13}$$

2.2.3 Magnets system

A particle of charge q moving with a velocity $\boldsymbol{v}$ in an electric and magnetic field $\boldsymbol{E}$ and $\boldsymbol{B}$ is experiencing the Lorentz force ($\boldsymbol{F}$):

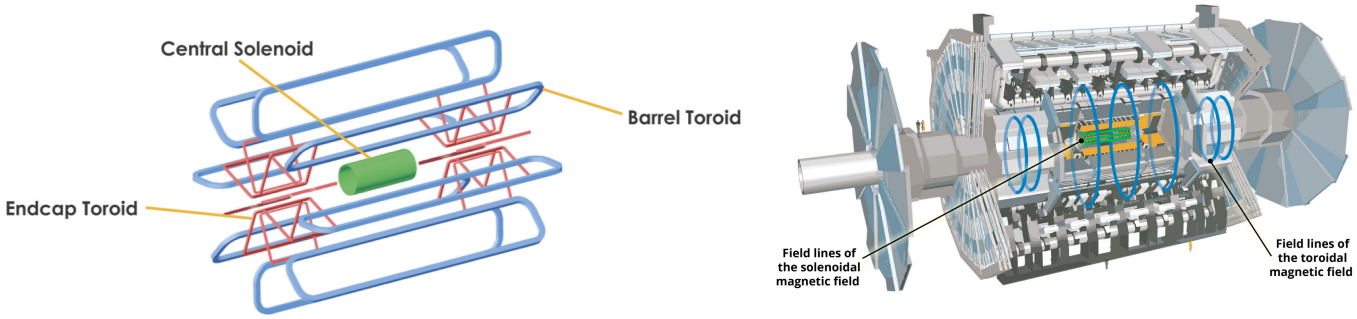
$$\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}). \tag{2.2.14}$$

Magnetic fields are thus playing a leading role for particle detectors as the charge (more precisely the ratio momentum over charge (p/q)) of a particle can be measured thanks to the bending of its trajectory.

ATLAS uses a hybrid magnet system [128] composed of a central solenoid magnetic field of 2 T. The solenoid magnet surrounds the inner detector hence creating a magnetic field parallel to the beam axis in the tracking part of the detector. A toroidal magnetic field *i.e.* rotating is created in the outer part of the detector, as shown in Figure 2.8, by the barrel and endcap toroids which are producing respectively a 3.9 T and 4.1 T magnetic field. The magnetic field inside ATLAS is hence non-homogenous. All the magnets are cooled with liquid helium.

The central solenoid is a cylinder of 5.8 meter length, 2.46 meter inner diameter and 2.56 meter outer diameter, whose magnetic field is induced by a nominal current of 7.73kA cooled to 4.5K. To achieve the desired calorimeter performance, the layout of the solenoid was carefully optimized to keep the material thickness in front of the calorimeter as low as possible, resulting in the solenoid assembly contributing in total to around 0.66 radiation lengths for particles arriving at normal incidence. The magnetic field flux is returned by the steel of the hadronic calorimeter structure. To be noted that the solenoidal magnetic field drops steeply in the inner detector at its extremity as shown in Figure 2.9.

4. The gamma (Γ) function is defined as $\Gamma(z) := \int_0^{+\infty} t^{z-1} e^{-t} dt$. For an integer $n \geq 1$ then $\Gamma(n) = (n-1)!$.



(a) **(b)**

Figure 2.8: (a) Schemes of the solenoid, barrel and endcap toroid magnets and (b) magnetic field lines they produce. The solenoid is creating a magnetic field parallel to the beam axis inside the tracker (straight magnetic field lines in green). The toroidal magnets are creating the rotating magnetic field (blue lines): the large radius central magnetic field lines are produced by the barrel toroid magnets while the smaller external ones are the sum of the endcap magnets field and the barrel toroid one (taken from Refs. [145, 146]).

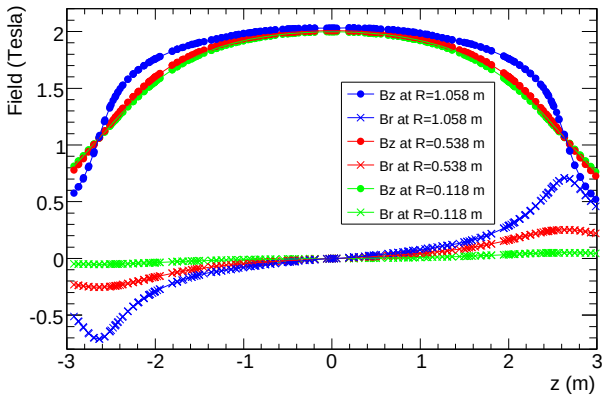


Figure 2.9: Radial (R) and longitudinal (z) dependence of the radial (B_r) and axial (B_z) magnetic field components in the inner detector cavity (taken from Ref. [128]).

The barrel toroid creates the magnetic field in the cylindrical volume surrounding the calorimeters but also in the two endcaps. The barrel toroid consists of eight individual stainless-steel coils with a nominal current of $20.5\ \mathrm{kA}$ circulating inside their "racetrack-shape": they are cooled down to $4.6\ \mathrm{K}$. It has an overall size of 25.3 meters in length, with respective inner and outer diameters of 9.4 meters and 20.1 meters. The net Lorentz forces generated per coil is approximately 1400 tonnes [128].

The endcap toroids generate the magnetic field required in the endcap region for optimizing the endcap muon spectrometer system. To improve the uniformity and the overlap across their magnetic fields and the barrel magnetic field, they are rotated of $22.5°$ with respect to the barrel toroid.

2.2.4 Inner detector and tracking system

The inner detector (ID) of ATLAS [147], schematized in Figure 2.10, covers the region $|\eta| < 2.5$. It is composed of 3 sub-detectors: the pixel detector, the Semi-Conductor Tracker (SCT) and the Transition Radiation Tracker (TRT). The inner detector is divided into three regions: the barrel and the two endcaps. In the barrel section the sub-detectors are arranged in concentric cylinders around the beam axis, while in the endcaps they are composed of disks perpendicular to the beam axis as presented in Figure 2.10c.

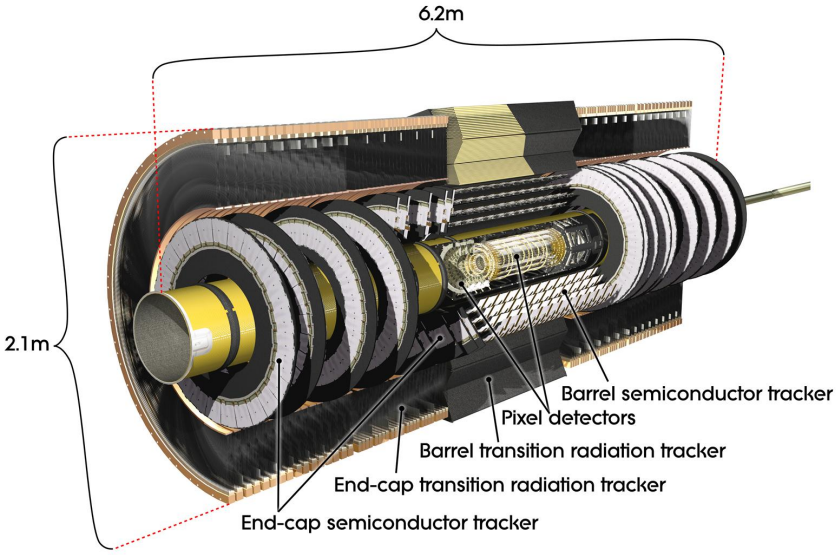
The ID is designed to identify charged particles and measure their momentum thanks to the $2\ \mathrm{T}$ solenoidal magnetic field that is bending their trajectory [128]. It provides electron identification for $|\eta| < 2.0$ for energies between $0.5\ \mathrm{GeV}$ and $150\ \mathrm{GeV}$ which is complementary to the information from the calorimeters. With its initial design (TDR, [147]), the inner detector has a track momentum resolution uncertainty [148] integrated over $|\eta|$ of:

$$\sigma\left(\frac{1}{p_{\mathrm{T}}}\right) = A \oplus \frac{B}{p_{\mathrm{T}}\sqrt{\sin(\theta)}} = \left(0.00036 \oplus \frac{0.0013}{p_{\mathrm{T}}\,[\mathrm{GeV}]\sqrt{\sin(\theta)}}\right)\ \mathrm{GeV}^{-1}, \tag{2.2.15}$$

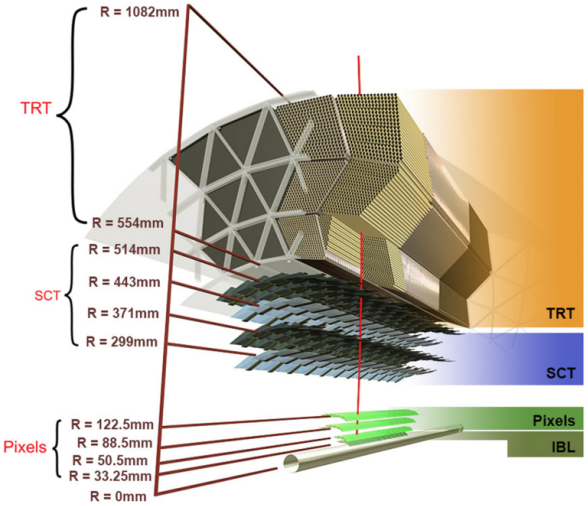
with A being linked to the intrinsic error on the track parameters and B represents the multiple scattering error.

The inner detector is highly segmented into cells in order to efficiently reconstruct and distinguish all collisions arising from pile-up. It enables to reconstruct both primary and secondary vertexes measuring charged particles tracks with transverse momenta above $0.5\ \mathrm{GeV}$.

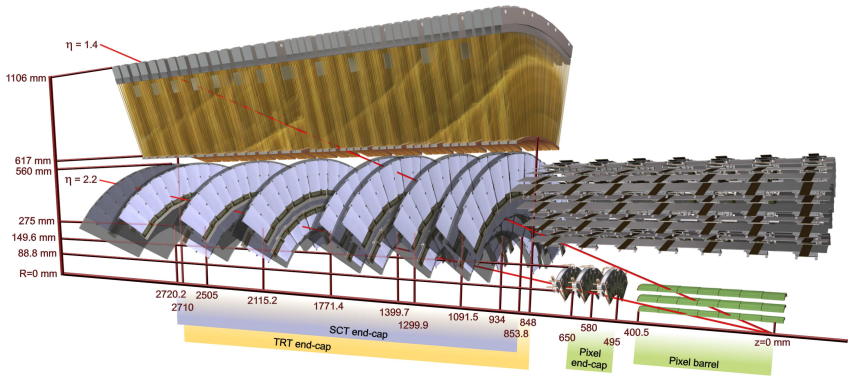
The inner detector layout is a trade-off between high performance and the thickness of the 3 sub-detectors which affects momentum and energy resolution measurement of the calorimeters due to multiple scattering. It has been designed to withstand high irradiation levels and has a maximum radiation length of 2.5 X_0 in the forward region as illustrated in Figure 2.11.



(a)



(b)



(c)

Figure 2.10: (a)Three dimensional scheme of the inner detector of ATLAS. (b) Transverse and (c) longitudinal views of the inner detector of ATLAS. The Insertable B-Layer (IBL) is only shown in the transverse view (taken from Refs. [128, 149, 150]).

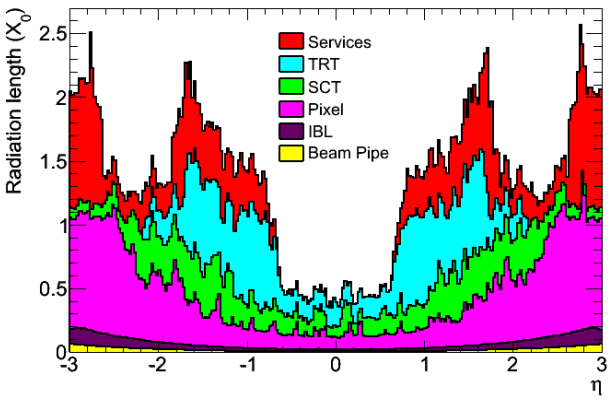


Figure 2.11: Radiation length (X_0) as a function of η for the different inner detector components (taken from Ref. [151]).

2.2.4.1 The pixel detector

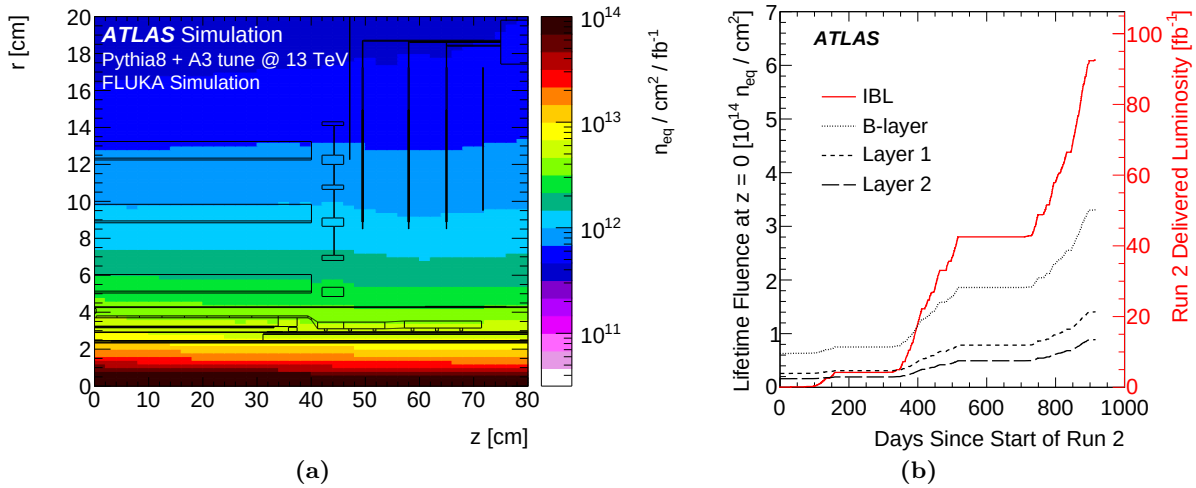
The pixel detector is the closest to beam-pipe sub-detector of ATLAS. It is composed of 92 million pixels divided in 4 silicon high granular ϕ and z segmented layers: the Insertable B-Layer (IBL) and 3 pixel layers referred to as b-layer, layer 1 and layer 2. The temperature of the silicon cells is maintained below $0\,^{\circ}\mathrm{C}$.

Typical particles at the LHC are nearly minimum-ionizing particles (MIP) [152]. Hence, the charged particles traversing the sensor cells deposit energy by ionizing the silicon as described by the Bethe-Bloch formula from Equation (2.2.9). Due to the semiconductor nature of the silicon, electron-holes pairs[5] are created along the particles path: for silicon the energy required to create such pair is $3.6\mathrm{eV}$. The electron-hole pairs are then drifting through the sensor due to the electric field created by the reverse bias voltage applied to the electrodes on the two surfaces of the silicon modules: this analogue signal is collected by the electrodes, digitized and read out using chip. On the other hand non-ionizing interactions caused by heavy particles and nuclei, lead to radiation damage in the sensor. Those type of interactions are affecting the sensor mainly causing displacement of a silicon atom out of its lattice site resulting in a silicon interstitial site and a leftover vacancy (Frenkel pair): those created defaults in the sensor affect the detection of MIPs as it reduces signal collection efficiency due to charge trapping. It also increases the sensor leakage current. As a result, the sensors are designed to resist high radiation especially the IBL and b-layer which are the closest from the interaction point: the layer 1 and 2 receive less than half of the radiations of the two former layers as presented in Figure 2.12.

The IBL [151, 153, 154] is the innermost layer of the pixel detector. This unique silicon layer, installed at a radius of 3.3 centimeters from the beam axis, is offering a coverage of $|\eta| < 3$. It was added during the long shut down between Run 1 and Run 2, in between the existing pixel detector layers and a new smaller radius beam-pipe. It increases the robustness of the tracking against dead pixel modules, the tracking precision and performances to cope with the increased peak luminosity of Run 2. It also improves the discrimination between b-jets and jets originating from gluons and light quarks as for instance the resolution of the track impact parameter[6], for tracks with $p_{\mathrm{T}} < 1\ \mathrm{GeV}$, was improved by about 40%. It is composed of a total of 12 million silicon cell pixels with size $50 \times 250\ \mu\mathrm{m}^2$ per cell for the barrel which is around 60% of the pixel size for the b-layer and layers 1 and 2. Two different sensor technologies were adopted: n^+-in-n planar sensors [155] and 3D n^+-in-p sensors [156]. Those cells can withstand a fluence[7] of $5 \times 10^{15}\ \mathrm{n_{eq}\,cm^{-2}}$ and a peak luminosity of the order of $10^{34}\ \mathrm{cm^{-2}\,s^{-1}}$ which are the expected conditions for the end of Run 3. Thanks to its granularity, the IBL achieves a transverse hit spatial resolution of $10\ \mu\mathrm{m}$ and $66.5\ \mu\mathrm{m}$ for the transverse (r-ϕ) and longitudinal (z) directions.

The b-layer, layer 1 and layer 2 are respectively located at 5.1, 8.9 and 12.3 centimeters from the beam pipe axis [128, 151]. They are composed of $50 \times 400\ \mu\mathrm{m}^2$ silicon cells. Each of them has a transverse hit spatial resolution of $10\ \mu\mathrm{m}$ and $115\ \mu\mathrm{m}$ for the transverse (r-ϕ) and longitudinal (z) directions for the barrel while for the endcaps it is $10\ \mu\mathrm{m}$ in the transverse (r-ϕ) direction and $10\ \mu\mathrm{m}$ in the radial (R) direction.

5. An electron-hole pair is created when an electron moves from the valence band to the conduction band of a semiconductor such as silicon.
6. The impact parameter is defined as the distance of closest approach of the track to the collision point. b-jets are characterized by higher impact parameters than light-jet, c-jets and gluon-jets which is a key feature for b-tagging algorithm. More details are provided in Chapter 3
7. The fluence also called radiant exposure is the number of particles traversing a unit area in a certain point in space over a certain period of time. For arbitrary particles with a specific energy distribution, if the fluence is expressed in $\mathrm{n_{eq}\,cm^{-2}}$ it is called the "neutron equivalent fluence": it corresponds to the fluence required for 1 MeV neutrons in order to produce the same damage in a detector material as induced by the arbitrary particles.



(a) (b)

Figure 2.12: (a) Simulated 1 MeV n_{eq} radial and longitudinal fluences for a slice ($z > 0$ and above the beam) of the detector. (b) Predictions for the lifetime fluence experienced by the four layers (IBL, *b*-layer, layer 1 and layer 2) of the ATLAS pixel detector as a function of time since the start of Run 2 at $z = 0$ up to the end of 2017 (NB: Run 2 finished in 2018). For the IBL, the lifetime fluence is only due to Run 2 and for the other layers, the fluence includes all of Run 1. The IBL curve represents both the fluence on the IBL (left axis) and the delivered integrated luminosity in Run 2 (right axis) (taken from Ref. [152]).

2.2.4.2 SemiConductor Tracker

The Semiconductor Tracker (SCT) [128, 157], shown in Figure 2.10, is a precise silicon microstrip detector that surrounds the pixel detector. It is used to reconstruct the tracks of charged particles thanks to 4088 p^+-on-n type silicon sensor modules and plays an important role in measuring their transverse momenta. It is composed of 4 cylindrical barrel layers ($|\eta| < 1.4$ region) with a hit spatial resolution of 17 µm and 580 µm for the transverse (r-ϕ) and longitudinal (z) directions. Each endcap ($1.4 < |\eta| < 2.5$) contains 8 planar discs per endcap which achieve a resolution of 17 µm in the transverse (r-ϕ) direction and 580 µm in the radial (R) direction. The SCT sensors are cooled to about $-7\,^{\circ}$C with C_3F_8 fluid.

2.2.4.3 Transition Radiation Tracker

The last layer of the inner detector is the Transition Radiation Tracker (TRT) [128, 158, 159] which is a straw-tube tracker. It consists of drift tubes with a diameter of 4 millimeters filled with a gas mixture of 70% Xe, 27% CO_2 and 3% O_2. Each tube contains a thin tungsten wire of 31 µm at its center which acts as a small proportional counter collecting charges as a voltage of around 1.5 kV is applied between the straw wall and the tungsten wire.

When a charged particle enters the TRT, it ionizes the active gas inside the straws. The resulting free electrons drift towards the tungsten wire where they are amplified thanks to the high voltage applied: a detectable current signal is hence induced.

The TRT barrel region (range of $|\eta| < 1$) contains 52 544 straw tubes of 1.5 meters length that are parallel to the beam axis. The wires are split electrically in order to be able to read out charges at both extremities of the straw. The TRT endcaps ($1 < |\eta| < 2$) contain radial 0.4 meters long straws that are arranged perpendicularly to the beam axis. Each endcap has of 122 880 straws that are read out at their outer end.

2.2.5 Electromagnetic and hadronic calorimeters

After the pixel detector, comes the calorimeter of ATLAS, shown in Figure 2.13, which is covering the region $|\eta| < 4.9$. It is also divided in two parts the barrel and the endcap calorimeters. The calorimeters are further split into two sub-detectors: the inner layer is the electromagnetic calorimeter (ECAL) and the outer layer is the hadronic calorimeter (HCAL).

Calorimeters are designed to absorb and measure energy of most known particles (except muons and neutrinos), produced by the proton-proton collisions, forcing them to deposit all of their energy and stop within the detector. Each calorimeter is composed of successive layers of "inactive" absorbing high-density material meant to stop incoming particles followed by "active" material that measures particles energy.

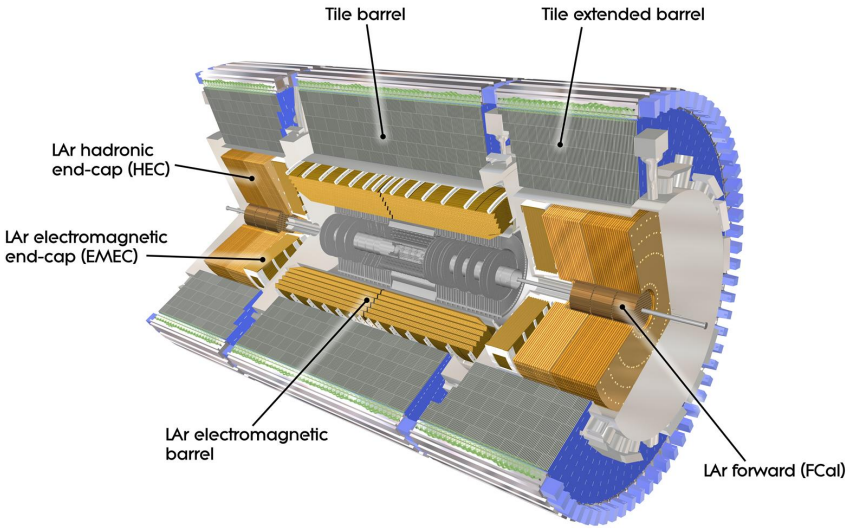


Figure 2.13: Scheme of the ATLAS calorimeter composed of electromagnetic and hadronic calorimeters (taken from Refs. [128, 160]).

2.2.5.1 Electromagnetic calorimeter

The electromagnetic calorimeter (ECAL) [128, 161] is a lead-liquid argon (LAr) sampling calorimeter with accordion geometry as presented in Figure 2.14. The calorimeter is divided in two parts the central ($|\eta| < 1.475$) barrel calorimeter and the two endcaps ($1.375 < |\eta| < 3.2$) containing each an electromagnetic endcap calorimeter (EMEC). The transition region ($1.37 < |\eta| < 1.52$) between the barrel and the endcaps calorimeter has a large amount of material upstream of the first calorimeter layer as illustrated in Figure 2.11. This section is instrumented with scintillators located between the barrel and endcap cryostats, and extending up to $|\eta| = 1.63$. The ECAL measures energy and direction of electromagnetic showers.

Liquid argon has been chosen as the active medium because of its intrinsic linear behaviour, its stability of response over time and its intrinsic radiation-hardness. The absorbers are made of lead plates, to which two stainless-steel sheets of 0.2 millimeters thickness are glued to provide mechanical strength. In the barrel, the lead plates are 1.53 millimeters thick for $|\eta| < 0.8$ and of 1.13 millimeters thick for $|\eta| > 0.8$. The change in lead thickness at $|\eta| = 0.8$ is to limit the decrease of the sampling fraction for increasing $|\eta|$. In the endcap calorimeters, the plates have a thickness of 1.7 millimeters for $|\eta| < 2.5$ and of 2.2 millimeters for $|\eta| > 2.5$. The readout electrodes are located in the gaps between the absorbers.

The accordion geometry of the calorimeter provides a full coverage around ϕ. It maximizes the number of interaction particles have with the absorber and active layers. It also enables a fast extraction of the signal at the rear or at the front of the electrodes. The total thickness of the barrel calorimeter varies between $22X_0$ to $33X_0$ as $|\eta|$ increases while it passes from $24X_0$ to $38X_0$ for the endcap calorimeters. In the barrel, the accordion waves are axial and runs in φ as depicted in Figure 2.14. The folding angles of the waves are varied with the radius in order to keep the liquid-argon gap constant. While in the endcaps, the waves are parallel to the radial direction and run axially. For the same reason, the wave amplitude and the folding angle of the absorbers and electrodes vary with radius. Thanks to all the characteristics cited above, very uniform performances in terms of linearity and resolution are obtained with respect to ϕ.

As illustrated in Figure 2.14, the barrel calorimeter is composed of 3 layers (front, middle and back) also called samplings. In the region $|\eta| < 1.8$, a pre-sampler LAr layer (before the front layer) is used to estimate the energy lost by electrons and photons upstream of the calorimeter. This layer is respectively 1.1 centimeters thick and 0.5 centimeters thick in the barrel and endcaps regions. The front layer has a $4.3X_0$ depth and a fine granularity along η that was optimized to differentiate neutral pions (π^0, decaying into 2 photons) from photons. The middle layer of $16X_0$ depth is the thickest: it collects most of the showers energy. The back layer of $2X_0$ depth is the thinnest it usually measures the tails of the shower but is also useful to estimate the amount of energy leaking beyond the ECAL.

Finally, the relative energy resolution of the ECAL is parametrized by

$$\frac{\sigma(E)}{E} = \frac{a}{\sqrt{E}} \oplus \frac{b}{E} \oplus c, \tag{2.2.16}$$

with a being the *sampling term* which is related to stochastic fluctuations in showers development in the calorimeter, it is modeled by simulation and was also measured in test-beams before the data taking: a good agreement between test-beam data and the simulation was found. b is referred to as *noise term*, it accounts for the electronic and pile-up noises measured in calibration runs. And c is the *constant term*, it includes non energy dependent contributions such as radiation damages, non uniformities of the detector, longitudinal leakage. Those three parameters are η dependent and their importance depends on the incident particle energy. For high energies, the constant term dominate as the two other terms are suppressed due to their energy dependence. At relatively low energy (few GeV), the sampling term is dominating. At lower energies, the noise term is dominating. The designed value for the sampling term and constant term are $a = 10\%$ and $c = 0.7\%$. The noise term was measured to be equal to $b = 350 \times \cosh(\eta)$ MeV for a 3×7 clusters in the barrel for a mean number of interaction per bunch crossing $\mu = 25$ [162].

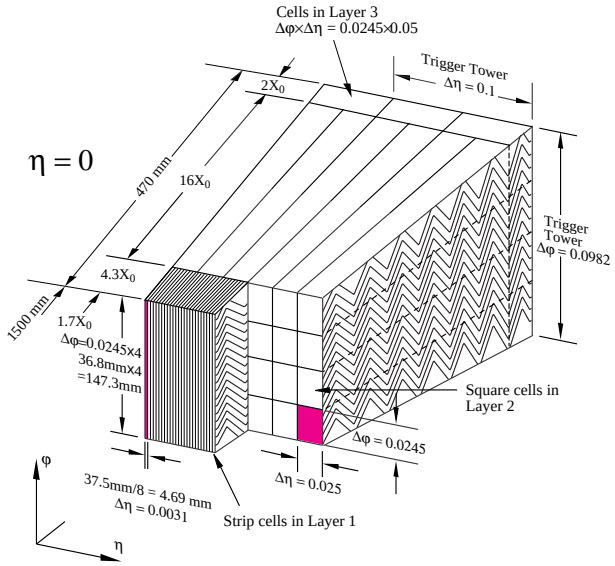


Figure 2.14: Scheme of a barrel module. The granularity in η and ϕ of the cells of each of the three layers and of the trigger towers is also shown (taken from Ref. [128]).

2.2.5.2 Hadronic calorimeter

The hadronic calorimeter (HCAL) [128] measures the energy of hadrons (protons, neutrons, pions and so on) which do not deposit all of their energy in the electromagnetic calorimeter (ECAL). This tile calorimeter surrounds the ECAL: it is divided in 3 regions the central barrel hadronic calorimeter ($|\eta| < 1.7$) called the Tile Calorimeter (TileCal), the Hadronic Endcap Calorimeters (HEC) in the $1.5 < |\eta| < 3.2$ region and the LAr hadronic Forward Calorimeter (FCal) covering $3.1 < |\eta| < 4.9$.

The TileCal uses steel as absorber and plastic scintillator as active medium. When particles hit the layers of steel, they generate a shower of new particles. The plastic scintillators in turn produce photons, which are converted into an electric current, by a photomultiplier, whose intensity is proportional to the original particle's energy. The hadronic calorimeter contains 420 000 plastic scintillator tiles and 9 500 photomultipliers.

The TileCal, which has a radial depth of $7.4X_0$, is composed of one central barrel and two extended barrels as shown in Figure 2.13. It is made of 64 modules, one module is schematized in Figure 2.15, fully covering the ϕ direction. Each tile is placed perpendicularly to the beam axis and spans radially.

The hadronic endcap calorimeters (HEC) are located behind the electromagnetic calorimeter endcaps (EMEC). This copper/liquid argon absorber/active material sampling calorimeter is composed of two cylindrical wheels per endcap and overlaps with the forward calorimeter. Each wheel is disposed perpendicularly to the beam axis, measures 4 meters of diameter and is constructed of 32 identical wedge-shaped modules.

The LAr hadronic forward calorimeter (FCal) is installed nearby the beam axis next to the endcap calorimeters. The close vicinity of these systems minimizes energy losses in cracks between the calorimeter systems but also prevents most of the backgrounds from reaching the muon spectrometer. The FCal contains 3 modules in each endcap: one electromagnetic module (FCal1) followed by two hadronic modules (FCal2 and FCal3). Copper was chosen as the absorber for FCAL1 while tungsten was chosen in FCal2 and FCal3 which provides containment and minimize the lateral spread of hadronic showers.

When combining the TileCal or HEC with the electromagnetic calorimeters for the measurement

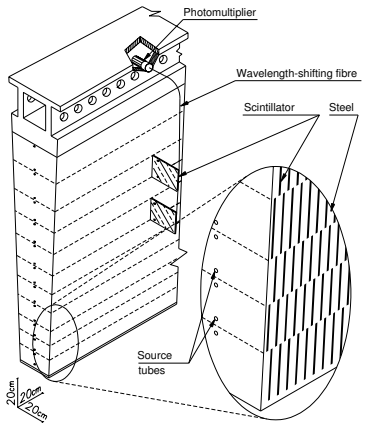


Figure 2.15: Hadronic tile calorimeter module scheme. The components of the optical readout (the tiles, the fibers and the photomultipliers) are shown (taken from Ref. [128]).

of hadronic particles, the following energy resolution for jets are obtained [128]:

$$\frac{\sigma(E)}{E} \approx \frac{50\%}{\sqrt{E}} \oplus 3\%, \tag{2.2.17}$$

while the FCal resolution is

$$\frac{\sigma(E)}{E} \approx \frac{100\%}{\sqrt{E}} \oplus 10\%. \tag{2.2.18}$$

2.2.6 Muon spectrometer

As muons are heavier than electrons, the energy radiated by bremsstrahlung effect is reduced: they usually pass through the inner detector and calorimeters with minimal energy loss.

The muon spectrometer (MS) [128] is surrounding the calorimeters. It is divided in 2 regions: the barrel spectrometer for $|\eta| < 1.2$ and the endcap spectrometers ($1 < |\eta| < 2.7$). The MS is also designed to trigger for muons in the region $|\eta| < 2.4$. For muons with energy greater or equal to 3 GeV, this spectrometer is able to measure on its own, their direction, electric charge and momentum thanks to the bending of their trajectory. For energies below 3 GeV, the energy loss in the calorimeter can prevent such "standalone" measurement and requires adding the information from the tracks reconstructed in the inner detector. The muon spectrometer has an excellent charge identification.

The toroidal magnetic field is orthogonal to the muon trajectory in the r-ϕ plane. Hence, muons trajectories are bent in the R-z plane which minimizes the degradation of the resolution due to multiple scattering when muon pass through the MS layers.

The MS is composed of four different detectors represented in Figure 2.16 and listed below: the tracking chambers (MDT and CSC) are meant for precise momenta measurement while the triggering chambers (RPC and TGC) have fast responses (few tens of nanoseconds after the passage of a particle) that are used by the online event triggering system. In the barrel part, the muon chambers are arranged in three concentric cylindrical layers around the beam axis at approximately $5, 7.5$ and 10 meters radii. In the two endcap regions, muon chambers form large wheels are disposed perpendicularly to the z-axis at distances $|z| = 7.4, 10.8, 14$ and 21.5 meters from the interaction point.

➢ **The Monitored Drift Tubes (MDT)**: measure curves of tracks. The MDTs cover $|\eta| < 2.7$ (except in the innermost endcap layer where they only cover $|\eta| < 2.0$) and allow for precision

momentum measurement. These chambers are composed of three to eight layers of drift tubes, operated at a pressure of 3 bar, which achieve an average resolution of 80 µm per tube leading to a precision of about 35 µm per chamber.

- **The Cathode Strip Chambers (CSC)**: measure with high precision coordinates in the forward part of the detector. In the forward region ($2 < |\eta| < 2.7$), the CSC are used in the innermost tracking layer due to their higher rate capability and time resolution as the particle fluxes and muon-track density are highest in those regions. This multiwire proportional chambers system has cathode planes segmented into strips in orthogonal directions. As a result the spatial resolution is 40 µm in the bending plane and about 5 mm in the transverse plane.
- **The Resistive Plate Chambers (RPC)**: are used for triggering and provide a second coordinate measurement in central region ($|\eta| < 1.05$). They are arranged in concentric cylinders. An RPC is a gaseous parallel electrode-plate detector. Two resistive plates are kept parallel to each other at a distance of 2 mm and an electric field of about 4.9 kV/mm between them is applied. When a muon enters the chamber it ionizes the gas creating an avalanche of charges along its track, charges that are collected at the anode.
- **The Thin Gap Chambers (TGC)**: are used for triggering and providing a second coordinate measurement (in the non-bending direction) in the forward region ($1.05 < |\eta| < 2.4$) to complement the measurement of the MDT chambers. The TGCs are multi-wire proportional chambers.

The performance of the muon spectrometer in terms of transverse momentum resolution varies between 3% to 12%, for p_{T} values between 10 GeV and 1 TeV.

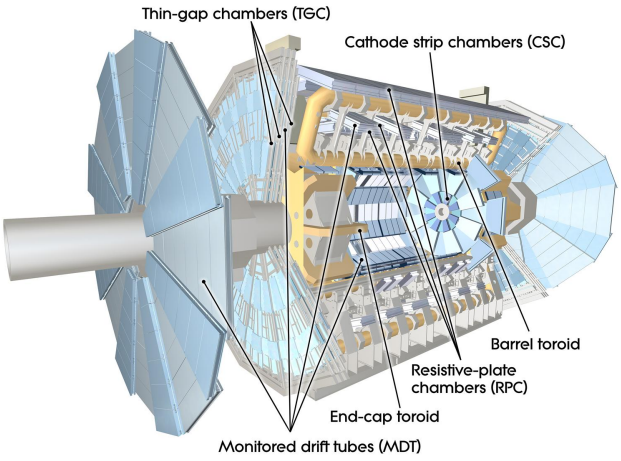


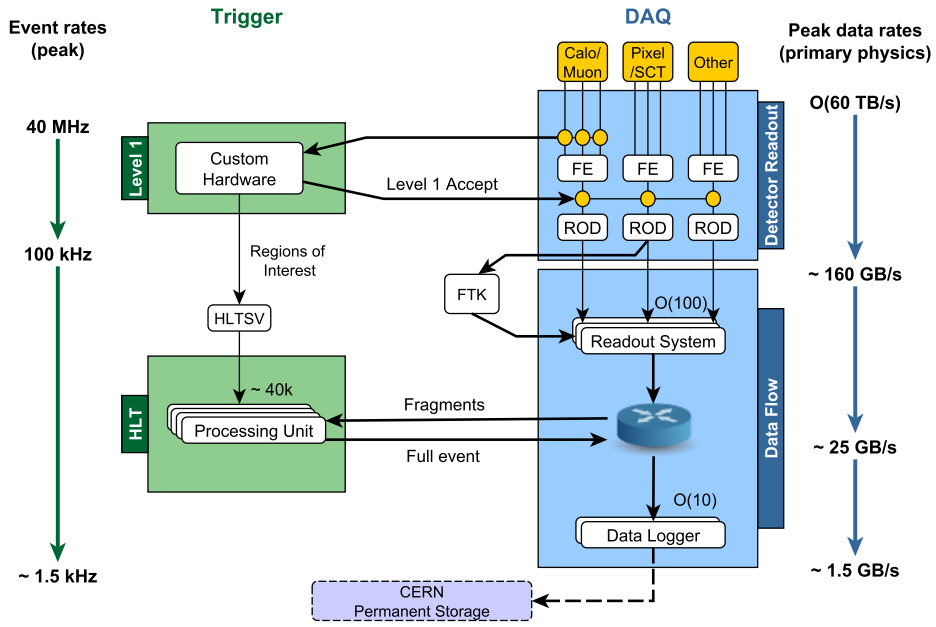
Figure 2.16: Muon spectrometer scheme (taken from Ref. [163]).

2.2.7 Trigger and data acquisition

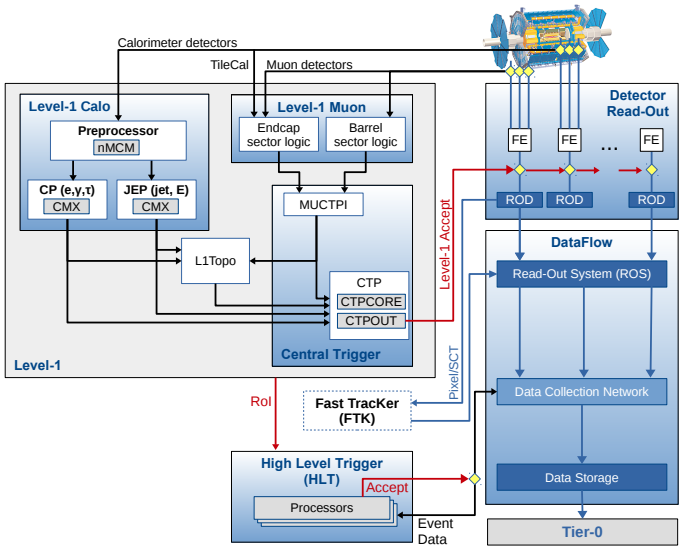
In the ATLAS detector, about 1.7 billions proton-proton collisions are taking place every second [164]. Saving that amount of information would require 60 million megabytes of storage per second which is obviously not possible. Indeed, only a small fraction of the observed collisions are interesting from a Physics point of view as the majority are simply elastic or low energy inelastic collisions. The Trigger and Data Acquisition (TDAQ) system [165, 166], which is shown in Figure 2.17, is taking care of selecting collision events potentially interesting for Physics and save them for study. The TDAQ is triggering recording of events based on inputs coming from the different parts of the detector.

For Run 1, the TDAQ was divided in three stages: the level 1 (L1), the level 2 (L2) triggers and the event filter. For Run 2 and Run 3, due to the increase of the instantaneous luminosity, the TDAQ was changed to only two decision steps: the level 1 (L1) trigger and a high level trigger

(HLT). The L1 trigger, located in the ATLAS detector, is a hardware trigger *i.e.* its decision are based on raw inputs from a subset of information coming from the calorimeters and the muon spectrometer searching for simultaneous characteristic signals. Its decision is performed in less than 2.5 microseconds after an event occurred: the event information during this time are being kept in buffer storage. The information rate is reduced from $40\,\mathrm{MHz}$ to $100\,\mathrm{kHz}$ thanks to the L1 trigger decision *i.e.* only 100 000 events per second are selected by the L1 trigger. If an event is accepted it is then examined by the high level trigger which further reduces the information rate to $1.5\,\mathrm{kHz}$. It corresponds to the amount of events saved for Physics analyses that are performed off-line. The HLT is using about 40 000 CPU cores and takes a decision in just 200 microseconds. It is relying on the full detector information available applying tighter selection criteria with reconstruction algorithms that are very close to off-line algorithms.



(a)



(b)

Figure 2.17: Trigger and data acquisition (TDAQ) system used by ATLAS for Run 2. (a) Diagram showing the expected peak rates and bandwidths through each component while (b) focuses on the components of the Level 1 (L1) trigger system (taken from Ref. [166]).

Chapter 3

Reconstruction and identification of objects in ATLAS

3.1 Vertex and track reconstruction . . . 82
3.2 Electrons . . . 86
3.2.1 Electron reconstruction . . . 86
3.2.2 Electron identification and isolation criteria . . . 89
3.2.3 Electron calibration . . . 91
3.3 Muons . . . 93
3.3.1 Muon reconstruction . . . 93
3.3.2 Muon identification and isolation criteria . . . 94
3.4 Taus . . . 96
3.5 Photons . . . 97
3.6 Jets . . . 97
3.6.1 Reconstruction of jets with the anti-k_t algorithm . . . 98
3.6.2 Small-R jets . . . 99
3.6.2.1 PFlow and EMTopo reconstruction algorithms . . . 100
3.6.2.2 Calibration of small-R jets . . . 100
3.6.3 Large-R jets . . . 106
3.6.3.1 Reconstruction of large-R jets . . . 106
3.6.3.2 Calibration of large-R jets . . . 107
3.6.4 VR track-jets and ghost association . . . 110
3.6.5 Flavour tagging . . . 112
3.6.6 Jet cleaning . . . 116
3.7 Overlap removal . . . 116
3.8 Missing transverse energy . . . 117

After being recorded, events are passed through offline event reconstruction and identification algorithms. The sub-detectors measurements are combined in order to reconstruct, identify and calibrate particles or physics objects such as vertices, tracks, leptons, photons, jets, missing transverse energy and so on. This chapter describes that full process which is performed both for data and event simulations. For instance, as presented by Figure 3.1, reconstructing an electron requires to combine hit measurements from the tracking part of the detector to first be able to reconstruct its track, direction and momenta to then combine those information with the electromagnetic calorimeter measurements to determine its energy.

Let's define what efficiency and rejection refers to as those two notions are used throughout this entire chapter. The efficiency (denoted ε) of a selection or an identification, is simply the ratio of a certain number after and before a selection is applied (*e.g.* ratio of the number of events or number of particles or number of jets or reconstructed energy versus true energy and so on). For example, the b-tagging efficiency (denoted ε_b) of a b-tagging algorithm is the ratio $\varepsilon_b := N_{b\text{-jet}}^{\text{tagged}}/N_{b\text{-jet}}^{\text{total}}$ of the number of true b-jets that were tagged ($N_{b\text{-jet}}^{\text{tagged}}$) and the total number of true b-jets ($N_{b\text{-jet}}^{\text{total}} = N_{b\text{-jet}}^{\text{tagged}} + N_{b\text{-jet}}^{\text{untagged}}$).

The rejection of a selection is simply the ratio $1/\varepsilon$ as for instance the c-jet rejection ($1/\varepsilon_c$) of a b-tagging algorithm. If a b-tagging algorithm has a c-jet rejection of 100, it means that out of a total of 100 true c-jets only one of them will be b-tagged. Hence, for a same b-tagging efficiency, the higher is the c- and light-jet rejection, the better the b-tagging algorithm is as only b-jets are selected while c- and light-jets are rejected.

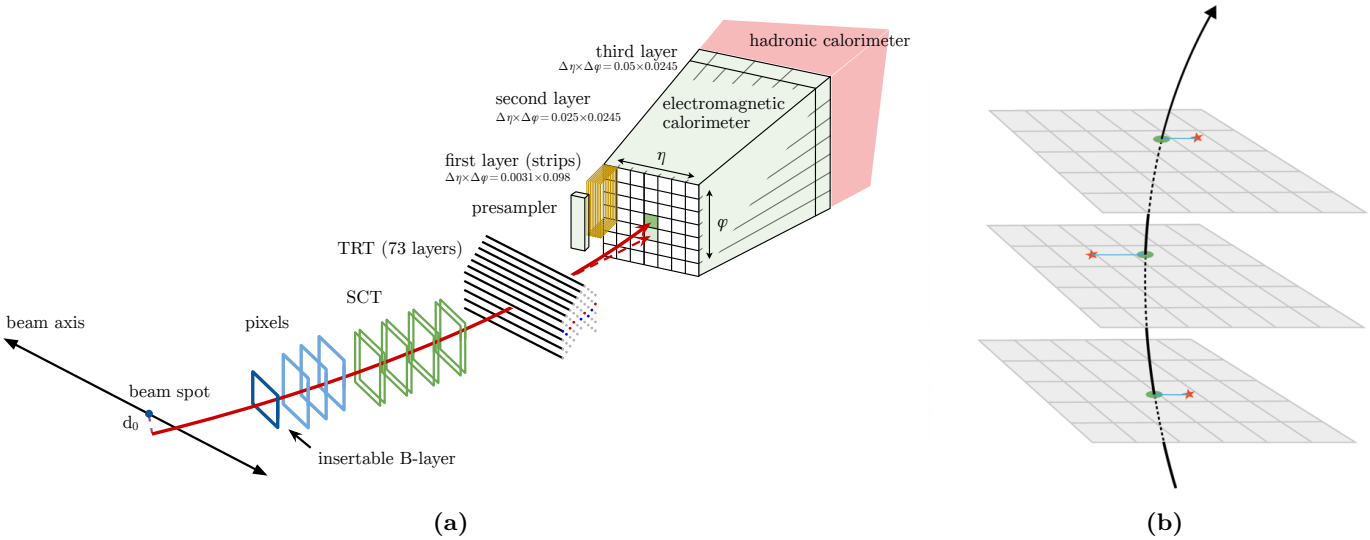


Figure 3.1: (a) Schematization of the path of an electron (red solid line) in the ATLAS detector. The electron first goes through the tracking system (pixel detectors, then silicon-strip detectors and TRT) and then enters the electromagnetic calorimeter where it deposits all its energy. The dashed red trajectory indicates the path of a photon produced by the interaction of the electron with the material in the tracking system. (b) Representation of a charged particle crossing sub-detector planes. The red stars indicate hit measurements in the detector layers. The fitted track is shown with a black arrow. The green dots indicate the intersection of the reconstructed track with each surface. The blue lines represent the track-to-hit distance for each layer (taken from Refs. [150, 167]).

3.1 Vertex and track reconstruction

A primary vertex (PV) is the initial collision point between two partons from a proton-proton collisions. To deal with the high multiplicity of primary vertices for Run 2 due to the high pile-up conditions, ATLAS uses a 3D imaging technique [168] to reconstruct the primary vertices. The 3D imaging algorithm is divided in two steps: vertex finding and vertex fitting. For the vertex finding stage, the algorithm tries to determine simultaneously all PV candidates along the beam axis taking into account all reconstructed tracks using spatial 3D Fourier transformations of the track path length in order to identify high density region which indeed correspond to primary vertices. At the vertex fitting step, tracks are iteratively assigned to the closest vertex candidate. An adaptive vertex fitting algorithm [169] is then used to estimate the position and its uncertainty removing incompatible tracks. The hard-scatter primary vertex is then determined as the one with the highest sum of the transverse momenta squared ($\sum p_{\mathrm{T}}^2$) of all associated tracks. The remaining vertices are from pile-up and secondary vertices. The efficiency to reconstruct and then select the hard-scatter primary vertex has been measured to be 99% in $t\bar{t}$ events for $\mu = 30$ [170].

Reconstruction of charged particle tracks [171, 172] plays a leading role in all physics analyses. For instance, identification of primary (PV) and secondary (SV) vertices are key ingredient to distinguish light- and c-jets from b-jets. As b-hadrons have a relatively long lifetime (typically $\tau \approx 1.5$ ps) [173], their flight distance ($\gamma c\tau$) is of the order of 450 µm up to few millimeters depending on their relativistic factor. As a result b-hadrons decay to other particles creating a

secondary vertex which is an intrinsic characteristic of b-jet as pictured in Figure 3.2b.

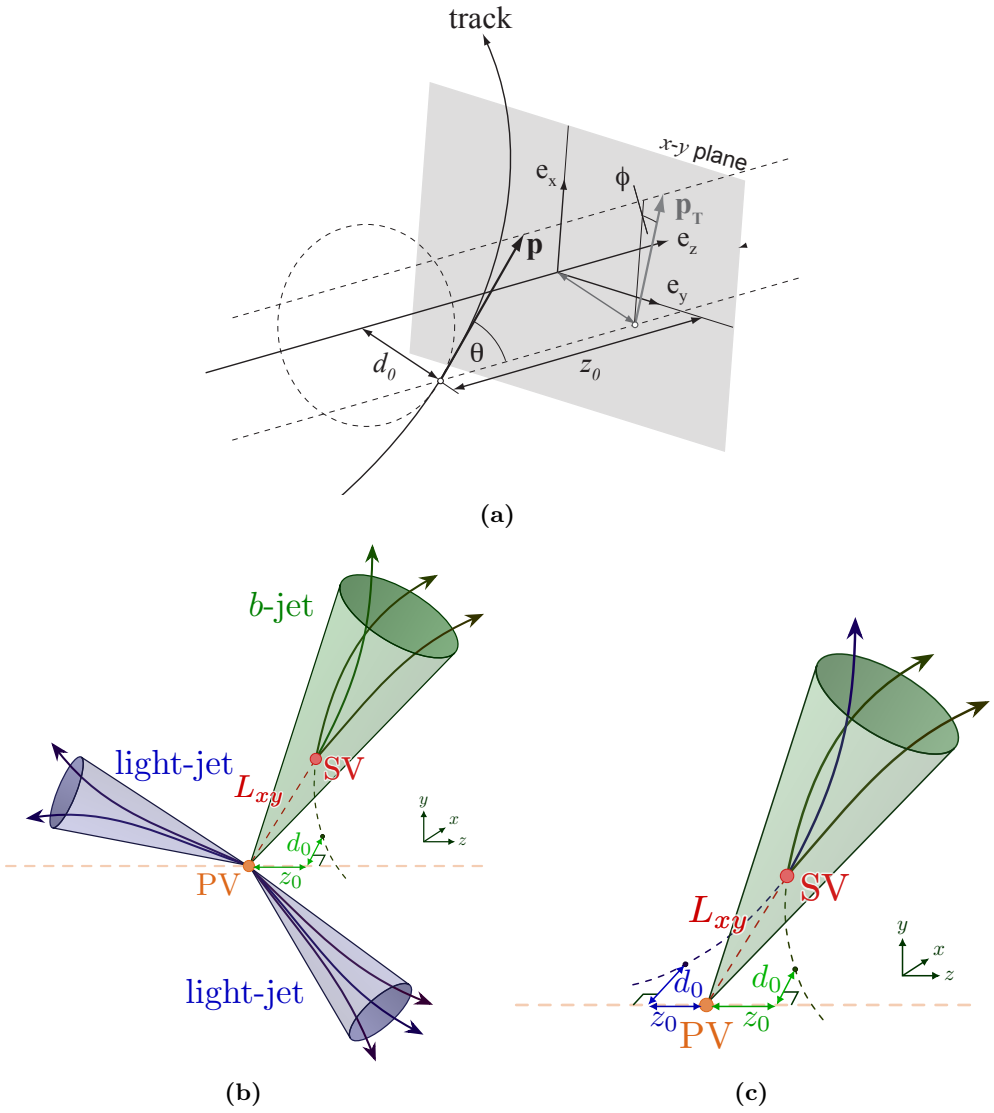


Figure 3.2: (a) Schematization of the track helix parameters used by ATLAS to fully describe tracks. (b) Light- and b-jet tracks scheme. Light-jets are produced at the primary vertex (PV) while b-hadrons are creating a secondary vertex (SV) which is displaced from the primary vertex. Tracks from b-jets are characterized by large longitudinal (z_0) and transverse (d_0) impact parameters. (c) Representation of the longitudinal and transverse impact parameters associated to two different tracks (blue and green) from a b-jet (taken from Refs. [45, 174]).

Charged particles in the ATLAS inner detector have a helicoidal trajectory due to the solenoidal magnetic field. Tracks can therefore be fully described by 5 helical parameters $(d_0, z_0, \theta, \phi, q/p)$ [172, 174], shown in Figure 3.2 and described below. The point of closest transverse approach of the track to the beam axis is called the perigee.

- **The transverse impact parameter** (d_0): is the transverse distance between the perigee and the beam axis.
- **The longitudinal impact parameter** (z_0): is the longitudinal distance between the perigee and the transverse plane (x-y) which is containing the nominal interaction point (or the primary vertex depending on the context).
- **The track azimuthal angle** (ϕ): is the azimuthal angle of the track direction at its perigee.
- **The track polar angle** (θ): is the polar angle of the track direction at its perigee.
- **The charge-momentum ratio** (q/p): defines the curvature of the track due to the 2 T solenoidal

magnetic field.

ATLAS uses an *inside-out* method to reconstruct tracks *i.e.* first the inner sub-detectors (pixel detector and SCT) information is used before the outer TRT information is taken into account. The process begins with assembling clusters from the raw measurements of the inner detector. In average the clusters size in the barrel is about two pixels in the r-ϕ plane and varies from one to three pixels in the longitudinal direction increasing with η. Thanks to those formed clusters, three-dimensional measurements are performed to reconstruct the point, referred as *space-point*, where the charged particle traversed the active material of the inner detector.

Once space-points are reconstructed, a track-seeding is performed which consists in forming triplets of space-points (referred to as a seed) either from the pixel or SCT subdetectors which are compatible with originating from a charged particle track. Some loose selections are applied in order to avoid seeds that would lead to low quality tracks. To reduce combinatorics, *search roads* (sets of detector modules that can be expected to contain clusters compatible with the seed) are built through the remaining detector thanks to the estimated seed trajectory.

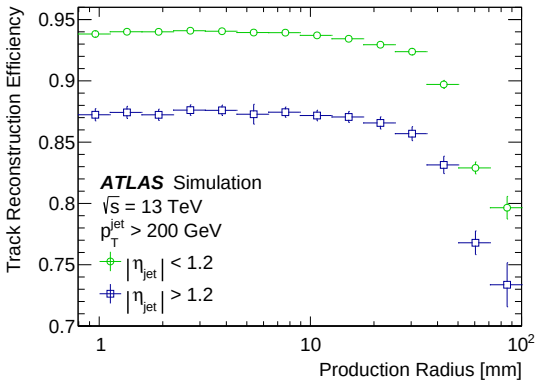
The track candidates are then formed extending radially the seeds thanks to the search roads with the Kalman filter [175] which searches both outwards and inwards adjacent clusters while trying to smooth the trajectory. An ambiguity resolution step is performed to resolve overlaps between track candidates and reject incorrect combinations of unrelated clusters (called "fake tracks"). Tracks candidate are ranked based on a track score which takes into account criteria such as the number of clusters assigned to the track, the track fit χ^2, the transverse momentum (to promote energetic tracks and suppress low momentum tracks that often have incorrectly assigned clusters). Shared hits are often an indicator of fakes or low quality tracks. In order to suppress fake tracks and to ensure a good spatial and momentum resolution, a minimum of 7 hit clusters (out of the 12 expected) summing the ones from the silicon tracker and SCT is required with as well no more than two shared clusters for a track. Finally, in order to benefit from additional measurements on the track, in particular for momentum resolution and particle identification, an extension to the TRT is attempted using again a road search and Kalman filter starting from the position estimation of the track candidate in the TRT volume. The TRT hits are then added to the tracks and the whole tracks are fitted again with the global χ^2 fitter in case the extension is successful. To be noted that the silicon-only tracks (*i.e.* only built with pixel detector and SCT information) are kept if the fit worsens with the addition of the TRT hits. The track extension to TRT can also be rejected based on TRT track quality criteria.

It is worth mentioning that potential seeds failing the track-finding requirements are checked for calorimeter compatibility. For a seed within a region of interest (ROI) in the calorimeter, the track procedure is again performed allowing for an additional "kink" in the track using the Gaussian Sum filter [176] instead of the Kalman filter. This increases electron efficiency reconstruction as bremsstrahlung radiations can cause non-negligible change in track directions: this procedure is referred to as *bremsstrahlung recovery*.

The inside-out procedure described above is efficient to reconstruct tracks from particles produced by the primary pp interactions. However, particles produced at greater distance from the beam-pipe such as electrons originating from photon conversions ($\gamma \rightarrow e^+e^-$) in the inner detector could fail the track reconstruction algorithm. To circumvent this issue an *outide-in* procedure is used *i.e.* using first TRT information before the pixel and SCT's ones. It uses the non already assigned to track hits from the inside-out algorithm. The track reconstruction is only performed in ROIs determined by the electromagnetic calorimeter and compatibility with the TRT hit segments is performed. The compatible segments are checked to be matched with short silicon track seeds formed of two space-points from the pixel or SCT subdetectors only taking into account hits close-by to the TRT. Finally, as previously, the seeds found are extended into tracks using the same protocol (road search, Kalman filter, ambiguity removal and global χ^2 fitter). The track candidates

are then again extended (back) into the TRT using the collected hit segments.

The Figure 3.3 shows the track reconstruction efficiency for charged particles inside jets in dijet simulated events.



s

Figure 3.3: Track reconstruction efficiency for charged primary particles in jets with $|\eta| < 1.2$ ($|\eta| > 1.2$) as a function of the jet transverse momentum in simulated dijet MC events (taken from Ref. [171]).

The Figure 3.4 shows the improvement of the transverse and longitudinal impact parameter resolution in data between Run 1 and Run 2 thanks to the insertion of the Insertable B-Layer (IBL) in the inner detector for the Run 2.

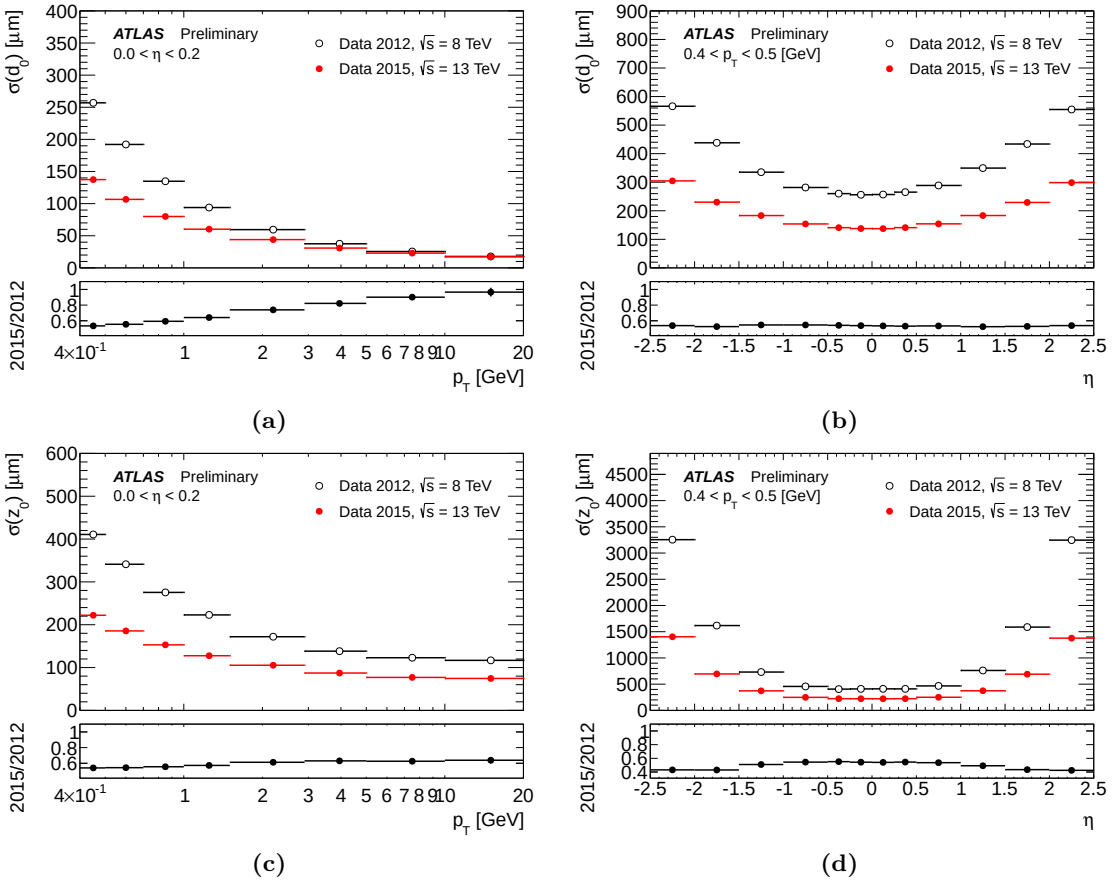


Figure 3.4: Comparison of the spatial resolution (σ) for the transverse (d_0) and longitudinal (z_0) impact parameters for Run 1 (2012, $\sqrt{s} = 8$ TeV) and Run 2 (2015, $\sqrt{s} = 13$ TeV) data as a function of (a), (c) p_{T} for $0 < \eta < 0.2$ or (b), (d) η for $0.4 < p_{\mathrm{T}} < 0.5$ GeV. The improvement observed is thanks to the addition of the Insertable B-Layer (IBL) in the inner detector for the Run 2 (taken from Ref. [177]).

3.2 Electrons

Electrons going through the ATLAS detector are reconstructed thanks to their track left in the inner detector and their (full) energy deposition in the electromagnetic (EM) calorimeter [178–180] as shown in Figure 3.1. The first step consists in reconstructing the electromagnetic shower of the electron candidate. Then information from the calorimeter and the inner detector are matched in the $\eta \times \phi$ phase space to enable the identification of electrons. Finally, a calibration of the electron is performed.

3.2.1 Electron reconstruction

Before 2017, electromagnetic showers were reconstructed with a fixed-size clustering algorithm called *sliding-window* algorithm [181] *i.e.* the formed clusters had a fixed size. Since then, ATLAS is now using a *dynamical topological cell clustering* (abbreviated topo-cluster) algorithm [178–180, 182] meaning that clusters have a variable size. The dynamic topo-cluster algorithm enables to recover energy from bremsstrahlung photons or from electrons from photon conversions which is the main drawback of the sliding window algorithm. Moreover, it is achieving similar energy response linearity and stability with respect to pile-up as the sliding window algorithm.

The topo-cluster algorithm works as follows:

It begins by iteratively forming *proto-clusters* from the electromagnetic and hadronic calorimeters using a set of noise thresholds in which the cell initiating the cluster is required to have a significance $\left|\zeta_{\text{cell}}^{\text{EM}}\right| > 4$ with:

$$\zeta_{\text{cell}}^{\text{EM}} = \frac{E_{\text{cell}}^{\text{EM}}}{\sigma_{\text{noise,cell}}^{\text{EM}}}, \tag{3.2.1}$$

where $E_{\text{cell}}^{\text{EM}}$ is the cell energy at the EM scale *i.e.* the correct energy deposited by the shower in the calorimeter cell after having subtracted expected noise contributions (electronic and pile-up noise). As a result $E_{\text{cell}}^{\text{EM}}$ can be negative due to noise subtraction. And $\sigma_{\text{noise,cell}}^{\text{EM}}$ is the expected cell noise uncertainty taking into account all known noises which are caused by electronics and pile-up (estimated using the average expected instantaneous luminosity for Run 2). All the neighbour cells with $\left|\zeta_{\text{cell}}^{\text{EM}}\right| > 2$ are added to the proto-cluster and each of them becomes a seed cell repeating the procedure. If two proto-clusters are sharing a common cell they are then merged. This procedure is robust to energy bias as it uses $|\zeta_{\text{cell}}|$ instead of ζ_{cell}, so negative energy cells are also taken into account hence proto-cluster energy is not overestimated. Finally, a "crown" of the nearest neighbour cells is added to the proto-clusters independently of their significance. In case there are several local maxima, the proto-cluster is split into separate clusters, a local maxima being a cell with $E_{\text{cell}}^{\text{EM}} > 500$ MeV, at least four neighbours and none of the neighbours has a larger signal. Cells from the pre-sampler and first layer of the LAr EM calorimeter are not considered at that stage to avoid formation of noise clusters. Afterwards, the proto-cluster procedure is extended to the whole calorimeter. In order to pre-select electron candidate clusters, the EM fraction (f_{EM}) can be defined as:

$$f_{\text{EM}} := \frac{E_{L1} + E_{L2} + E_{L3} + w(E_{E4} + E_{\text{PS}})}{E_{\text{clus}}}, \quad w = \begin{cases} 1 & \text{if } 1.37 < |\eta| < 1.63 \\ 0 & \text{otherwise} \end{cases} \tag{3.2.2}$$

where E_{Lx} is the cluster energy in layer x of the calorimeter and the term $E_{E4} + E_{\text{PS}}$ is only considered for clusters in the transition region $1.37 < |\eta| < 1.63$, since the energy deposition of electrons is non-negligible in the pre-sampler and scintillator layers. And E_{clus} (referred to as the EM energy of the cluster) is the weighted sum of the cell energy constituting the cluster, the weight being determined by the geometric structure of the cluster.

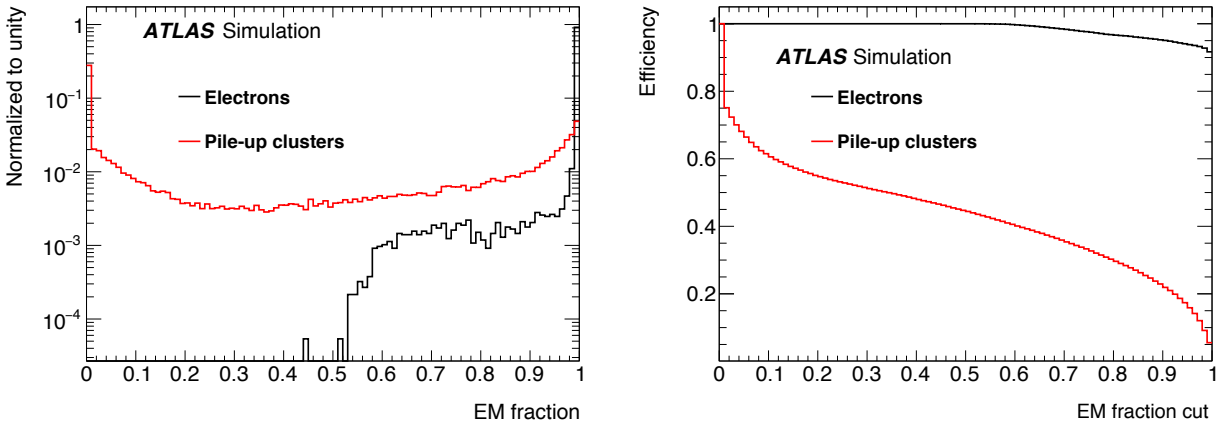
The f_{EM} represents the fraction of energy left by a particle in the electromagnetic calorimeter. Since electrons deposit almost all their energy in the electromagnetic calorimeter one would expect f_{EM} to be close to 1. As a result, electron candidate clusters are required to have $f_{\text{EM}} > 0.5$ because this cut enables to reject around 60% of the pile-up clusters without affecting the efficiency for true electron topo-clusters selections as demonstrated in Figure 3.5. Moreover, only clusters with EM energy greater than $400\ \text{MeV}$ are considered.

All physical properties of the proto-cluster, its direction, momentum, energy, center of gravity etc are determined using weighted sums taking into account the energy of each cell/cluster and their geometrical weight inside the proto-cluster [182].

To reconstruct electron tracks, the EM clusters passing loose shower shape requirements in terms of energy leakage in the hadronic calorimeter downstream of the electromagnetic calorimeter and energy distribution in η are used to define the region of interest (ROI). Within the ROI, the usual track reconstruction is performed. The tracks are then matched to the EM clusters by means of tight requirements in η and ϕ. In case several tracks are found they are ranked in the following order: tracks with hits in the pixel detector, then tracks with hits in the SCT but not in the pixel detector. Tracks with the better ΔR with respect to the cluster in the second layer of the calorimeter are preferred. Other more refine selections are performed in some particular cases.

Then *superclusters* are reconstructed in two stages: EM topo-clusters are tested as seed cluster candidates and EM topo-clusters near the seed candidates are identified as *satellite*. Satellites can be the result of bremsstrahlung effect: if they pass some requirement they are added to the supercluster. To build the superclusters, the topo-clusters are sorted by decreasing transverse energy $E_{\text{T}} := \sqrt{m^2 + p_{\text{T}}^2} \approx E/\cosh(\eta)$. Each cluster starting from the one with the highest E_{T} to the lowest one is considered as an electron seed candidate, they are required to have $E_{\text{T}} > 1\ \text{GeV}$, be matched to a track with at least four hits in the silicon tracking subdetectors. If a cluster satisfies those requirements, a search for satellite clusters is performed within a window $\Delta\eta \times \Delta\phi = 0.075 \times 0.125$ around the cluster barycenter. An extended search is also performed for $\Delta\eta \times \Delta\phi = 0.125 \times 0.300$ to catch bremsstrahlung electrons this time requiring that the best matched track is shared between the seed cluster and satellite cluster candidate. The search window is larger in ϕ than in η because trajectories of charged particles are bent the in the R-ϕ plane resulting in a larger shower spread in that direction. In that iterative process, clusters that are found to be satellite clusters of other clusters are not considered as electron seeds.

A calibration of the supercluster is then performed matching tracks in the same way as for EM topo-clusters but with the supercluster: the corresponding electron object for analysis is created. The Figure 3.6 present the reconstruction efficiencies for the different steps of an electron reconstruction as a function of the transverse energy.



(a) **(b)**

Figure 3.5: (a) Distribution of the EM fraction (f_{EM}) and (b) reconstruction efficiency as a function of the f_{EM} selection cut for simulated true electron (black) and pile-up (red) clusters (taken from Ref. [178]).

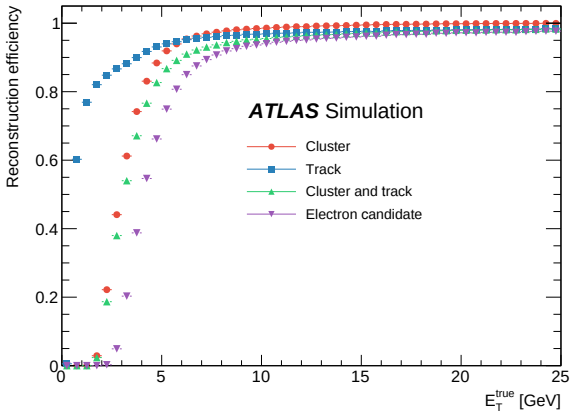
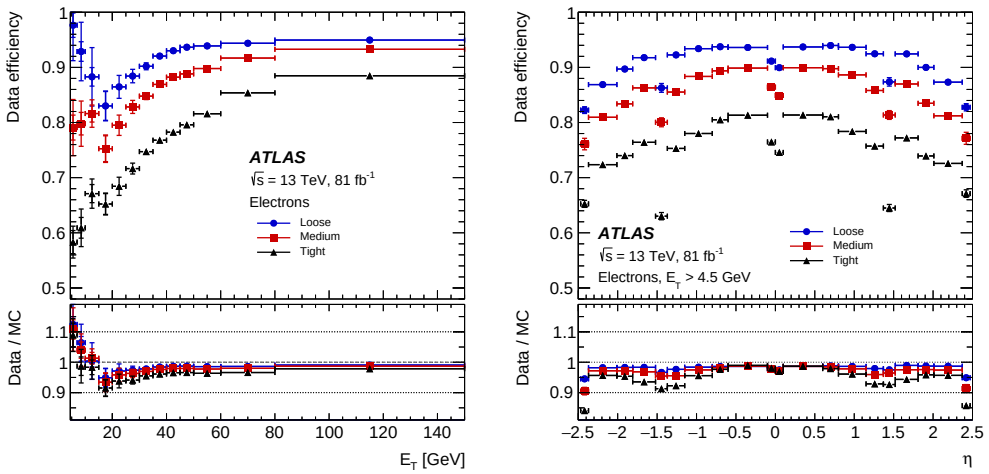


Figure 3.6: Reconstruction efficiencies for cluster, track, cluster and track, and electron as a function of the generated electron E_{T} (taken from Ref. [178]).

3.2.2 Electron identification and isolation criteria

Once reconstructed, quality criteria for the electron can be defined to improve the purity of the selected electrons [178–180] and reject fake electrons. The discrimination variables used to determine electron quality can be grouped into several categories: lateral, longitudinal shower development profile, track of the electron and compatibility of the track-cluster association. Here are a non-exhaustive list of discriminating variables used: the transverse impact parameter (d_0) of the track, its significance ($|d_0/\sigma(d_0)|$), its number of hits, the lateral shower energy deposition variables in the η and ϕ directions, lateral width, longitudinal shower shape variable, the energy leakage in the hadronic calorimeter, the ratio E/p.

In ATLAS, 3 electron quality criteria (also called working point) are defined *LHTight* (also denoted *TightLH*), *LHMedium* (or *MediumLH*) and *LHLoose* (or *LooseLH*) with LH standing for likelihood. The likelihood is built using the probability density functions from the discriminating variables (typically 13 variables) and is divided into 9 bins for η and 12 bins for E_{T}. The likelihood provides a unique output discriminant denoted d_L. The working point are corresponding to three decreasing fixed values of the LH discriminant d_L resulting in identification efficiencies of 80%, 88% and 93% for electrons with $E_{\mathrm{T}} = 40$ GeV respectively for the LHTight, LHMedium and LHLoose working points (WP) as shown in Figure 3.7.



(a) (b)

Figure 3.7: Electron identification efficiency in $Z \to e^+e^-$ events in data as a function of (a) E_{T} and (b) η for the Loose, Medium and Tight operating points. For both plots, the bottom panel shows the data-to-simulation ratios (taken from Ref. [178]).

One can distinguish *prompt* electrons *i.e.* coming from the main collision from a *non-prompt* electrons which are non isolated electrons because they are emitted for instance in the semi-leptonic decay of heavy-hadrons (electrons emitted in jets). Non-prompt electrons are not rejected by quality criteria as they are real electrons. The isolation can basically be quantified with the energy deposits in the calorimeter around the electron (*calorimeter isolation*) or from the tracks of nearby charged particles (*track isolation*) or combining the calorimeter and track isolation criteria.

ATLAS provides different working points for the calorimeter and track isolation that are presented in Table 3.1. The working points can be defined in two ways, either targeting a fixed value of efficiency (*Gradient* WP) or with fixed cuts on the isolation variables (*HighPtCaloOnly, Loose and Tight* WPs). The Figure 3.8 shows the efficiency of the different isolation WPs. The isolation criteria needed are specific to each physics analysis in particular it depends on the topology of the events targeted. The choice of the isolation is hence a compromise between an efficient identification of prompt isolated electrons or produced in a busy environment, and a good rejection of

electrons from heavy-flavour decays or light hadrons misidentified as electrons.

For calorimeter isolation, the energy around the electromagnetic shower ($E_{\mathrm{T}}^{\mathrm{coneXX}}$) is determined by using a cone $\Delta R = XX/100$ around the electron cluster barycenter. For example if $XX = 20$ then $\Delta R = 0.2$ and $E_{\mathrm{T}}^{\mathrm{cone20}}$ is measured. $E_{\mathrm{T}}^{\mathrm{coneXX}}$ is computed by subtracting the sum of the energy of the core EM shower, the transverse energy leakage outside of the core and contribution from pile-up to the total energy contained in the cone [178].

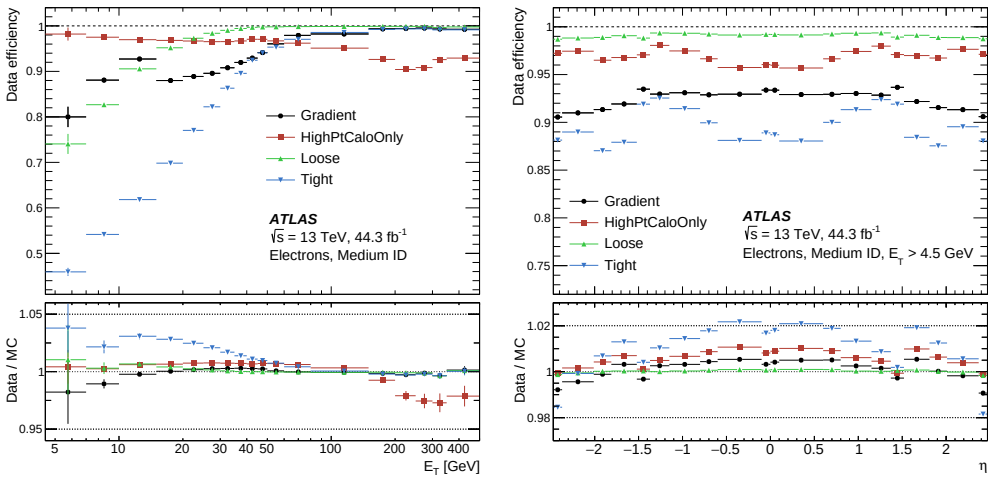
For track isolation, the transverse momentum ($p_{\mathrm{T}}^{\mathrm{varconeXX}}$) is obtained by summing the transverse momentum of selected tracks within the cone centered around the electron track excluding the tracks matched to the electron. The cone in that case is defined with a variable radius ΔR, defined in Equation (3.2.3), which shrinks for large transverse momenta electron. The tracks considered are required to have $p_{\mathrm{T}} > 1$ GeV, $|\eta| < 2.5$, at least 7 silicon hits (Pixel+SCT subdetectors) and a loose vertex association they should have either been used in the primary vertex fit otherwise they should satisfy $|z_0| \sin(\theta) < 3$ mm with z_0 the longitudinal impact parameter distance to the primary vertex.

$$\Delta R = \min\left(\frac{10}{p_{\mathrm{T}}[\mathrm{GeV}]}, \Delta R_{\mathrm{max}}\right), \tag{3.2.3}$$

with ΔR_{max} is the maximum cone size taken to be $\Delta R_{\mathrm{max}} = 0.2$. The track isolation is taking advantages of the much smaller tracker granularity which allows for narrower cone sizes: for calorimeter isolation it would not be possible to build cones with a radius much smaller than $\Delta R = 0.2$ due to the finite granularity of the calorimeter.

Working point	Calorimeter isolation	Track isolation
Gradient	$\epsilon = 0.1143 \times p_{\mathrm{T}} + 92.14\%$ (with $E_{\mathrm{T}}^{\mathrm{cone20}}$)	$\epsilon = 0.1143 \times p_{\mathrm{T}} + 92.14\%$ (with $p_{\mathrm{T}}^{\mathrm{varcone20}}$)
HighPtCaloOnly	$E_{\mathrm{T}}^{\mathrm{cone20}} < \max(0.015 \times p_{\mathrm{T}}, 3.5\ \mathrm{GeV})$	-
Loose	$E_{\mathrm{T}}^{\mathrm{cone20}}/p_{\mathrm{T}} < 0.20$	$p_{\mathrm{T}}^{\mathrm{varcone20}}/p_{\mathrm{T}} < 0.15$
Tight	$E_{\mathrm{T}}^{\mathrm{cone20}}/p_{\mathrm{T}} < 0.06$	$p_{\mathrm{T}}^{\mathrm{varcone20}}/p_{\mathrm{T}} < 0.06$

Table 3.1: Electron isolation working points and isolation efficiency ε definitions. In the Gradient working point definition, the unit of p_{T} is GeV. For all working points, a cone size of $\Delta R = 0.2$ for calorimeter and $\Delta R_{\mathrm{max}} = 0.2$ for track isolations has been used The HighPtCaloOnly WP is only based on calorimeter isolation while the other WPs require both calorimeter and tracks isolation criteria. (taken from Ref. [178]).



(a) **(b)**

Figure 3.8: Efficiency of the different isolation working points for electrons passing the Medium identification criterion in inclusive $Z \to e^+e^-$ events as a function (a) E_{T} and (b) η. The electrons are required to pass the LHMedium identification criterion. The bottom panel shows the ratio of the efficiencies measured in data and in MC simulations (taken from Ref. [178]).

3.2.3 Electron calibration

Once electrons are reconstructed, a multivariate analysis (MVA) regression algorithm is used [178] to calibrate them both in data and simulations. A regression algorithm is a machine learning algorithm taking input variables to predict one or several output variables such as for instance corrections for the energy and transverse momentum of a particle. Regression algorithms are hence a powerful tool to reduce biases, improve resolution and so on: they can be used to calibrate particles such as electrons.

The MVA regression algorithm is trained in simulated samples containing single electrons without pile-up. Phase space is divided in $(|\eta|, E_{\mathrm{T}})$ regions. In that case the MVA should estimate the energy of the electrons based on the energy deposits in the calorimeter. The set of input variable provided is: the energy deposited in the calorimeter, the energy deposited in the pre-sampler, the ratio of the energies deposited in the first and second layers (E_{L_1}/E_{L_2}) of the EM calorimeter, the η of the impact point of the shower in the calorimeter, and the angular distances in η and in ϕ between the impact point of the shower and the center of the closest cell in the second calorimeter layer. In the transition region between the barrel and endcap calorimeters, $1.4 < |\eta| < 1.6$, the amount of material before the first active layer of the calorimeter is large which degrades electron energy resolution. To mitigate this effect, the information from the scintillators (layer 4) are used: the MVA regression algorithm. In this region the energy (E_{L4}) deposited in the cells of the scintillators and the difference in position between the cluster position and the center of the scintillator cells are added to the set of input variables.

An energy correction is then only applied in data to correct for potential miscalibrations of the detector and imperfections of the energy reconstruction procedure. It is determined by means of a template method using $Z \to e^+e^-$ events by comparing the invariant mass distribution of the e^+e^- system in data and simulation to derive the energy correction [178–180]. The same method is fully described in Chapter 4 but was applied to determine an energy correction for b-jets.

In practice the energy scale correction is applied to the data, and the resolution correction is applied to the simulation:

$$E^{\text{data corrected}} = E^{\text{data}}/(1+\alpha), \quad \left(\frac{\sigma_E}{E}\right)^{\text{MC corrected}} = \left(\frac{\sigma_E}{E}\right)^{\text{MC}} \oplus c, \tag{3.2.4}$$

where α is the energy scale factor and c the constant resolution factor. Usually those corrections are derived in η-bins corresponding to the different part of the ATLAS calorimeter. The energy response of the calorimeter is expected to vary with η as different technologies are used. The energy scale factor and constant resolution factor are also measured per year because variations of the LAr temperature and the increase of the instantaneous luminosity are affecting the corrections to be applied. The Figure 3.9 shows the measurement for the α and c corrections as well as the invariant mass distribution comparison between data and simulation after having applied their respective correction.

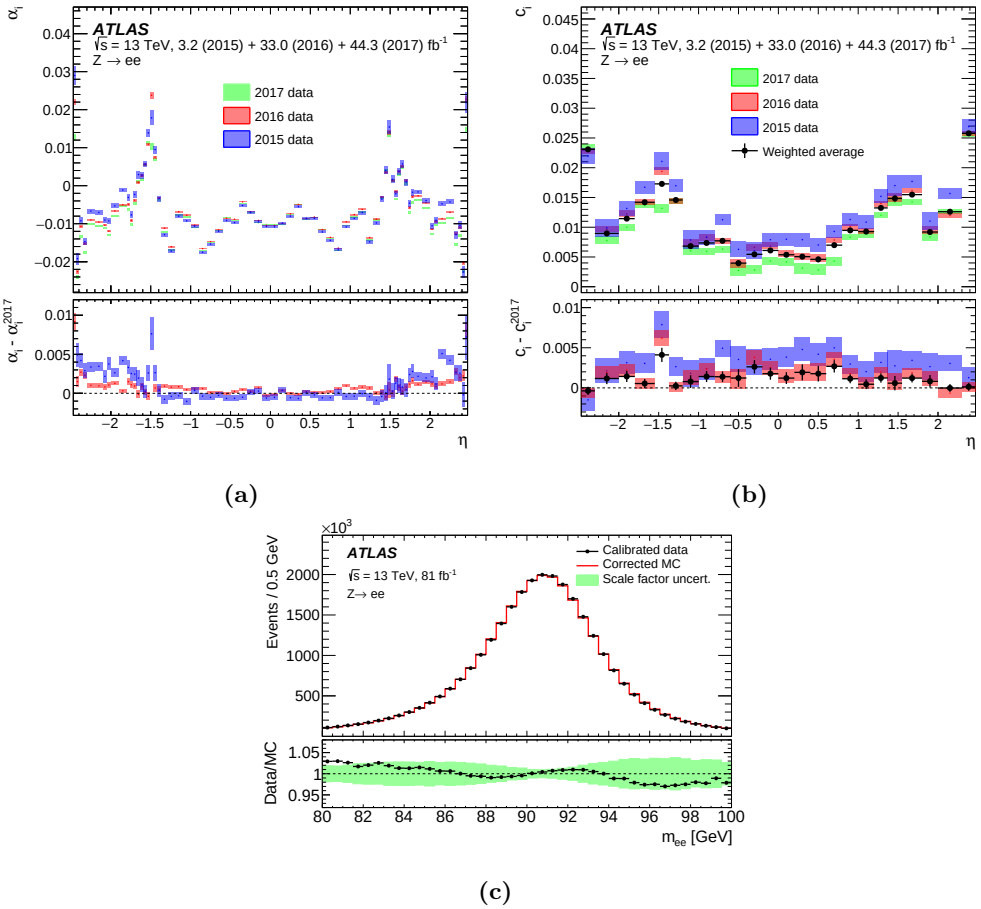


Figure 3.9: (a) Energy scale factors α and (b) additional constant term c as a function of η. The shaded areas correspond to the statistical uncertainties. For (a) and (b), the bottom panels show the differences between measured values in a given data-taking period and the measurements using 2017 data. (c) Comparison between data and simulation of the m_{ee} invariant mass distribution for the selected $Z \to e^+e^-$ candidates after the calibration and resolution corrections are respectively applied to the data and the simulation. The total number of events in the simulation is normalized to the data. The uncertainty band of the bottom plot represents the impact of the uncertainties in the calibration and resolution correction factors (taken from Ref. [178]).

3.3 Muons

Muon candidates are reconstructed using information from the inner detector (ID) and the muon spectrometer (MS) [183, 184]. The first step consists in reconstructing tracks of the muon candidate in the muon spectrometer. Then information from the MS and the inner detector are matched together to enable the identification of muons.

3.3.1 Muon reconstruction

The muon spectrometer measures deflection of muons trajectories in the r-z plane due to the toroidal magnetic field. Since in the r-ϕ plane, muon trajectories are straight lines, the first step of the track reconstruction in the MS consists in identification of short straight line track segments from hits in one of the individual MS stations *i.e.* parts of the sub-detectors. This identification is performed with a Hough transform [185] which enables to recognize patterns of points in an image. The segments are then combined into track candidates requiring a loose pointing constraint based on the IP and assuming perfect parabolic trajectories in the bending plane which is a first order approximation. Information from the muon track detector and muon trigger detectors are combined to obtain a 3D track candidate. A global χ^2 fit of the muon trajectory is performed through the whole muon spectrometer taking into account possible interactions of the muon with the detectors.

Then using the muon trajectory from the global fit, outliers hits are removed and hits along the trajectory that were not assigned to track are included. The fit is then performed again with the updated track information. In case tracks share a large amount of hits, the ambiguity is resolved by keeping only the higher quality tracks. The only exception is for identical tracks in two stations but not sharing hits in a third station in order to increase the reconstruction efficiency of boosted low-mass di-muon systems. Finally, a third fit is performed with a loose constraint on the interaction point additionally taking into account the energy loss related to the calorimeters. The p_{T} of the muon candidate at the interaction point is then extrapolated from that fit.

Finally, the muon reconstruction based on the full detector information is performed. There are five main strategies used to reconstruct muons leading to five "types" of muon described below.

- **Combined (CB) muons**: are reconstructed by matching MS and ID tracks performing a combined track fit based on hits in those two subdetectors taking into account energy loss in the calorimeters. The track in the MS may be again updated during the fit and afterwards the whole track fit can be repeated. For $|\eta| > 2.5$ the MS tracks can be combined with short tracks segments from the pixel and SCT detectors those muons are called silicon associated forward (SiF) muons.
- **Inside-out (IO) combined muons**: a complementary inside-out algorithm is used to extrapolate ID tracks to the MS looking for at least three loosely-aligned hits in the MS. The ID tracks, energy loss in the calorimeters and MS hits are then combined in track fit. The MS reconstructed tracks are not used in that reconstruction which enables to gain in efficiency in regions where the MS coverage is limited or for low-p_{T} muons that may not reach the middle MS station.
- **Muon spectrometer extrapolated (ME) muons**: in case a muon MS track cannot be matched with an ID track, then its parameters are extrapolated to the beamline and used to define a ME muon. It enables to extend the muon reconstruction to the full coverage $|\eta| < 2.7$ of the MS outside of the ID acceptance.
- **Segment-tagged (ST) muons**: are obtained by matching an extrapolated ID track to the MS with a tight angular requirement for at least one segment reconstructed in the MS. The ID track is then considered as muon candidate and muon parameters are computed from the ID track fit.
- **Calorimeter-tagged (CT) muons**: are obtained by extrapolating tracks through the calorime-

ters to search for energy deposits consistent with a MIP. The deposits are then used to tag the ID track as a muon: parameters of the muon are taken from the ID track. Only tracks with a $p_{\mathrm{T}} > 5\,\mathrm{GeV}$ are considered to avoid large background contamination whereas for other type of muon ID tracks with $p_{\mathrm{T}} > 2\,\mathrm{GeV}$ are taken into account.

3.3.2 Muon identification and isolation criteria

Once reconstructed, high quality muon candidates should be identified. As for electron the identification is analysis specific and depend on the targeted topology of the events [183, 184].

ATLAS provides several working points to identify muons and distinguish prompt from non-prompt muons. A distinction is made between non-prompt muons from semi-leptonic decay of light-hadrons and those from heavy flavour hadrons decays. The identification working points are targeting rejection of muon from the light hadron in-flight decays whose track quality is in general lower due to the change in trajectory arising from in-flight decay in the detector. Those muons have often a "kink" in their reconstructed track due to the momentum carried away by the undetected neutrino. As a consequence, it is expected that the fit quality of the resulting combined track will be poor and that the momentum measured in the ID and MS may not be compatible. Bottom and charm hadrons on the other hand are producing good-quality muons: they can be distinguished from prompt-muons thanks to isolation requirement and applying a distance to primary vertex requirement since prompt muons are expected to originate from the PV.

The *Loose*, *Medium* and *Tight* WPs are the 3 identification working points defined by ATLAS for increasing tighter selections. The Loose WP has been optimized for the reconstruction of the decay of the Higgs boson into 4 muons ($H \to ZZ^* \to 4\mu$). The Loose WP has a large signal efficiency which is appropriate for the study of the Higgs in that channel due to the high signal to background ratio. The Medium WP is meant to be used by a large range of analyses, it comes with a good efficiency and rejection of background. The Tight WP achieves the highest purity despite a few percent prompt muon efficiency loss. It brings a significant improvement of the background rejection and is hence meant to be used by analyses that are limited by non-prompt muon background. Finally, two additional WPs corresponding to extreme phase space regions are introduced: the *High-*p_{T} WP which is optimal for muon with $p_{\mathrm{T}} > 100\,\mathrm{GeV}$ and the *Low-*p_{T} WP targeting the lowest p_{T}-muons (split in two method: a cut-based one and a multivariate one). The Table 3.2 and Figure 3.10 summarize the efficiency and misidentification efficiencies for prompt and non-prompt muons for the different working point.

The identification are based on requirements on the number of hits in the MS stations and ID sub-detectors. The q/p compatibility ($\sigma^{\text{comp.}}(q/p)$) for CB and IO muons defined as:

$$\sigma^{\text{comp.}}(q/p) := \frac{|q/p_{\mathrm{ID}} - q/p_{\mathrm{MS}}|}{\sqrt{\sigma^2(q/p_{\mathrm{ID}}) + \sigma^2(q/p_{\mathrm{MS}})}} \tag{3.3.1}$$

is also a discriminating criterion. Where q/p_{ID} and q/p_{MS} are the measurement of the ratio charge-momentum (q/p) in the ID and MS expressed at the interaction point. $\sigma(q/p_{\mathrm{ID}})$ and $\sigma(q/p_{\mathrm{MS}})$ are the associated uncertainties on those measurements. For CB and IO muons the absolute difference between the ID and MS p_{T} measurements (ρ') is also a relevant criterion:

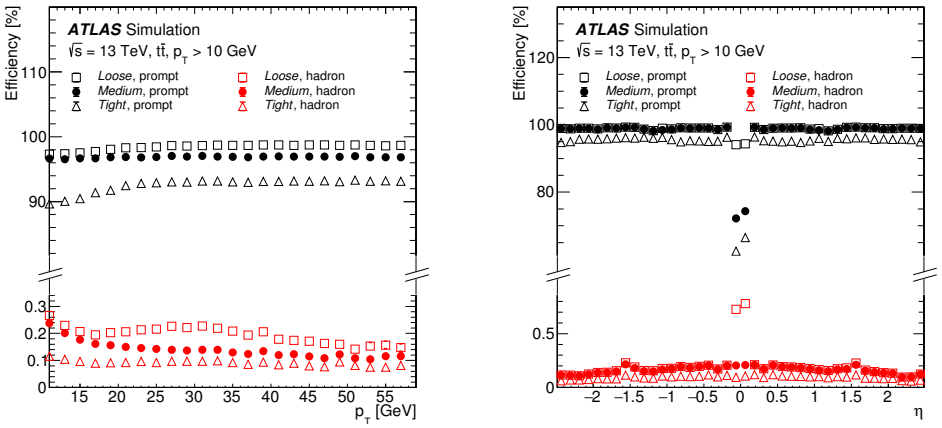
$$\rho' := \frac{\left|p_{\mathrm{T,ID}} - p_{\mathrm{T,MS}}\right|}{p_{\mathrm{T,CB}}}, \tag{3.3.2}$$

with $p_{\mathrm{T,ID}}$, $p_{\mathrm{T,MS}}$ and $p_{\mathrm{T,CB}}$ that are respectively the muon-p_{T} measured in the ID, MS and obtained from the combined track fit. For muons without an ID or MS track $\sigma^{\text{comp.}}(q/p)$ and ρ' are not considered as requirements. The Tight and Medium WP accepts only CB and IO muons. The details of the identification requirements for the different WPs can be found in Reference [183].

The vertex association of the muon candidate is performed imposing the transverse impact parameter significance to satisfy $|d_0|/\sigma(d_0) < 3$. The muon track is checked to be compatible with the reconstructed primary vertex with the $|z_0|\sin(\theta) < 0.5$ mm requirement.

	$3 < p_T$ [GeV] < 5		$5 < p_T$ [GeV] < 20		$20 < p_T$ [GeV] < 100		$p_T > 100$ GeV	
Working point	ϵ_μ [%]	ϵ_{had} [%]	ϵ_μ [%]	ϵ_{had} [%]	ϵ_μ [%]	ϵ_{had} [%]	ϵ_μ [%]	ϵ_{had} [%]
Loose	90	1.17	98	1.06	99	0.25	98	0.12
Medium	70	0.63	97	0.85	97	0.17	97	0.07
Tight	36	0.15	90	0.38	93	0.12	93	0.04
Low-p_T (cut-based)	86	0.82	95	0.71	97	0.17	97	0.07
Low-p_T (multivariate)	88	0.73	96	0.66	97	0.17	97	0.07
High-p_T	45	0.34	79	0.60	80	0.13	80	0.05

Table 3.2: Prompt-muon efficiencies (ε_μ) and light-hadron misidentification rates (ε_{had}) for the different identification working points, evaluated in a $t\bar{t}$ MC sample in different p_T-regions in the region $|\eta| < 2.5$. By design, the Tight working point does not select any muons with $p_T < 4$ GeV, which is reflected in the $3 < p_T < 5$ GeV region. The statistical uncertainties are at least one order of magnitude smaller than the last digit reported (taken from Ref. [183]).



(a) (b)

Figure 3.10: Efficiency as a function of (a) p_T and (b) for the Loose, Medium and Tight identification WP in simulated $t\bar{t}$ events for tracks with $p_T > 10$ GeV shown separately for prompt muons and muons from hadron decays (taken from Ref. [183]).

The muon isolation criteria are based on the same idea as for the electron isolation: prompt muons can be distinguished from non-prompt muons by looking at the vicinity activity around them. A track-based isolation and a calorimeter-based isolation variables are defined.

For the track based isolation variable, as for electron (see corresponding Section 3.2 for more details), a variable size radius cone is defined with $\Delta R = \min(10\text{ GeV}/p_T, \Delta R_{\text{max}})$ with p_T the transverse momentum of the muon and $\Delta R_{\text{max}} = 0.3$. The transverse momentum around the muon $p_T^{\text{varcone30}}$ is computed as the scalar sum of the transverse momenta of tracks with $p_T > 1$ GeV excluding the muon track. The $p_T^{\text{varcone30}}$ is optimized for topologies where jets or other leptons are expected to be close to an energetic muon. Some other isolation WPs are using a fixed radius cone $\Delta R = 0.2$ with associated transverse momenta denoted p_T^{cone20}.

The calorimeter based isolation is defined with the $E_T^{\text{topocone20}}$ which corresponds to the sum of the transverse energy of the topological cell clusters within $\Delta R = 0.2$ around the extrapolated position of the muon in the calorimeter and removing energy deposition from the muon. $E_T^{\text{topocone20}}$ is also noise corrected similarly to what is done for electrons and jets.

The Table 3.3 presents the different isolation working point available for ATLAS. The WP *PLBDT-*

Isolation WP	Definition	Track p_{T} requirement
*PflowLoose**	$(p_{\mathrm{T}}^{\mathrm{varcone30}} + 0.4 \cdot E_{\mathrm{T}}^{\mathrm{neflow20}}) < 0.16 \cdot p_{\mathrm{T}}^{\mu}$	$p_{\mathrm{T}} > 500$ MeV
*PflowTight**	$(p_{\mathrm{T}}^{\mathrm{varcone30}} + 0.4 \cdot E_{\mathrm{T}}^{\mathrm{neflow20}}) < 0.045 \cdot p_{\mathrm{T}}^{\mu}$	
*Loose**	$p_{\mathrm{T}}^{\mathrm{varcone30}} < 0.15 \cdot p_{\mathrm{T}}^{\mu}$, $E_{\mathrm{T}}^{\mathrm{topocone20}} < 0.3 \cdot p_{\mathrm{T}}^{\mu}$	$p_{\mathrm{T}} > 1$ GeV
*Tight**	$p_{\mathrm{T}}^{\mathrm{varcone30}} < 0.04 \cdot p_{\mathrm{T}}^{\mu}$, $E_{\mathrm{T}}^{\mathrm{topocone20}} < 0.15 \cdot p_{\mathrm{T}}^{\mu}$	
HighPtTrackOnly	$p_{\mathrm{T}}^{\mathrm{cone20}} < 1.25$ GeV	$p_{\mathrm{T}} > 1$ GeV
*TightTrackOnly**	$p_{\mathrm{T}}^{\mathrm{varcone30}} < 0.06 \cdot p_{\mathrm{T}}^{\mu}$	
PLBDTLoose (PLBDTTight)	$p_{\mathrm{T}}^{\mathrm{varcone30}} < \max(1.8\text{ GeV}, 0.15 \cdot p_{\mathrm{T}}^{\mu})$ BDT cut to mimic *TightTrackOnly* (*Tight*) efficiency	$p_{\mathrm{T}} > 1$ GeV

Table 3.3: Definitions of the muon isolation WPs. The criteria used are listed in the second column, while the requirement on the minimum track p_{T} is shown in the third column. The WPs marked with "*" exist in two variants: one using $p_{\mathrm{T}}^{\mathrm{varcone30}}$ (*i.e.* variable ΔR cone) and the other using $p_{\mathrm{T}}^{\mathrm{cone20}}$ (*i.e.* fixed ΔR cone) for $p_{\mathrm{T}} > 50$ GeV (taken from Ref. [183]).

Loose and *PLBDTTight* are using a prompt lepton BDT to achieve respectively the same prompt muon efficiency as the *TightTrackOnly* and *Tight* isolation WPs. To be noted that all criteria including a calorimeter-based isolation are also applying a track-based selection because the former is more sensitive to pile-up [183]. The efficiencies obtained for the isolation WPs are summarized in Table 3.4.

	$3 < p_{\mathrm{T}}$ [GeV] < 5		$5 < p_{\mathrm{T}}$ [GeV] < 20		$20 < p_{\mathrm{T}}$ [GeV] < 100		$p_{\mathrm{T}} > 100$ GeV	
Working point	ϵ_{μ} [%]	ϵ_{HF} [%]	ϵ_{μ} [%]	ϵ_{HF} [%]	ϵ_{μ} [%]	ϵ_{HF} [%]	ϵ_{μ} [%]	ϵ_{HF} [%]
Loose	63	14.3	86	7.2	97	6.1	99	12.7
Tight	53	11.9	70	4.2	89	1.0	98	1.6
PflowLoose	62	12.9	86	6.8	97	5.0	99	9.1
PflowTight	45	8.5	63	3.1	87	0.9	97	0.8
HighPtTrackOnly	92	35.9	92	17.2	92	4.5	92	0.6
TightTrackOnly	80	19.9	81	7.0	94	3.2	99	3.3
PLBDTLoose	81	17.4	83	5.1	93	1.3	98	1.7
PLBDTTight	57	9.6	69	2.7	87	0.5	98	1.7

Table 3.4: Efficiencies for prompt muons (ε_{μ}) and muons from bottom and charm semileptonic decays ($\varepsilon_{\mathrm{HF}}$) for the different isolation working points, evaluated in a $t\bar{t}$ MC sample in different p_{T}-regions for tracks satisfying the Medium identification and the vertex association criteria. The isolation working points considered correspond to the variants with the cone size remaining constant at $\Delta R = 0.2$ for $p_{\mathrm{T}} > 50$ GeV. The statistical uncertainties are at least one order of magnitude smaller than the last digit reported (taken from Ref. [183]).

3.4 Taus

Due to their mean lifetime of 290.3 fs [18], tau leptons decay before reaching the ATLAS detector. Taus can decay either leptonically ($\tau \to l\nu_l\nu_\tau, l = e, \mu$) or hadronically ($\tau \to \nu_\tau +$ hadrons) with respective branching ratio of 34% and 66%. Leptonic decays produce isolated electrons or muons, and neutrinos that cause momentum imbalance in the transverse plane (see Section 3.8). Hadronic decays can be classified in two categories, the 1 prong and 3 prong decays, which contain respectively 1 and 3 charged particles (mostly $\pi^{\pm}$) with respective probabilities of 72% and 22%. In 68% of cases those hadronic decays contain at least one neutral pion.

In the ATLAS detector, hadronic tau decays are reconstructed as jets with the anti-k_t $R = 0.4$ clustering algorithm (the anti-k_t algorithm is discussed in Section 3.6.1) which are required to have $p_{\mathrm{T}} > 10$ GeV, $|\eta| < 2.5$ and one or three charged tracks within a cone of $\Delta R = 0.2$ around the jet axis [186]. The neutrino from the hadronic tau lepton decay cannot be reconstructed and the

combination of all visible decay products is referred to as $\tau_{\text{had-vis}}$.

In order to reject the background from quark- and gluon-initiated jets, ATLAS was previously using a boosted decision tree (BDT) based algorithm [187] and recently switched to a recurrent neural network (RNN) identification algorithm [188]. The Figure 3.11 presents a comparison of the background rejection versus the tau identification efficiency for the RNN and BDT based algorithm: an increase of performances is obtained with the RNN.

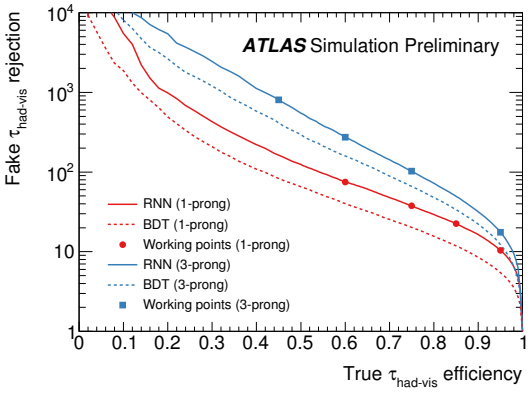


Figure 3.11: Rejection of quark and gluon jets misidentified as hadronic τ-decays (fake $\tau_{\text{had-vis}}$) depending on the true $\tau_{\text{had-vis}}$ efficiency. The 1-prong (red) and 3-prong (blue) ROC curves are shown using the RNN-based (full line) and the BDT-based (dashed line) identification algorithms. The markers correspond to the four defined working points *Tight*, *Medium*, *Loose* and *Very loose* with increasing signal selection efficiencies (taken from Ref. [188]).

3.5 Photons

Since in this thesis photons are not used, only a brief description of their reconstruction and identification is provided.

Reconstruction of photon EM showers is performed in the same way as for electrons (described in Section 3.2) [178]. Photons can be split in two categories the converted and unconverted ones distinguished from one another by using tracking information. Converted photons interacted with the tracker material in photon conversion process ($\gamma \to e^+e^-$) which results in having at least one track originating from a vertex, whereas unconverted photons do not have such a track: the energy deposit for converted photons tends to be wider because the e^+e^- pair aperture is amplified by the solenoidal magnetic field. The identification of photons is based on cuts related to the EM shower exploiting its shape and energy deposition characteristics. For instance, a search for local energy maxima within the transverse energy profile of the EM shower is performed in the finely segmented layer 1 of the EM calorimeter to distinguish prompt photons (1 maximum) from fakes (typically $\pi^0 \to 2\gamma$) which have 2 maxima. Three identification working points (Loose, Medium and Tight) optimizing the cuts applied on the discriminant distributions in different η bins are proposed by ATLAS. Similarly to electrons, isolation criteria are defined for photons.

3.6 Jets

A jet is the result of the hadronization of a parton (quark or a gluon) leading to a spray of collimated hadrons in the detector. Jets in ATLAS are reconstructed using calorimeter based and/or pixel detector based information. Afterwards, a full calibration of the jets is performed to correct for various effects.

3.6.1 Reconstruction of jets with the anti-k_t algorithm

Jets in ATLAS are reconstructed thanks to 3D topological clusters (topo-clusters), the reconstruction of topo-clusters being fully explained in Section 3.2.1. For each event, a correction to account for the position of the primary vertex referred to as *origin correction* is applied to every topo-cluster based on its depths within the calorimeter and its pseudorapidity: they initially points back to the origin of the ATLAS detector.

Then the anti-k_t clustering algorithm [189] which is a fast, infrared[1] safe and collinear[2] safe algorithm is used to iteratively merge the topo-clusters. For the high transverse momentum jets it creates a cone of circular radius in the η-ϕ plane. Other reconstruction algorithms are not necessarily meeting those stability criteria.

The association of clusters is performed computing the distance (d_{ij}) between two clusters i and j and the distance (d_{iB}) between the cluster i and the beam B:

$$d_{ij} := \min\left(\frac{1}{p_{\mathrm{T,i}}^2}, \frac{1}{p_{\mathrm{T,j}}^2}\right)\frac{\Delta R_{ij}^2}{R^2}, \tag{3.6.1}$$

$$d_{iB} := \frac{1}{p_{\mathrm{T,i}}^2}, \tag{3.6.2}$$

with R the radius parameter to be chosen which determines the final size of the jets, $\Delta R_{i,j}$ the angular distance between the clusters (defined in Equation (2.2.7)), $p_{\mathrm{T,i}}$ and $p_{\mathrm{T,j}}$ being the transverse momentum of the cluster i and j.

The association of clusters is performed in the following way. For each cluster i:

- ➢ compute d_{iB} and the smallest distance d_{ij} among all clusters j.
  - If the smallest distance is d_{ij} then the cluster i and j are merged in a single new cluster.
  - Otherwise, if the smallest distance is d_{iB}, then the cluster i is considered as a jet, and it is removed from the list of clusters considered.
- ➢ That procedure is repeated with the updated list of objects until no objects are remaining.

The iterative process and $\min(p_{\mathrm{T,i}}^{-2}, p_{\mathrm{T,j}}^{-2})$ term ensures that clusters associated to hard particles are predominant compared to soft particles ones: the jets formed are hence infrared safe. The uncalibrated physical quantities associated to the jets (direction, energy, transverse momentum...) are obtained by means of weighted sum as for the electron reconstruction. Afterwards, a full calibration chain is applied to jets as described in Sections 3.6.2.2 and 3.6.3.2.

Depending on the value of R chosen, one can distinguish small radius jets from large radius jets respectively referred to as small-R jets and large-R jets corresponding in ATLAS to radius $R = 0.4$ and $R = 1.0$. The Figure 3.12 presents a reconstruction of large-R jets seen in the η-ϕ plane.

The use of small-R jets or large-R jets is an analysis dependent choice as it is mainly based on the topology of the events targeted. Indeed, for a particle A with mass m_A, transverse momentum p_{T}^A decaying $A \to ij$ into a pair of particles i, j, then their angular distance ΔR_{ij} for a quasi-collinear splitting can be approximated by [190, 191]:

$$\Delta R_{ij} \approx \frac{1}{\sqrt{z(1-z)}}\frac{m_A}{p_{\mathrm{T}}^A}, \tag{3.6.3}$$

1. An observable/measurement is infrared safe if it is independent from low energy particles (or alternatively long distance physics). Infrared safety implies computation can be performed with perturbative quantum field theory and be compared to experiments. A jet algorithm is infrared safe if the reconstructed quantitative properties of a jet such as its four-momentum is stable in case of emission of additional particles of negligible energy (soft emissions) such as soft gluons.
2. An observable/measurement is collinear safe if it is independent from collinear emission *i.e.* splitting into several collinear particles. A jet algorithm is collinear safe if the reconstructed quantitative properties of a jet are stable with respect to the number of emitted collinear particles.

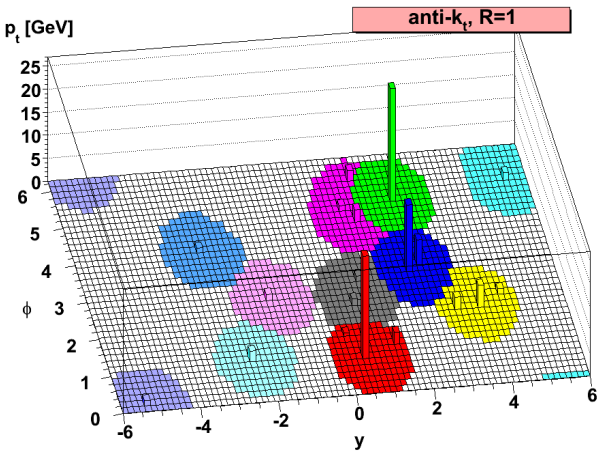


Figure 3.12: Large-R jets ($R = 1.0$) reconstructed with the anti-k_t algorithm. They form circular cones in the η-ϕ plane (taken from Ref. [189]).

where $z = p_{\mathrm{T}}^{i}/p_{\mathrm{T}}^{A}$ is the transverse momenta fraction of the particle i. For a decay into two particles of equal mass, such as the decay of the Higgs boson into a pair of b-quarks ($H \to b\bar{b}$), the Equation (3.6.3) can be further simplified as $z \approx 1/2$:

$$\Delta R_{ij} \approx \frac{2m_A}{p_{\mathrm{T}}^{A}}. \tag{3.6.4}$$

For instance, for the $H \to b\bar{b}$ decay, for low momentum Higgs bosons the events can be reconstructed identifying two small-R jets. For boosted Higgs bosons *i.e.* with high transverse momentum, the jets tends to be more collimated and the former reconstruction could be difficult: reconstructing the 2 hadrons in a single large-R jet as schematized in Figure 3.13 is more appropriate. Those two event topologies are respectively called resolved and boosted topologies: the transition between the two regimes being for transverse Higgs momenta of few hundreds of GeV as $m_H = 125$ GeV.

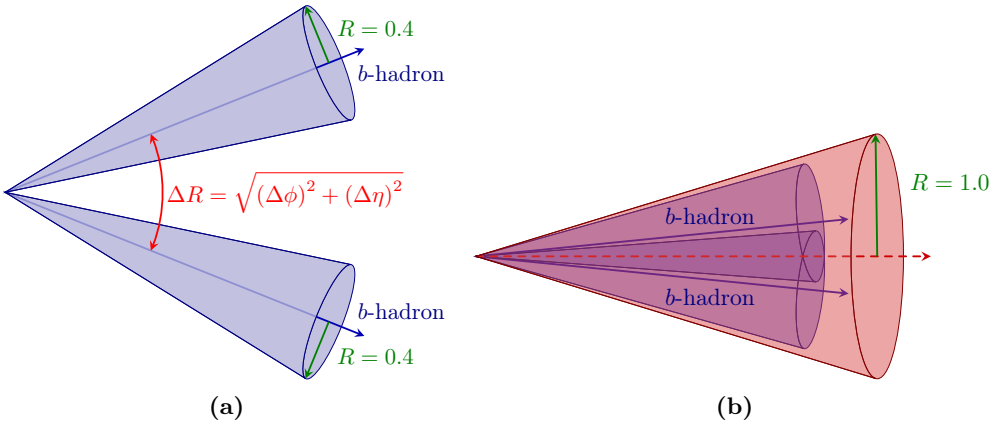


Figure 3.13: $H \to b\bar{b}$ decay: (a) resolved and (b) boosted topologies. Events are respectively reconstructed with small-R jets ($R = 0.4$) and a large-R jet ($R = 1.0$). The angular distance between the b-hadrons being: $\Delta R(b, b) \approx \frac{2m_H}{p_{\mathrm{T}}^{H}}$.

3.6.2 Small-R jets

The small-R jets are reconstructed using a radius parameter $R = 0.4$. Several energy reconstruction algorithm can be used as detailed below. Small-R jets undergo a calibration procedure to improve

their energy reconstruction, direction and so on.

3.6.2.1 PFlow and EMTopo reconstruction algorithms

The jet calibration algorithm in ATLAS has recently been changed passing from electromagnetic topological (EMTopo) jets to particle flow (PFlow) jets [192–194].

The EMTopo algorithm exploits only information from the calorimeter: jets are only reconstructed from topo-clusters [195] calibrated at the EM energy scale, the EM scale being the energy depositions from electromagnetic showers measurement after correction of pile-up and electronic noise corrections.

On the other hand, the PFlow algorithm combines the measurements from the inner detector and calorimeter to reconstruct the energy flow of the event. The energy deposited in the calorimeter by charged particles is subtracted from the observed EM energy scaled topo-clusters and replaced by the momenta of tracks that are matched to those topo-clusters. The PFlow-jets comes with several improvements in particular a better reconstruction efficiency, jet energy and angular resolutions are expected with as well a higher suppression of pile-up contamination compared to EMTopo-jets as explained below.

The energy resolution of the calorimeter and inverse transverse momentum resolution of the tracker for a single charged pion in the center of the detector are [192]:

$$\frac{\sigma(E)}{E} = \frac{50\%}{\sqrt{E}} \oplus 3.4\% \oplus \frac{1\%}{E}, \quad \sigma\left(\frac{1}{p_{\mathrm{T}}}\right) = 0.036\% \oplus \frac{1.3\%}{p_{\mathrm{T}}}, \tag{3.6.5}$$

with both E and p_{T} expressed in GeV. As a consequence, for low energy charged particles, the tracker has a significantly better resolution than the calorimeter. Also, those low charged energy particles could be bent too much by the magnetic field to be within the jet cone by the time they reach the calorimeter. With the PFlow jets, using their perigee coordinate, they can be matched to the calorimeter topo-cluster and are hence taken into account. Moreover, thanks to its high granularity, the angular resolution of the tracker is better than the calorimeter one as well as pile-up discrimination is easier as vertices are reconstructed thanks to the inner detector.

Comparison between EMTopo and PFlow jets can be found in the next Section in Figures 3.18, 3.19, 3.20 and 3.21.

3.6.2.2 Calibration of small-R jets

Once fully reconstructed, the PFlow jets go through a set of corrections [192, 193, 196] whose steps are presented in Figure 3.14 and detailed below.

➢ **Pile-up correction and residual pile-up correction**: two pileup related corrections are applied to the jets to obtain the corrected transverse jet momentum ($p_{\mathrm{T}}^{\mathrm{corr}}$) from the reconstructed ($p_{\mathrm{T}}^{\mathrm{reco}}$) one before pile-up correction is applied:

$$p_{\mathrm{T}}^{\mathrm{corr}} = p_{\mathrm{T}}^{\mathrm{reco}} - \rho \times A - \alpha \times (N_{\mathrm{PV}} - 1) - \beta \times \mu. \tag{3.6.6}$$

- The first correction consists in a subtraction of the per-event-pile-up contribution to the jet p_{T}. The pile-up contribution is estimated, in the η-ϕ plane, from the median p_{T} density ($\rho := \langle p_{\mathrm{T}}/A \rangle$) with A the jet area which is computed from the number of ghost-associated particles after its clustering *i.e.* particles matched to the jet (the ghost association is detailed in Section 3.6.4). The jet area is a measurement of how much a jet can be affected by pile-up. In order to have a correct estimation, the p_{T} density is only computed in the central region

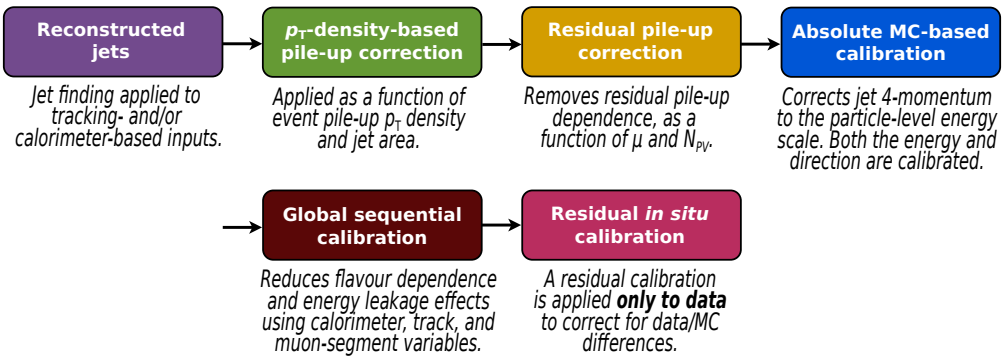
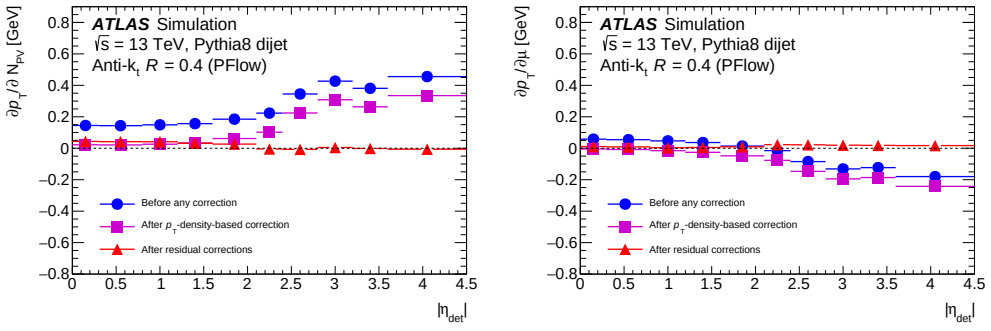


Figure 3.14: Different steps of jet calibration (taken from Ref. [193]).

$|\eta| < 2$ because the detectors in the forward region have a lower acceptance that would bias the median p_{T} density if included in the measurement.

- The second correction is because the pile-up density (ρ) is derived in central region which have a lower occupancy as most pile-up is expected in forward regions. An additional correction is needed to account for residual discrepancies between the reconstructed jet p_{T} and its truth value. That correction is derived as a function of the number of primary vertices (N_{PV}) and the number of inelastic collisions (μ) corresponding respectively to in-time and out-of-time pile-up[3]. The dependence was observed to be linear for both terms. Two coefficient α and β respectively associated to N_{PV} and μ correction terms are derived in bins of truth p_{T} ($p_{\mathrm{T}}^{\mathrm{true}}$) and pseudorapidity detector region (η_{det}) by means of fits in the simulation.

The impact of the pile-up corrections is shown in Figure 3.15.



(a) (b)

Figure 3.15: Dependence of PFlow small-R jets p_{T} with respect to (a) in-time pile-up (N_{PV} averaged over μ) and (b) out-of-time pile-up (μ averaged over N_{PV}) as a function of $|\eta_{\mathrm{det}}|$ for $p_{\mathrm{T}}^{\mathrm{true}} = 25$ GeV (taken from Ref. [193]).

➢ **Absolute MC-based calibration**: an absolute jet energy scale and η calibrations is then applied to correct the reconstructed jet four-momentum to the particle-level energy and correct biases in the jet η reconstruction. Those discrepancies, as shown in Figure 3.16, are coming from transitions between the different calorimeter regions and as well from the different calorimeter technologies and granularities used. The correction are only relying on MC simulations of the detector response.

- Concerning the jet energy correction: the jet energy response ($\mathcal{R} := E^{\mathrm{reco}}/E^{\mathrm{true}}$) defined as the ratio between the reconstructed energy and the true energy of jets is the correction factor applied to the jet four-momentum. This factor is obtained from a Gaussian fit of the

3. In-time pile-up corresponds to additional proton-proton collisions occurring in the same bunch-crossing as the collision of interest while out-of-time pile-up are collisions occurring before or after the collision of interest.

$E^{\text{reco}}/E^{\text{true}}$ distribution.

- For the η calibration, the difference between the reconstructed (η^{reco}) and the truth (η^{true}) jet pseudorapidity is observed and affects energy estimation. That correction is derived as a function of the signed difference $\eta^{\text{reco}} - \eta^{\text{true}}$ as a function of the E^{true} and η_{det}.

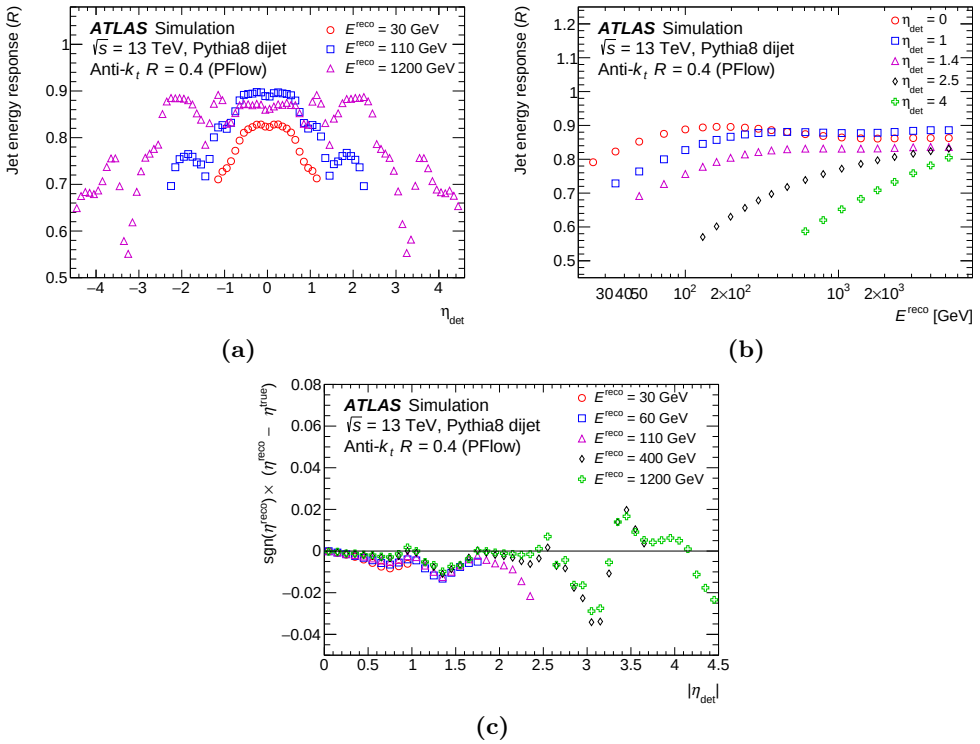


Figure 3.16: PFlow small-R jet average energy response ($\mathcal{R} = E^{\text{reco}}/E^{\text{true}}$) as a function of (a) the detector pseudorapidity η_{det} and (b) reconstructed jet energy E^{reco}. (c) Signed difference between the reconstructed (η^{reco}) and the truth (η^{true}) jet pseudorapidity. All results are derived with the PYTHIA 8 MC sample for jets with $p_{\text{T}} > 20$ GeV (taken from Ref. [193]).

After having applied the full corrections described above (origin corrections, pile-up corrections, absolute jet energy scale and η calibration) the jets are referred to be at PFlow+JES scale (or EM+JES for EMTopo jets).

➢ **Global Sequential Calibration (GSC)**: after obtaining PFlow+JES scale calibrated jets, there are still corrections to be applied to the jets. Indeed, the jet response depend on the flavour and energy distribution of particles constituting the jet. The shower development and energy profile is also dependent on the initiating particle. Quark initiated jets often include hadrons penetrating more in the calorimeter as they carry a higher fraction of transverse momentum compared to gluon initiated jets which tend to have a higher particle multiplicity with lower p_{T} fraction per particle inducing a larger transverse profile and a lower calorimeter response.

The Global Sequential Calibration is a set of multiplicative successive corrections aiming for a reduction of the fluctuation effects mentioned previously. The GSC, whose impact on the jet energy response resolution is illustrated in Figure 3.17, consists in the 6 independent corrections based respectively on the variables below (listed in the order in which they are applied). All corrections are derived as a function of p_{T} except the punch-through correction:

- f_{charged} is the fraction of the jet p_{T} measured from ghost-associated tracks with $p_{\text{T}} > 500\,\text{MeV}$ ($|\eta_{\text{det}}| < 2.5$).
- f_{Tile0} is the fraction of the jet p_{T} measured in the first layer of the hadronic Tile calorimeter ($|\eta_{\text{det}}| < 1.7$).
- f_{Lar3} is the fraction of the jet p_{T} measured in the third layer of the electromagnetic LAr

calorimeter ($|\eta_{\text{det}}| < 3.5$).

- n_{trk} is the number of tracks with $p_{\text{T}} > 1$ GeV ghost-associated to the jet.
- w_{trk} is the track width correction. It corresponds to the average p_{T}-weighted transverse distance between the jet axis and all tracks of $p_{\text{T}} > 1$ GeV ghost-associated to the jet ($|\eta_{\text{det}}| < 2.5$).
- n_{segments} the number of muon track segments ghost-associated to the jet ($|\eta_{\text{det}}| < 2.7$). This correction is called the punch-through correction, it allows to reduce the tails of the response distribution due to high-p_{T} jets whose shower is not fully contained in the calorimeter inducing an energy leakage. For this reason, the correction is derived as a function of the energy which is more relevant than the p_{T}.

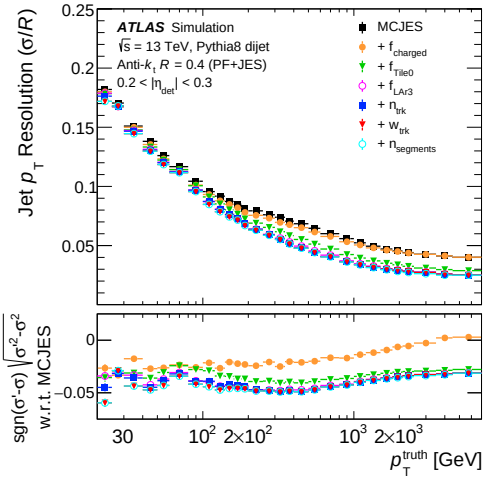


Figure 3.17: Fractional jet resolution ($\sigma_{\mathcal{R}}/\mathcal{R}$) of PFlow+EM small-R calibrated jets for $0.2 < |\eta_{\text{det}}| < 0.3$ measured in PYTHIA 8 dijet MC simulation after each step of the global sequential calibration (GSC). All jet flavours, including b-jets, are considered. The lower panel shows the difference in quadrature between the resolution before any GSC correction is applied (σ) and after the corresponding GSC step is applied (σ'). The fractional jet resolution represents the fluctuation of the jet energy: the figure shows the more corrections are applied the more the fluctuation is reduced (taken from Ref. [193]).

➢ **Residual in-situ calibrations**: this correction is only applied to data once jets have been through the GSC corrections. That final step consist in correcting differences between the jet response in data and simulation.

The jet response is obtained by balancing the jet p_{T} of the jet to calibrate with respect to a reference object or system. This procedure called the *jet balance method* it is exploiting the conservation of momentum: before the collision partons inside the proton do not have transverse momentum so the sum of transverse momenta after the collision should also cancel. The response $\mathcal{R}_{\text{in situ}}$ is obtained from the average ratio of the jet p_{T} to calibrate over the jet p_{T} of the reference object. $\mathcal{R}_{\text{in situ}}$ is derived in p_{T}-bins of the reference object by means of Gaussian fit of the peak. Indeed, the ratio (c) of the response in data and MC simulation is used:

$$c = \frac{\mathcal{R}^{\text{data}}_{\text{in situ}}}{\mathcal{R}^{\text{MC}}_{\text{in situ}}}, \tag{3.6.7}$$

which is more robust to potential biases effects such as additional radiation of jets (initial and final state radiations).

There are three steps for the in-situ calibration (applied in the listed order):

- the η-intercalibration: the energy scale of forward jets ($0.8 \leq |\eta_{\text{det}}| \leq 4.5$) is corrected to match the central jets one ($|\eta_{\text{det}}| < 0.8$) *i.e.* the jet balanced method is used with one forward and one central jet taking the central jet as a reference: : it is referred to as *dijet balance* method.

- The jet balance method is then used in Z+jet and γ+jet events using the calibrated Z-boson or photon as reference to correct energy of central jets only.
- For high-p_{T} jet to calibrate, the so-called *multijet balance* (MJB) method is used: the jet to calibrate is recoiled against several well calibrated lower p_{T}-jets referred to as the recoil system.

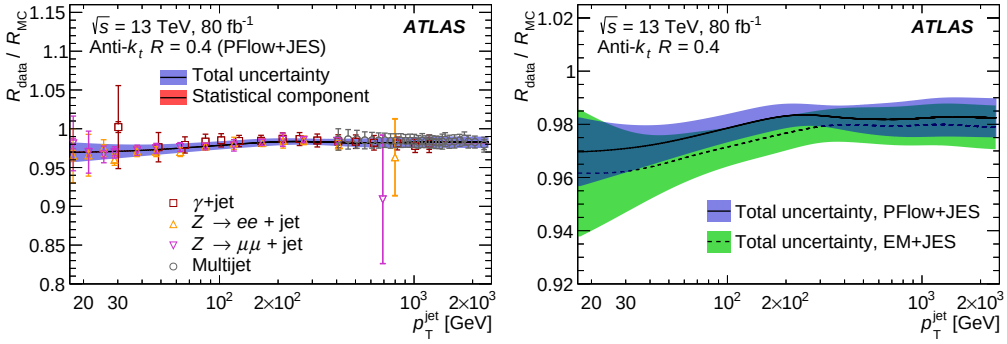
The combination of both Z/γ+jets and MJB measurement result in a smooth calibration over the full p_{T}-range measured in the ATLAS detector of the jet energy scale. Since those steps are performed sequentially systematic uncertainties are propagated from one step to the next one.

The jet energy scale (JES) refers to all jet energy related corrections, its uncertainties are related to the in situ calibration method, flavour of jets that account for the uncertainties on the composition of jets and detector response, the pile-up uncertainties and η intercalibration uncertainties. The Figure 3.18, 3.19 and 3.21 present the jet energy response measurement for PFlow and EMTopo jets, a comparison of the EMTopo versus PFlow JES uncertainties and the PFlow breakdown of uncertainties. The JES uncertainties are comparable for PFlow and EMTopo jets.

Finally, once jets have been through the full calibration procedure described above, the jet energy resolution (JER) is estimated. The JER plays a leading role for many Standard Model precision measurement and can be expressed as:

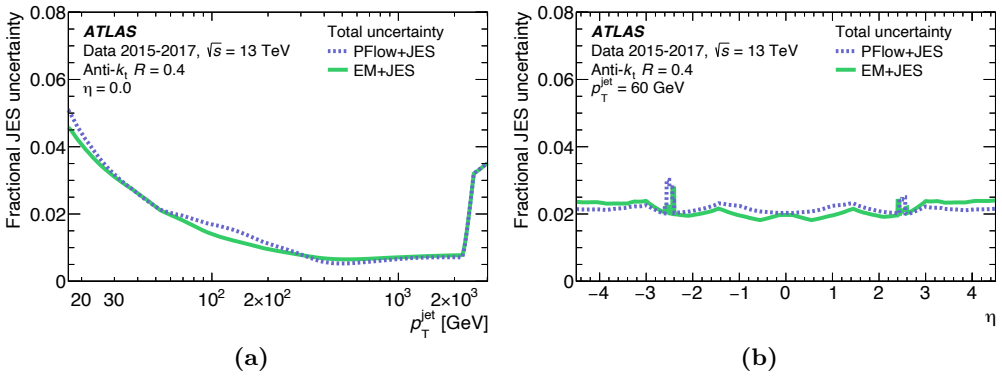
$$\frac{\sigma(p_{\mathrm{T}})}{p_{\mathrm{T}}} = \frac{N}{p_{\mathrm{T}}} \oplus \frac{S}{\sqrt{p_{\mathrm{T}}}} \oplus C, \tag{3.6.8}$$

with N the noise term, S the stochastic coefficient and C the constant term. The JER is obtained from measurement in dijet, γ+jet and Z+jet events as described in Reference [193], the constant term being measured in data events with a random cones technique to estimate fluctuations of the deposited energy in the calorimeters. The Figure 3.20 and 3.21 shows a comparison of the jet energy resolution for EMTopo and PFlow jets and the PFlow breakdown of uncertainties. The jet energy resolution is improved for PFlow jets.



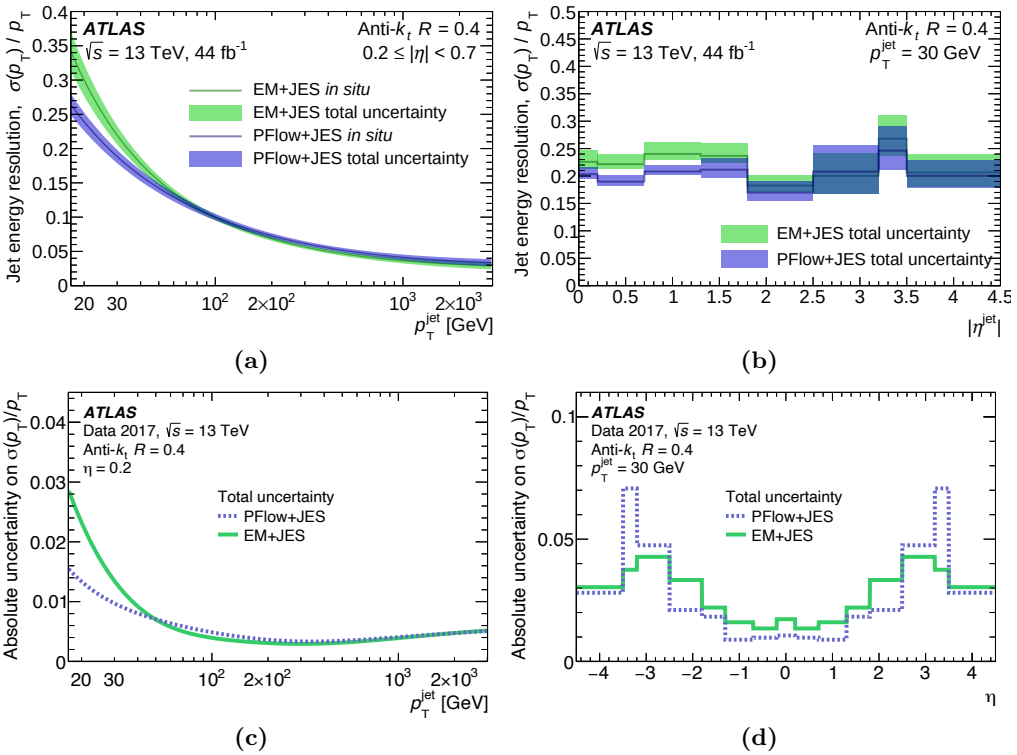
(a) **(b)**

Figure 3.18: (a) Jet response ratio ($\mathcal{R}^{\mathrm{data}}_{\text{in situ}}/\mathcal{R}^{\mathrm{MC}}_{\text{in situ}}$) as a function of the jet p_{T} for PFlow+JES small-R jets obtained combining Z+jet, γ+jet and multijet in situ calibrations. The final calibration is obtained from the inverse ratio ($\mathcal{R}_{\mathrm{MC}}/\mathcal{R}_{\mathrm{data}}$) which is taken as the jet energy scale factor and is applied in data (b) Comparison of the jet response ratio for PFlow+JES and EM+JES small-R jets (taken from Ref. [193]).



(a) (b)

Figure 3.19: Fractional jet energy scale (JES) systematic uncertainty (a) as a function of p_{T} for $\eta = 0$ and (b) as a function of η for $p_{\mathrm{T}} = 60\,\mathrm{GeV}$. The total uncertainty is shown for both EMTopo and PFlow jets (taken from Ref. [193]).



(a) (b)

(c) (d)

Figure 3.20: Comparison of the jet energy resolution (JER) for PFlow+JES jets (blue curve) and EMTopo+JES small-R jets (green curve) (a) as a function of p_{T} and (b) as a function of η. Fractional JER systematic uncertainty (c) as a function of jet p_{T} at $\eta = 0.2$ and (d) as a function of η at $p_{\mathrm{T}} = 30\,\mathrm{GeV}$. The total JER uncertainty is shown for both EM+JES and PFlow+JES small-R jets (taken from Ref. [193]).

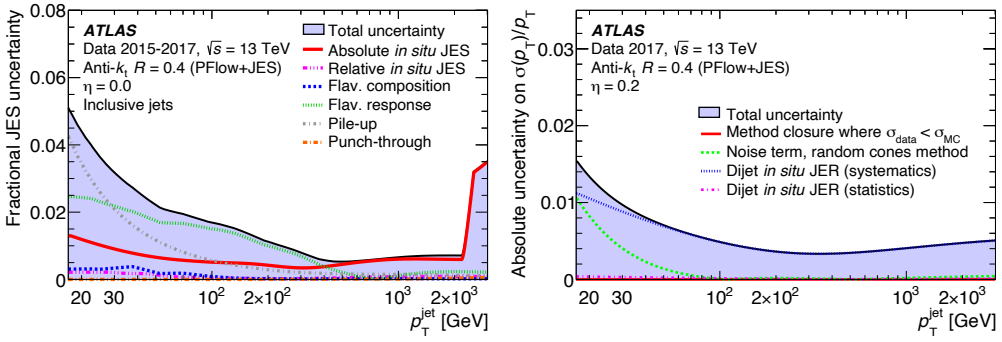


Figure 3.21: Fractional (a) jet energy scale (JES) and (b) jet energy resolution (JER) uncertainties for PFlow+JES small-R jets as a function of p_{T} respectively for $\eta = 0$ and $\eta = 0.2$. The relative in situ JES uncertainty refers to the η intercalibration uncertainties while all other in situ measurements are combined into the absolute in situ JES uncertainty (taken from Ref. [193]).

3.6.3 Large-R jets

As for small-R jets, large-R jets are passed though reconstruction and calibration algorithms. The explanations in this Section are more succinct compared to the previous Section 3.6.2 because the full process is pretty similar to small-R jets. The Figure 3.22 is summarizing the different steps of the procedure.

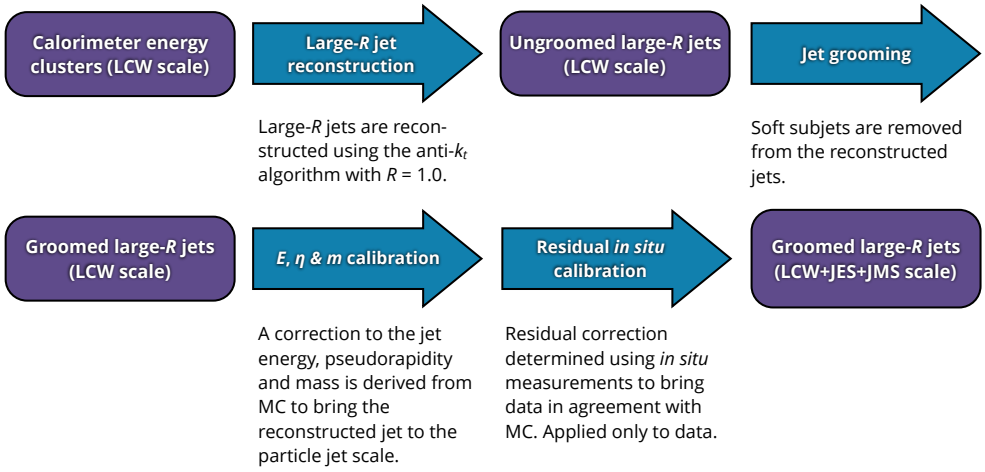


Figure 3.22: Large-R jet reconstruction and calibration procedure (taken from Ref. [197]).

3.6.3.1 Reconstruction of large-R jets

Large-R jets are reconstructed with the anti-k_t algorithm using a radius parameter $R = 1.0$ [197, 198] in ATLAS. The calorimeter topo-clusters are used to reconstruct the large-R jets which are then adjusted to point at the event's primary vertex. The large-R jets are formed with topo-clusters calibrated at the local hadronic cell weighting (LCW) energy scale [182] instead of the EM scale for small-R jets. The LCW energy scale is an additional calibration on top of the EM energy scale.

This LCW procedure aims at correcting:

➢ **the non-compensation of the calorimeter response:** in ATLAS, the signal induced by hadrons in the detector is smaller than the one of electromagnetic particles for a same energy deposition. As a result specific local corrections for electrons and pions energy deposit are needed to

improve the jet resolution and the linearity of the energy response.

➢ **Signal losses due to clustering**: the noise suppression correction can induce some signal loss due to clustering in case of wrong estimation of the pile-up conditions. An additional correction is helping to recover some signal that would be lost otherwise.

➢ **Energy deposition in dead material within the detector**: energy can be lost in inactive material and this should be taken into account in a correction.

Each topo-cluster is classified either as electromagnetic- or hadronic-like using a likelihood function which is based on geometrical considerations of the topo-cluster and the energy deposition. The clusters are then energy corrected and weighted based on the likelihood and their intrinsic characteristics before recomputing again the jet properties (energy, direction and so on). The jets obtained are equivalently referred to as Topo+LCW, Topo+LC or LCTopo large-R jets.

In order to reduce the effect of pileup and soft emission particles, a grooming algorithm [199] is used for large-R jets. In ATLAS, the grooming is performed with a trimming procedure [200]. As illustrated in Figure 3.23, it consists in reclustering constituents of the large-R jet into subjets (of radius $R = 0.2$) and remove the ones satisfying $p_{\mathrm{T}}^{\text{subjet}}/p_{\mathrm{T}}^{\text{large-R jet}} < f_{\text{cut}}$ with $f_{\text{cut}} = 5\%$. The remaining constituents form the large-R jet. The obtained large-R jets are mentioned hereafter as groomed LCTopo large-R jets.

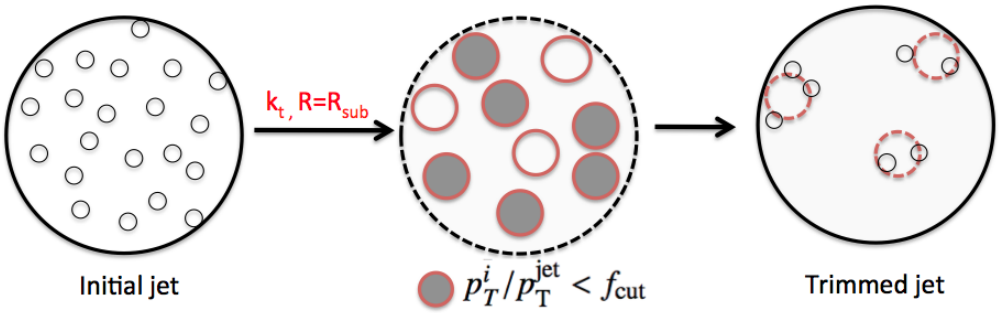


Figure 3.23: Large-R jet trimming procedure. Large-R jet ($R = 1.0$) are reconstructed with the anti-k_t algorithm (left). The topo-clusters constituents of the large-R jet are reclustered into subjets (center) using the k_t algorithm [201] with a radius parameter $R_{\text{sub}} = 0.2$. Any subjet whose transverse momentum fraction is smaller than $f_{\text{cut}} = 5\%$ (gray subjets in central image) of the large-R jet is removed (right) (taken from Ref. [199]).

3.6.3.2 Calibration of large-R jets

The obtained groomed LCTopo large-R jets need to be calibrated.

The reconstructed energy (E) and mass (m) of the large-R jets are first corrected using reco-to-truth jet energy and mass response ratio ($E_{\text{reco}}/E_{\text{true}}$ and $m_{\text{reco}}/m_{\text{true}}$) based on MC simulations. The pseudorapidity of the jet is also corrected in the same way as for small-R jets. Once those particle-level corrections are applied, the large-R jets are referred to as "JES+JMS", JES being the jet energy scale and JMS the jet mass scale.

To improve the mass resolution, a combination of the jet mass measurement from the calorimeter and the measurement of charged particle components of the jet within the inner detector is performed. The track-jets (description in next Section 3.6.4) are required to be ghost-associated to the large-R jet and must satisfy $p_{\mathrm{T}} > 500$ MeV. For the mass combination, the track-assisted (m^{TA}) mass of the large-R jet is defined as:

$$m^{\text{TA}} := \frac{p_{\mathrm{T}}^{\text{calo}}}{p_{\mathrm{T}}^{\text{track}}} m^{\text{track}}, \tag{3.6.9}$$

where m^{track} is the mass of the large-R jet measured in the tracker, $p_{\text{T}}^{\text{calo}}$ and $p_{\text{T}}^{\text{track}}$ are respectively the transverse momenta measured in the calorimeter and tracker. The m^{TA} mass benefits from a better resolution for high-p_{T} jets where the ratio m/p_{T} is low.

The combined mass (m^{comb}) is obtained from the weighted sum:

$$m^{\text{comb}} = w_{\text{calo}} m^{\text{calo}} + w_{\text{TA}} m^{\text{TA}}, \tag{3.6.10}$$

with w_{calo} and w_{TA} the two normalized weights (*i.e.* $w_{\text{calo}} + w_{\text{TA}} = 1$) defined as:

$$w_{\text{calo}} = \frac{\sigma_{\text{calo}}^{-2}}{\sigma_{\text{calo}}^{-2} + \sigma_{\text{TA}}^{-2}}, \quad w_{\text{TA}} = \frac{\sigma_{\text{TA}}^{-2}}{\sigma_{\text{calo}}^{-2} + \sigma_{\text{TA}}^{-2}}, \tag{3.6.11}$$

with σ_{calo} and σ_{TA} the respective mass resolution of the large-R jet of the calorimeter and track-assisted measurements.

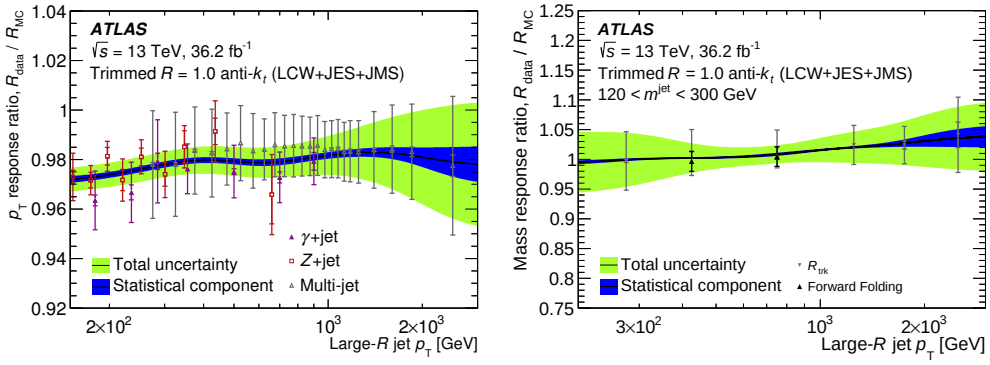
The large-R jets are then calibrated in-situ in the same way as small-R jets with the η-intercalibration procedure for $0.8 \leq |\eta| \leq 4.5$ using dijet events. Then the dijet or multijet balance method is used in Z+jets, γ+jets and multijet events, the results from those measurements are combined to derive the jet energy scale and jet energy resolution and their associated uncertainties.

Finally, an in situ jet mass calibration is performed combining results of the $\mathcal{R}_{\text{trk}}$ and forward folding methods [197]:

- **The $\mathcal{R}_{\text{trk}}$ method**: is the double ratio method with mass response factor $\mathcal{R}_{\text{trk}} = m^{\text{calo}}/m^{\text{track}}$. It is relying on the two independent large-R jet mass-measurements in the calorimeter and tracker. The comparison of the double ratio $\mathcal{R}_{\text{trk}}^{\text{data}}/\mathcal{R}_{\text{trk}}^{\text{MC}}$ enables to determine an mass scale correction for large-R jets in data. The large size of dijet samples enables to measure corrections for several p_{T} and m/p_{T} regions.
- **The forward folding method**: relies on hadronic decays of W boson and top quarks fitting their mass and jet response peaks in data and simulation to derive the in situ mass scale. Those fits are performed using $t\bar{t}$ lepton+jets events *i.e.* one of the top quark decays hadronically ($t \to Wb \to qqb$) and the other decays leptonically ($t \to Wb \to l\nu b$).

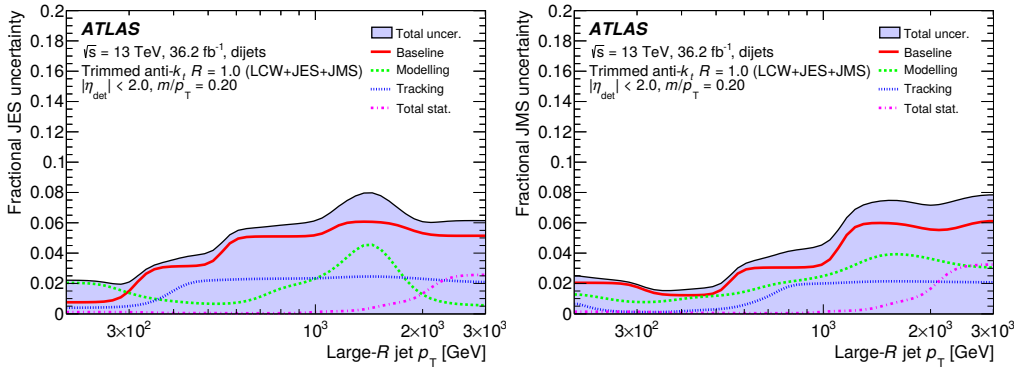
Both measurement are performed after the in-situ energy scale calibration is applied and are affecting the jet mass scale. Finally, the jet energy resolution is also obtained with dijet, γ+jet and Z+jet events as for small-R jets.

The Figure 3.24 and 3.25 are presenting the jet p_{T} and jet energy response for large-R jets as well as the JES and JMS uncertainty measurements obtained with the $\mathcal{R}_{trk}$ method.



(a) **(b)**

Figure 3.24: Ratio of the average (a) jet p_{T} response and (b) jet mass response as a function of p_{T} for trimmed LCTopo+JES+JMS large-R jets ($R = 1.0$). For the jet response, the overall results (band) are obtained from the combination of the Z+jets, γ+jets and multijet balance methods measurements. For the jet mass response the results are obtained from the combination of the $\mathcal{R}_{trk}$ and the forward folding methods for large-R jets with mass $120 < m^{\mathrm{jet}} < 300\,\mathrm{GeV}$ (taken from Ref. [197]).



(a) **(b)**

Figure 3.25: Fractional (a) jet energy scale and (b) jet mass scale uncertainties, obtained with the $\mathcal{R}_{trk}$ method, as a function of the jet p_{T} for trimmed LCTopo+JES+JMS large-R jets with $m/p_{\mathrm{T}} = 0.2$ and $|\eta_{\mathrm{det}}| < 2$ (taken from Ref. [197]).

3.6.4 VR track-jets and ghost association

Track-jets [202] are jets constructed with tracks-cluster with the anti-k_t algorithm typically with a radius parameter $R = 0.2$.

Track-jets can also be defined with a variable radius (referred to as VR track-jets) [203], they are obtained by slightly modifying the anti-k_t implementation replacing for the cluster-beam distance (d_{iB}), the radius R_{ij} by R_{eff} in Equation (3.6.2):

$$R_{\text{eff}} := \frac{\rho}{p_{\text{T}}}, \tag{3.6.12}$$

with ρ determining how fast the jet size decreases with the transverse momentum of the jet. R_{eff} enables to cope with how the ΔR angular distance between particles emitted in a jet evolves with respect to transverse momenta of the initiating particle: the larger is its p_{T}, the more its decay particles will be emitted in collimated directions, close-by to each others as described for instance by the Equation (3.6.3). Two parameters are introduced R_{min} and R_{max} in order to limit the jet radius size from becoming too large for low p_{T}-jets or too small, below the detector resolution, for high-p_{T} jets. Finally, the tracks are exclusively associated to a track-jet based on their angular distance:

$$\Delta R(\text{track}, \text{track-jet}) < 0.239 + e^{-1.220-1.64\times10^{-5}p_{\text{T}}[\text{MeV}]}. \tag{3.6.13}$$

For the $VH, H \to b\bar{b}$ analysis, the large-R calorimeter jet is used for physical interpretations but the track-jets within the large-R jet are used for b-tagging the large-R jet. Track-jets are preferred to small-R jets because of the better spatial resolution of tracks with respect to the calorimeter clusters. The parameters chosen for the $VH, H \to b\bar{b}$ analysis, for the modified anti-k_t clustering algorithm are $\rho = 30$ GeV, $R_{\text{min}} = 0.02$ and $R_{\text{max}} = 0.4$ which have been optimized for double b-tagging in $H \to b\bar{b}$ boosted decays [202] as shown in Figure 3.26 at truth-level double b-labelling.

For the $VH, H \to b\bar{b}$ analysis, the VR track-jets are matched to the large-R calorimeter jets via ghost-association [204]. Ghost-association consists in adding track-jet to the large-R jet and consider them as soft particles: "ghost" is referring to the fact that only spatial information of track-jet is kept. Afterwards, the large-R jet is reclustered with the anti-k_t algorithm with calorimeter clusters and the ghost track-jets. Track-jets whose ghost version is contained in the reclustered large-R jet are the so-called ghost-associated track-jets to the large-R jet. This procedure ensures the stability of the final large-R jet reconstruction algorithm with respect to soft particle contamination and very low p_{T} tracks.

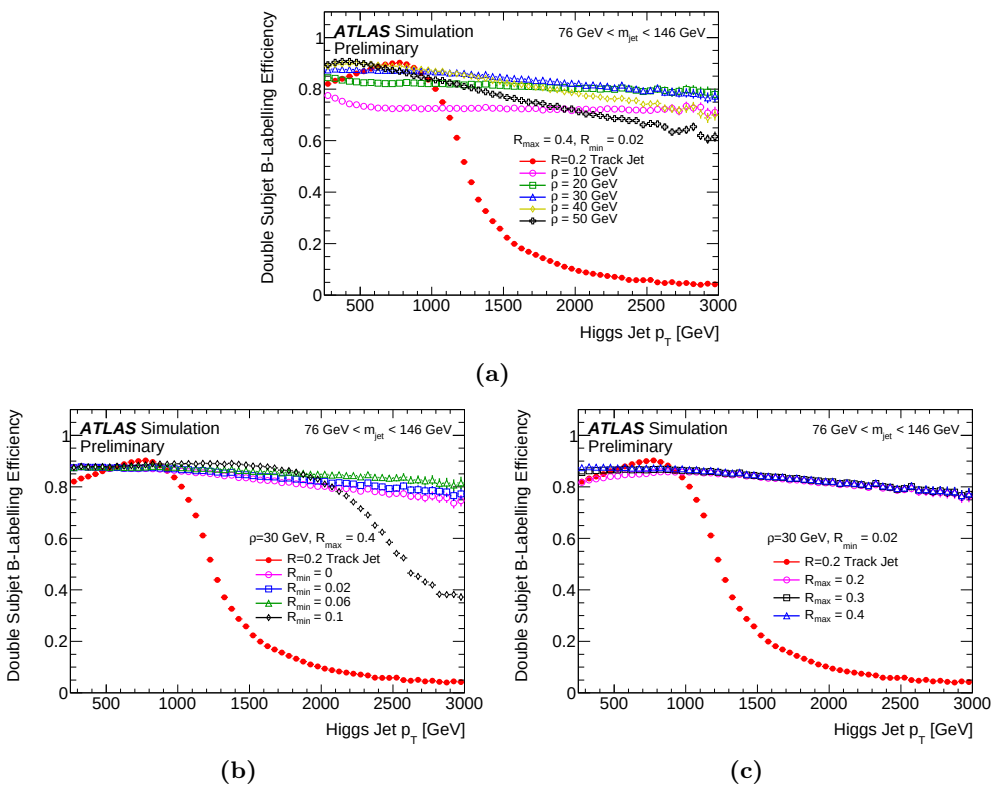


Figure 3.26: Efficiency of the track-jets double truth flavour b-labelling of a Higgs large-R jet as a function of the large-R jet-p_{T} for (a) several ρ values and $R_{\mathrm{min}} = 0.02$ and $R_{m}ax = 0.4$ for VR track-jets. (b) The efficiency for VR track-jets with $\rho = 30\,\mathrm{GeV}$ and $R_{\mathrm{max}} = 0.4$ for different values of R_{min}. (c) The efficiency for VR track-jets with $\rho = 30\,\mathrm{GeV}$ and $R_{\mathrm{min}} = 0.02$ for different values of R_{max}.For comparison, the efficiency for fixed radius parameter $R = 0.2$ track-jets is also included in all plots. For the $VH, H \to b\bar{b}$ analysis, $\rho = 30\,\mathrm{GeV}$, $R_{\mathrm{min}} = 0.02$ and $R_{\mathrm{max}} = 0.4$ are used as they correspond to the optimal parameters in terms of double b-tagging efficiency for the $H \to b\bar{b}$ boosted decays (taken from Ref. [202]).

3.6.5 Flavour tagging

Identification of heavy flavour b- and c-hadron initiated jets is a key feature for the study of the Higgs decays respectively into a pair of b or c quarks. Such b- and c-jet identifications are performed thanks to flavour tagging algorithms that are discussed in this section.

In MC simulation, the truth flavour of a jet is identified in the following way: a jet is b-labelled if a truth b-hadron with $p_{\mathrm{T}} > 5$ GeV is found within a fixed radius cone of $\Delta R = 0.3$ around the jet axis. If none is found, then a c-hadron (with $p_{\mathrm{T}} > 5$ GeV) is searched in the same way, if so then the jet is c-labelled. Otherwise, if a truth τ-lepton is found it is then τ-labelled. If that is not the case, then the jet is light-flavoured labelled.

At reconstruction level, a b- and c-tagging algorithm is used. ATLAS was previously using a boosted decision tree[4] (BDT) [205] algorithm called MV2c10 tagger [206, 207]. Since then, ATLAS has switched to a new flavour tagging algorithm, based on deep neural network[5] (DNN), referred to as DL1r tagger [207–212]. A schematization of the architecture of a BDT and a DNN is provided in Figure 3.27.

To identify b- or c-jets the following four low-level taggers below are used either by the MV2c10 or DL1r tagger whose inputs are summarized in Figure 3.28.

- **IP2D and IP3D** [206]: are two log-likelihood ratio (LLR) discriminants relying on the track impact parameters which is separating tracks associated to jets according to whether or not they are compatible with the primary vertex. IP2D only uses the transverse impact parameter significance (d_0/σ_{d_0}) while the IP3D uses both the transverse and the longitudinal ($z_0 \sin(\theta)/\sigma_{z_0 \sin(\theta)}$) impact parameter significances in order to take into account correlations between those two parameters. LLR separating b- and light-flavour jets, b- from c-jets and c-jets from light-flavour jets are computed.
- **SV1** [213]: aims at finding and reconstructing secondary vertices inside jets. The algorithm is first finding all two-track vertices, then it converts the set of two-track vertices into multi-track vertices by merging the two-track vertices that are close in space. To remove ambiguity for tracks associated to multiple vertices, each of those tracks is assigned to the vertex with the smallest track-vertex association χ^2. Finally, a track cleaning is performed for each vertex, removing iteratively the track with the largest χ^2 of the track-vertex association. The vertex fit is repeated until an acceptable vertex χ^2 and a vertex invariant mass lower than 6 GeV are obtained. If a higher value is found then the vertex is discarded. The SV1 tagger enables also to remove tracks with a high probability of coming from decays of other long-lived particles like K_S and Λ, photon conversions or material interactions. The SV1 algorithm outputs 8 discriminating variables associated to the secondary vertex such as the invariant mass of all tracks associated to the SV, the number of tracks matched to the SV, and the distances between PV and SV.
- **JetFitter** [214]: tries to reconstruct the full decay chain of b- and c-hadrons by looking for 2

4. A decision tree is a classifier performing repeated left/right (yes/no) decisions dividing phase space based on input variables until a stop criterion is fulfilled. Events are hence classified either as signal or background: a probability of being a signal event is assigned to them by the tree. Boosted decision trees is simply the extension from one to several decision trees which form a *forest*. The trees are derived from the same training ensemble by reweighting events (especially misclassified ones), and are finally combined into a single classifier which is given by a weighted average of the individual decision trees. Boosting enables to reduce overfitting peculiarities and prevents sensitivity to fluctuations in the training sample hence it enhances performances with respect to a single tree and avoids overtraining.

5. A neural network (NN), also known as artificial neural network (ANN), has a structure inspired by the human brain, mimicking the way that biological neurons signal to one another. ANNs are composed of layers with nodes, one input layer, one or more hidden layers, and an output layer. Each node, referred to as artificial neuron, connects to another and has an associated weight and threshold. If the output of any individual node is above a specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network. The output of the ANN is a weighted sum of the response of each neuron. As for BDTs, ANNs can be used a classifier for signal versus background discrimination. The term "deep" in DNN name is referring to the depth of layers in the neural network: an ANN with more than three layers is a DNN.

vertices lying on the same direction from the primary vertex *i.e.* on the b-hadron flight direction. The JetFitter computes also 8 discriminating variables which are similar to the SV1 tagger.

➢ **RNNIP** [215]: is a recurrent neural network (RNN) exploiting a sequence of track-by-track variables thus being sensitive to their correlations. Indeed, the IP3D algorithm assumes that the per-track flavour conditional likelihood can be computed independently from other tracks inside the jet which is a non-valid hypothesis. The RNNIP does hence improve significantly the discrimination between b-, c- and light-jets compared to IP2D and IP3D. The RNNIP outputs 4 probability variables respectively corresponding to the probabilities of a jet to be a b-jet, c-jet, hadronic τ-jet and a light-jet.

The MV2c10 and DL1r (high-level) taggers are using the output variables summarized in Table 3.5 from the IP2D, IP3D, SV1, JetFitter and RNNIP low level taggers. Since only the DL1r tagger uses the RNNIP tagger which comes with increased discriminating power, the DL1r achieves a better c- and light-jet rejections than the MV2c10 tagger as shown in Figure 3.29 both for small-R and VR track-jets. As a result for the $VH, H \to b\bar{b}$ analysis, an increased sensitivity to signal thanks to a better background rejection is expected using the DL1r tagger both for the resolved and boosted topologies. To be noted that flavour tagging can only be performed within the tracker acceptance *i.e.* for jets with $|\eta| < 2.5$ as track related information are primordial for discrimination.

The MV2c10 is a BDT classifier it assigns scores between -1 and 1 to jets, the closer the score is to 1 the more the tagger considers it is as a b-jet. On the other hand, the DL1r tagger outputs 3 probabilities p_b, p_c and p_{light} to jets for their respective probabilities to be a b-, c- and light-jet. For b-tagging, those probabilities are afterwards combined into a single discriminant variable:

$$\mathcal{D}_{DL1r} = \log\left(\frac{p_b}{f_c \times p_c + (1 - f_c) \times p_{\text{light}}}\right), \tag{3.6.14}$$

where $f_c = 0.018$ controls the importance of the c-jet rejection, it corresponds to the c-jet fraction used for the training sample and which was optimized for the DL1r discriminant. The similar discriminant for c-tagging can be obtained switching "b" and "c" in the Equation (3.6.14).

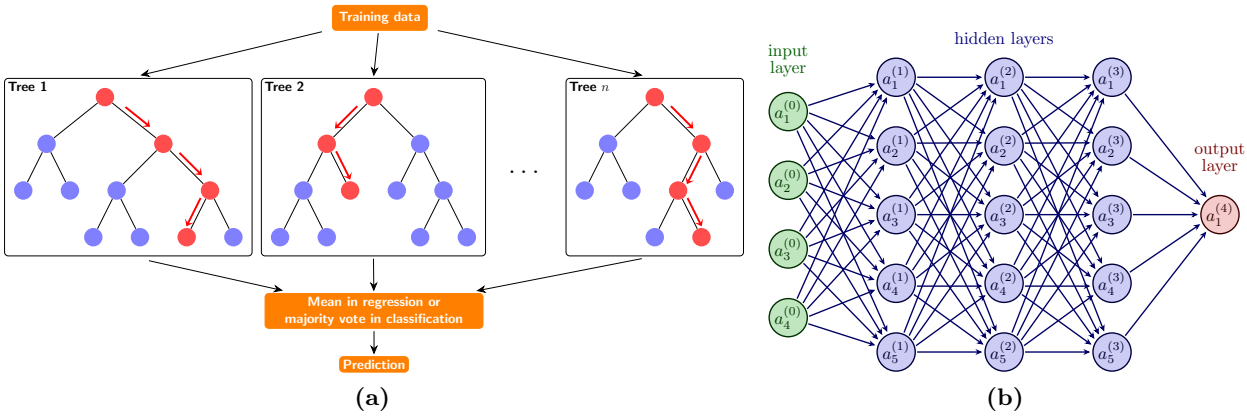


Figure 3.27: Scheme of (a) a Boosted Decision Tree (BDT) and (b) a Deep Neural Network (DNN) architecture. For the BDT scheme, each node represents a region of phase space. At each step the phase space is divided in sub-regions based on input variables. The red dots symbolize regions of phase space containing most of the signal while blue dots represent region of phase space mostly containing background events (highly inspired by Refs. [45, 216])

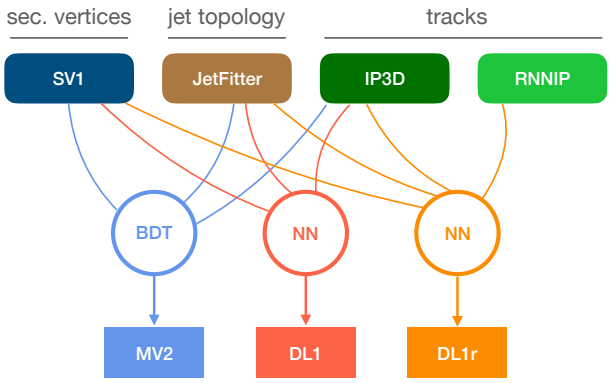


Figure 3.28: Low level taggers used by the MV2, DL1 and DL1r tagger (taken from Ref. [208]).

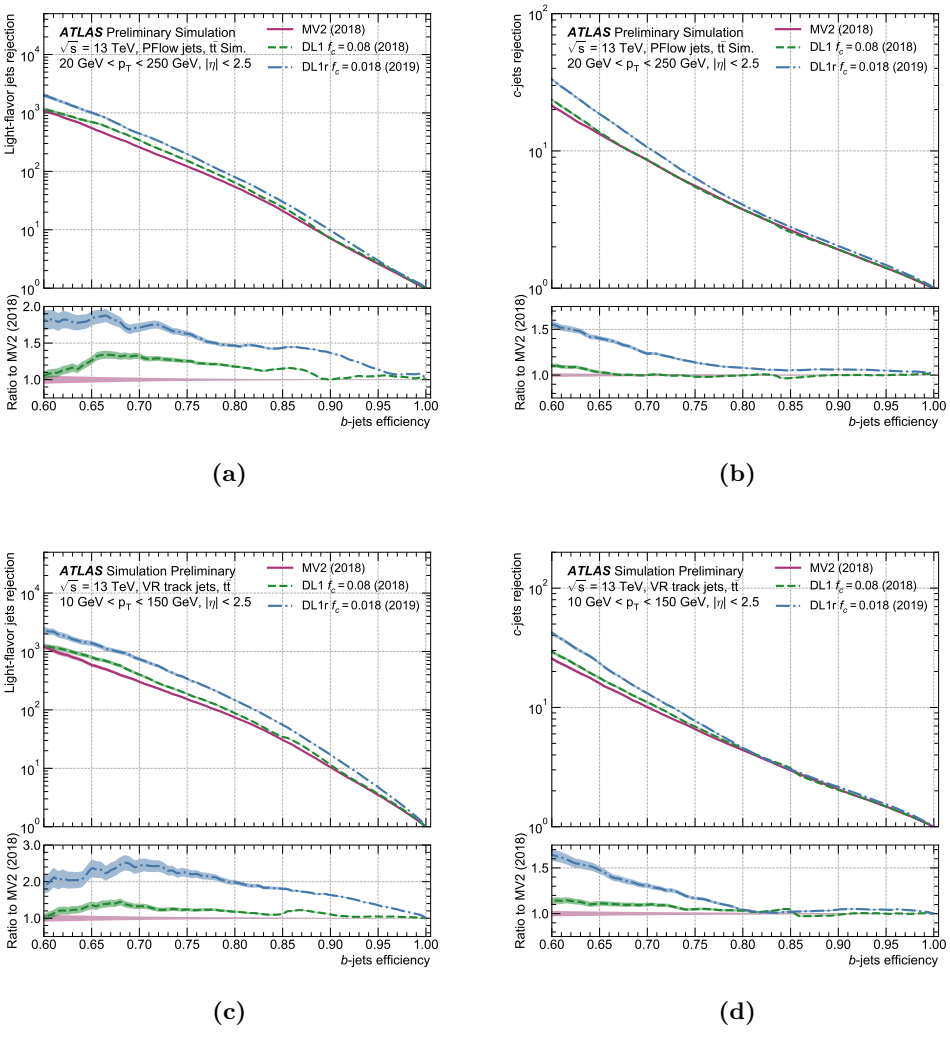


Figure 3.29: Comparison of ROC curves for 2018 recommended versions of MV2 and DL1, and 2019 DL1r optimization. The MV2c10 and DL1 algorithms are not using the RNNIP inputs variables while the DL1r algorithm does. (a) The light-flavor jets rejection ($1/\varepsilon_l$) and (b) c-jets rejection ($1/\varepsilon_c$) are shown as a function of the b-jet efficiency (ε_b). The performance is evaluated on PFlow small-R jets with $20 < p_{\mathrm{T}} < 250\,\mathrm{GeV}$ and $|\eta| < 2.5$, on a simulated dataset of $t\bar{t}$ events. (c) The light-flavor jets rejection ($1/\varepsilon_l$) and (d) c-jets rejection ($1/\varepsilon_c$) are shown as a function of the b-jet efficiency (ε_b). The performance is evaluated on VR track-jets with $10 < p_{\mathrm{T}} < 250\,\mathrm{GeV}$ and $|\eta| < 2.5$, on a simulated dataset of $t\bar{t}$ events. The shaded bands represent the statistical uncertainty (taken from Refs. [209, 210]).

Input	Variable	Description
Kinematics	p_{T}	Jet p_{T}
	η	Jet $\lvert\eta\rvert$
IP2D/IP3D	$\log(P_b/P_{\mathrm{light}})$	Likelihood ratio between the b-jet and light-flavour jet hypotheses
	$\log(P_b/P_c)$	Likelihood ratio between the b- and c-jet hypotheses
	$\log(P_c/P_{\mathrm{light}})$	Likelihood ratio between the c-jet and light-flavour jet hypotheses
SV1	$m(\mathrm{SV})$	Invariant mass of tracks at the secondary vertex assuming pion mass
	$f_E(\mathrm{SV})$	Energy fraction of the tracks associated with the secondary vertex
	$N_{\mathrm{TrkAtVtx}}(\mathrm{SV})$	Number of tracks used in the secondary vertex
	$N_{\mathrm{2TrkVtx}}(\mathrm{SV})$	Number of two-track vertex candidates
	$L_{xy}(\mathrm{SV})$	Transverse distance between the primary and secondary vertex
	$L_{xyz}(\mathrm{SV})$	Distance between the primary and the secondary vertex
	$S_{xyz}(\mathrm{SV})$	Distance between the primary and the secondary vertex divided by its uncertainty
	$\Delta R(\boldsymbol{p}_{\mathrm{jet}}, \boldsymbol{p}_{\mathrm{vtx}})(\mathrm{SV})$	ΔR between the jet axis and the direction of the secondary vertex relative to the primary vertex.
JetFitter	$m(\mathrm{JF})$	Invariant mass of tracks from displaced vertices
	$f_E(\mathrm{JF})$	Energy fraction of the tracks associated with the displaced vertices
	$\Delta R(\boldsymbol{p}_{\mathrm{jet}}, \boldsymbol{p}_{\mathrm{vtx}})(\mathrm{JF})$	ΔR between the jet axis and the vectorial sum of momenta of all tracks attached to displaced vertices
	$S_{xyz}(\mathrm{JF})$	Significance of the average distance between PV and displaced vertices
	$N_{\mathrm{TrkAtVtx}}(\mathrm{JF})$	Number of tracks from multi-prong displaced vertices
	$N_{\mathrm{2TrkVtx}}(\mathrm{JF})$	Number of two-track vertex candidates (prior to decay chain fit)
	$N_{\text{1-trk vertices}}(\mathrm{JF})$	Number of single-prong displaced vertices
	$N_{\geq\text{2-trk vertices}}(\mathrm{JF})$	Number of multi-prong displaced vertices
JetFitter c-tagging (DL1r tagger only)	$L_{xyz}(2^{\mathrm{nd}}/3^{\mathrm{rd}}\mathrm{vtx})(\mathrm{JF})$	Distance of 2^{nd} or 3^{rd} vertex from PV
	$L_{xy}(2^{\mathrm{nd}}/3^{\mathrm{rd}}\mathrm{vtx})(\mathrm{JF})$	Transverse displacement of the 2^{nd} or 3^{rd} vertex
	$m_{\mathrm{Trk}}(2^{\mathrm{nd}}/3^{\mathrm{rd}}\mathrm{vtx})(\mathrm{JF})$	Invariant mass of tracks associated with 2^{nd} or 3^{rd} vertex
	$E_{\mathrm{Trk}}(2^{\mathrm{nd}}/3^{\mathrm{rd}}\mathrm{vtx})(\mathrm{JF})$	Energy fraction of the tracks associated with 2^{nd} or 3^{rd} vertex
	$f_E(2^{\mathrm{nd}}/3^{\mathrm{rd}}\mathrm{vtx})(\mathrm{JF})$	Fraction of charged jet energy in 2^{nd} or 3^{rd} vertex
	$N_{\mathrm{TrkAtVtx}}(2^{\mathrm{nd}}/3^{\mathrm{rd}}\mathrm{vtx})(\mathrm{JF})$	Number of tracks associated with 2^{nd} or 3^{rd} vertex
	$Y_{\mathrm{trk}}^{\mathrm{min}}, Y_{\mathrm{trk}}^{\mathrm{max}}, Y_{\mathrm{trk}}^{\mathrm{avg}}(2^{\mathrm{nd}}/3^{\mathrm{rd}}\mathrm{vtx})(\mathrm{JF})$	Min., max. and avg. track rapidity of tracks at 2^{nd} or 3^{rd} vertex
RNNIP (DL1r tagger only)	P_b	b-jet probability
	P_c	c-jet probability
	P_τ	hadronic τ-jet probability
	P_{light}	light-jet probability

Table 3.5: List of input variables used by the MV2c10 and the DL1r algorithms. The JetFitter c-tagging and the RNNIP variables are only used by the DL1r algorithm (modified table taken from Ref. [207]).

3.6.6 Jet cleaning

As already mentioned, the energies of the reconstructed jets are corrected estimating the average pileup density (ρ) to avoid contamination of in-time and out-of-time pileup. Despite these corrections, local fluctuations due to those pile-up fluctuations can induce *spurious* jets. These jets are identified and removed from the events thanks to a Jet Vertex Tagger (JVT) [217–219].

The JVT is a 2 dimensional likelihood discriminant, relying on tracks information, trained in simulated dijet events. It is based on a k-nearest neighbour (kNN) algorithm that uses 2 input variables: $R_{p_{\mathrm{T}}}$ and corrJVF. $R_{p_{\mathrm{T}}}$ is the scalar sum of the p_{T} of the tracks that are associated with a jet and originate from the hard-scatter vertex (denoted PV_0) divided by the fully calibrated jet p_{T}, including pile-up subtraction:

$$\frac{\sum_k p_{\mathrm{T},k}(PV_0)}{p_{\mathrm{T}}^{\mathrm{jet}}}, \tag{3.6.15}$$

and corrJVF, standing for corrected jet vertex fraction, is defined in Reference [219], it is a corrected version of the JVF computation as it is independent from the number of primary vertices. The JVF is a variable used to identify the primary vertex from which a jet is originated from. The JVF is defined for each couple (jet, PV) in an event: it is computed as the ratio of the scalar sum of the p_{T} of tracks, matched to a jet, that originate from a given PV to the scalar sum of p_{T} of all matched tracks in the jet, independently of their origin. As a result, JVF(jet,PV) should be close to 0 if the jet is not originating from the primary vertex in question and should be close to unity otherwise.

The JVT discriminant outputs values between 0 and 1, where jets with values close to 1 are signal jet candidate while jets with scores close to 0 are considered as coming from pile-up. By means of cuts on the JVT value, pile-up jets can hence be vetoed. The JVT algorithm achieves a discriminating power stability within 1% with respect to the number of interaction per bunch crossing up to $\mu = 35$. Using the JVT, pile-up fake rates achieved are 0.4%, 1.0% and 3% respectively for signal jet efficiencies of 80%, 90% and 95%.

As the JVT relies on tracks it can only be used for jets within the inner detector acceptance $|\eta| < 2.5$. A forward jet vertex tagger (fJVT) is used to complete the JVT in forward regions [220]. A cut $\mathrm{JVT} < 0.2$ is first applied in order to remove all hard scattered central jets (98.8% of them are rejected by that cut). For jets passing that selection, the quantity $\mathrm{fJVT}(\mathrm{PV}_i)$ for each primary vertex (PV_i) is computed as:

$$\mathrm{fJVT}(\mathrm{PV}_i) = \frac{\boldsymbol{p}_{\mathbf{T,i}}^{\mathbf{miss}} \cdot \boldsymbol{p}_{\mathbf{T}}^{\mathbf{fj}}}{\left|p_{\mathrm{T}}^{\mathrm{fj}}\right|^2}, \tag{3.6.16}$$

with $\boldsymbol{p}_{\mathbf{T,i}}^{\mathbf{miss}}$ being the transverse missing momentum associated to PV_i without taking into account forward jets and $\boldsymbol{p}_{\mathbf{T}}^{\mathbf{fj}}$ is the transverse momentum of the forward jet. The fJVT score of the forward jet is then obtained from:

$$\mathrm{fJVT} = \max_i(\mathrm{fJVT}(\mathrm{PV}_i)). \tag{3.6.17}$$

For a forward pile-up jet, its energy will be balanced by $p_{\mathrm{T,i}}$ leading to fJVT values close to 1. On the other hand, for hard-scatter forward jets fJVT values closer to 0 are expected: the forward jets can be tagged as pile-up if their fJVT value is above a given threshold.

3.7 Overlap removal

In ATLAS, the reconstruction of objects is non-exclusive. The overlap removal (OR) [221], is a procedure applied, after all objects are selected with the loosest identification requirements and fully calibrated, in order to deal with multiple analysis objects being reconstructed from the same

detector response. After applying the overlap removal, the ambiguity is solved only keeping one object and removing all the other overlapping ones from the event.

The OR is a sequential procedure, it can be specific to each analysis, but it follows general concepts described below.

- ➢ **τ versus electron and muon OR**: hadronic τ candidates found withing a cone $\Delta R = 0.2$ of an electron or muon are removed from the event.
- ➢ **Muon versus τ OR**: if a muon not identified as a *combined* muon is found within $\Delta R = 0.2$ of a τ-lepton candidate of $p_{\mathrm{T}} > 50$ GeV then the muon is removed instead of applying the τ *versus muon OR*.
- ➢ **Electron versus muon OR**: if an electron and muon have a common track in the inner detector, then the electron is removed because it is considered as coming from a converted photon radiated by the muon.
- ➢ **Muon versus electron OR**: if a *calorimeter tagged* muon shares a common track with an electron, instead of applying the *electron versus muon OR*, the muon is removed, and the electron is kept.
- ➢ **Jet versus electron OR**: jets closer than $\Delta R = 0.2$ of an electron are removed because EM calorimeter information are both used to reconstruct jets and electrons. Otherwise, almost all the electrons would be double-counted as jets.
- ➢ **Electron versus jet OR**: after applying the *jet versus electron OR*, if an electron is found in a cone $\Delta R = \max\left[0.2, \min(0.4, 0.04 + 10/p_{\mathrm{T}}^{e}[\mathrm{GeV}])\right]$ around a jet then the electron is removed because it is considered as coming from semi-leptonic decays of heavy-flavoured hadrons.
- ➢ **Jet versus muon OR**: if a muon is found within $\Delta R = 0.2$ then the jet is removed if it has less than three tracks as such case is compatible with radiation from the muon.
- ➢ **Muon versus jet OR**: after applying the *Jet versus muon OR*, if a muon is found to be closer than $\Delta R = 0.2$ of a jet then the muon is considered as coming from a heavy flavour semi-leptonic hadron decay. The muon is referred to as *muon in jet*: it is removed and its energy is added to the final jet energy.

3.8 Missing transverse energy

Neutral weakly interacting particles such as neutrinos or potential dark matter particles cannot be directly detected by the ATLAS detector as they exit the detector without leaving a signature. However, the presence of such undetected particles in a collision can be inferred from the transverse momentum imbalance as the law of conservation of momentum imposes that after the collision the vectorial sum of transverse momenta of all particles should cancel since before the collisions partons in the protons have a negligible transverse momentum. As a result the vectorial sum of transverse momenta of undetected particles can be computed as the negative vectorial sum of all detected particles. The aforementioned vectorial sum is called the missing transverse momentum vector. Its magnitude is a scalar and is called missing transverse energy ($E_{\mathrm{T}}^{\mathrm{miss}}$) or momentum ($p_{\mathrm{T}}^{\mathrm{miss}}$) depending on whether it is computed from the calorimeter energy deposits or from the ID tracks. In the later case, only charged particles are taken into account, so the calorimeter-based transverse missing energy ($E_{\mathrm{T}}^{\mathrm{miss}}$) is in general preferred to the track-based transverse missing momentum ($p_{\mathrm{T}}^{\mathrm{miss}}$). To be noted that the parton longitudinal momentum is unknown before the collision as each parton carries a fraction of the proton momentum hence only the conservation of the transverse momentum can be exploited.

The $E_{\mathrm{T}}^{\mathrm{miss}}$ is computed from all the hard objects (electrons, muons, hadronic taus, photons and jets) combining information from the calorimeters and the muon spectrometer [194, 222, 223]. All the

reconstructed additional signals not associated with a reconstructed physics object are contributing to the so-called $E_{\mathrm{T}}^{\mathrm{miss}}$ "soft term" denoted $E_{\mathrm{T}}^{\mathrm{miss,soft}}$. The total missing energy E_u^{miss} in the direction $u = x, y$ can be computed as:

$$E_u^{\mathrm{miss}} = -\left(\sum_{i \in \{\text{hard objects}\}} p_{u,i} + \sum_{i \in \{\text{soft signal}\}} p_{u,i} \right), \tag{3.8.1}$$

$$E_u^{\mathrm{miss}} = E_u^{\mathrm{miss},e} + E_u^{\mathrm{miss},\mu} + E_u^{\mathrm{miss},\tau_{\mathrm{had}}} + E_u^{\mathrm{miss},\gamma} + E_u^{\mathrm{miss,jets}} + E_u^{\mathrm{miss,soft}}, \tag{3.8.2}$$

with $p_{u,i}$ the u component of the momentum of the reconstructed and calibrated objects or soft signal measured with the calorimeter. $E_u^{\mathrm{miss,obj}} = -p_{u,\mathrm{obj}}$ is the negative component of the transverse momentum for each objects along the u-axis. The total transverse missing energy is then obtained from E_x^{miss} and E_y^{miss} as:

$$\boldsymbol{E}_{\mathbf{T}}^{\mathbf{miss}} = (E_x^{\mathrm{miss}}, E_y^{\mathrm{miss}}, 0), \tag{3.8.3}$$

$$E_{\mathrm{T}}^{\mathrm{miss}} = \left| \boldsymbol{E}_{\mathbf{T}}^{\mathbf{miss}} \right| = \sqrt{(E_x^{\mathrm{miss}})^2 + (E_y^{\mathrm{miss}})^2}, \tag{3.8.4}$$

$$\phi^{\mathrm{miss}} = \tan^{-1}\left(\frac{E_y^{\mathrm{miss}}}{E_x^{\mathrm{miss}}} \right). \tag{3.8.5}$$

The double counting of various clusters is avoided by considering objects in the following order: muons, electrons, photons, hadronic taus and finally jets.

The soft terms can either be obtained with track-based soft terms (TST) or calorimeter-based soft terms (CST). The track-based terms are reconstructed from the ID tracks not associated with physics objects which are mainly coming from low momentum tracks. The calorimeter-based soft terms are computed from energy deposits in calorimeter cells. This method has the advantage to account for both the charged and the neutral components of the soft terms, but it has a much higher pile-up contamination dependence. As a result, the TST method is used for the soft-term.

Finally, the overall resolution of $E_{\mathrm{T}}^{\mathrm{miss}}$ depends on the resolution of the single objects and is dominated by the JER contribution. The RMS width of the $E_u^{\mathrm{miss}}, u = x, y$ increases as $\mathrm{RMS}\left(E_u^{\mathrm{miss}}\right) \propto \sqrt{E_u^{\mathrm{miss}}}$.

The Figure 3.30 shows a comparison of the $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution obtained in $Z \to \mu^+\mu^-$ events (no missing transverse energy is expected in such event) reconstructed with EMTopo and PFlow jets proving that a better accuracy is obtained with PFlow jets. It also shows that a better RMS resolution is obtained with PFlow jets in $Z \to e^+e^-$ events.

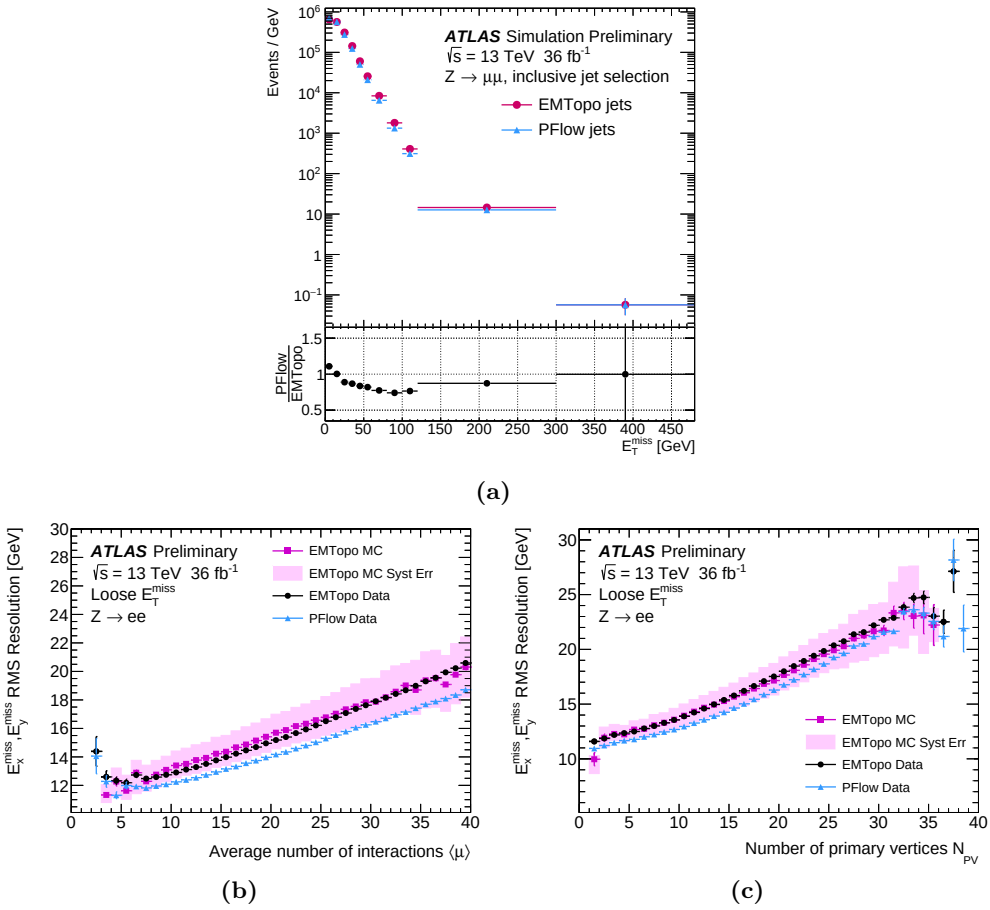


Figure 3.30: (a) Comparison of the EMTopo versus PFlow $E_{\mathrm{T}}^{\mathrm{miss}}$ distributions in $Z \to \mu^{+}\mu^{-}$ simulation using a dedicated event selection and the Loose $E_{\mathrm{T}}^{\mathrm{miss}}$ working point. While the RMS resolution obtained from the combined distributions of E_{x}^{miss} and E_{y}^{miss} is shown as a function of (b) the number of inelastic collisions per bunch crossing (μ) and (c) the number of primary vertex per bunch crossing (N_{PV}) using the Loose $E_{\mathrm{T}}^{\mathrm{miss}}$ working point for data with EMTopo jets (circular marker) and PFlow jets (triangular marker) and MC simulation with EMTopo jets (square marker) in $Z \to e^{+}e^{-}$ event selection (taken from Ref. [194]).

Chapter 4

Energy scale calibration of b-tagged jets using $t\bar{t}$ lepton+jets events

4.1 Introduction to the b-JES measurement . . . 122
4.2 Data and simulation samples . . . 127
4.3 Event selection . . . 129
4.4 In situ b-JES measurement with a template method . . . 132
4.4.1 $\chi^2(\alpha)$ curves determination . . . 133
4.4.2 b-JES statistical uncertainty . . . 138
4.4.3 b-JES systematic uncertainties and total uncertainty . . . 138
4.5 Results of the b-JES measurement . . . 142
4.5.1 Template method validation with pseudo-data . . . 143
4.5.2 Inclusive measurement in data . . . 146
4.5.3 Differential p_{T} measurement in data . . . 149
4.6 Conclusion . . . 153

This chapter presents a feasibility study and the technical details of the implementation of a measurement *in situ* of the energy scale of jets reconstructed from the hadronization of b-quarks (b-jets) and passing flavour-tagging requirements in the ATLAS experiment. The measurement uses 139 fb^{-1} of proton–proton collisions collected in 2015–2018 during the second data-taking period of the Large Hadron Collider at CERN.

A sample enriched in b-jets is selected in events with a $t\bar{t}$ lepton+jets signature consistent with the process $pp \to t\bar{t} \to \ell\nu q\bar{q}b\bar{b}$ ($\ell = e, \mu$). The b-tagged jets are reconstructed with the anti-k_t $R = 0.4$ algorithm based on calorimeter information and have pseudorapidity $|\eta| < 2.5$.

The measurement relies on the measured value of the top-quark mass and is based on a template method applied to the distribution of the difference $m_{qqb} - m_{qq}$ between the three-body invariant mass of the two non b-tagged jets and one b-tagged jet (m_{qqb}) and the two-body invariant mass (m_{qq}) of the two non b-tagged jets in four-jet events.

The measurement is both performed inclusively for b-tagged jets transverse momenta above 20 GeV, and for the first time, thanks to the large size of the sample under study, differentially as a function of the b-tagged jet transverse momentum between 30 and 500 GeV. The specific b-tagged jet energy correction is consistent with unity, with an uncertainty of 2.5% for the inclusive measurement, and decreasing from 4% to 1.9% for b-tagged jet transverse momenta increasing between 30 and 500 GeV for the differential measurement.

This chapter is organized as follows. The Section 4.1 is an introduction to the b-jet energy scale (b-JES) measurement. The data and simulation samples used for the measurement are described in Section 4.2. The selection applied to the data and simulation events is detailed in Section 4.3. Section 4.4 provides the details of the template method used in this study, including how the templates are computed, and the statistical procedures used to determine the nominal b-JES and

its statistical and systematic uncertainties. The results of the measurement are given in Section 4.5, which is followed by some concluding remarks. Some technical details about the analysis are provided in Appendix A.

The results I have obtained, have been published in a dedicated conference note [7] and have been presented [8, 9] at Moriond 2022 Electroweak Interactions & Unified Theories conference [10] during a Young Scientist Forum (YSF) session.

4.1 Introduction to the *b*-JES measurement

Jets of hadrons are produced abundantly in the proton-proton collisions at the Large Hadron Collider, and many signatures of interest for Standard Model measurements and for searches of beyond-SM phenomena include at least one jet in the final state under study.

Since top quarks almost always decay into a W boson and a b-quark, b-jets are present in any processes that involves in its final state a single top quark or top quark pair, alone or in association with other particles. Moreover, b-jets can be found in heavy flavour decays of the electroweak bosons, as well as in a large fraction of the Higgs boson decays, since the Higgs boson mainly decays into a pair of b-quarks with a branching ratio $\mathrm{BR}(H \to b\bar{b}) \approx 58\%$ [18]. New resonances could also decay to top or bottom quarks or to Higgs bosons which again would result in the presence of b-jets in the final state. To properly reconstruct the kinematics of processes producing b-jets, an accurate calibration of their energy is important.

The jet energy scale (JES), as explained in Section 3.6.2.2, refers to a correction to be applied on the energy of jets to correct for potential miscalibration of the detector and imperfections of the jet calibration procedure. The JES corrections in ATLAS are determined, *in situ i.e.* are only applied to data, in bins of jet transverse momentum and pseudorapidity from events with jets recoiling against well calibrated objects in samples enriched in γ+jet, Z+jet, or multi-jet events. These samples are dominated by jets originating from light-quarks ($q = u, d, s$), c-quarks, or gluons (g).

Due to the larger b-quark mass and the larger semileptonic branching fractions of b-hadrons, the radiation pattern of showers induced by b-quarks is different than for light-quarks or gluons, which in principle can cause a different energy response in the ATLAS detector. Currently, ATLAS prescribes an additional uncertainty to b-jets based on differences in response between different simulated samples to cover possible differences between the modelling of the response of b-jets compared to jets originating from light-quarks and gluons.

This chapter presents a measurement of a residual differential correction on top of the default JES calibration, for b-tagged jets, referred to hereafter as b-JES, and of an associated uncertainty, using the b-jets produced by the process $pp \to t\bar{t} \to \ell\nu q\bar{q}b\bar{b}$, $\ell = e, \mu$. A template method [224] is used to determine multiplicative factors, $1/(1+\alpha)$, which need to be applied to the data to calibrate these jets. The templates are based on the reconstructed distribution of the difference between the invariant mass of the jets produced by the top quark decaying hadronically ($t \to Wb \to q\bar{q}b$) and the invariant mass of the two jets from the light-quarks in the same decay, exploiting the measured values of the top-quark and W-boson masses. A similar technique was used by the ATLAS collaboration to study the energy scale of jets from the harmonization of the quarks produced by W-boson decays [225]. The b-JES is measured only for jets with pseudorapidity $|\eta| < 2.5$ passing a b-flavour-tagging algorithm. The differential measurement is performed for b-tagged jet transverse momenta between 30 and 500 GeV, profiting from the large sample of b-tagged jets populating this wide transverse momentum range. An inclusive measurement is also performed, including also b-tagged jets with transverse momenta down to 20 GeV. Previously, inclusive factors for the JES scaling of b-jets have been determined in the specific context of top-quark mass measurements by ATLAS with uncertainties of 2.4% [226] and 2% [227].

The use of the template method and the choice of the $t\bar{t}$ lepton+jets sample provide complementarity to alternative methods and processes that could be studied for the same purpose of measuring *in situ* the b-JES, such as applying the momentum-balance techniques to $Z+b$-jets [228] or $\gamma+b$-jets events.

Jets produced by b-quarks can be identified with good efficiency and large rejection of light-flavoured jets by exploiting the long lifetime and relatively large mass of b-hadrons. b-tagging algorithms are used to search for jets pointing to a secondary vertex that is displaced from the primary vertex (where the collision occured), and for tracks with large impact parameter and multiplicity as shown in Figure 3.2.

Jets are reconstructed by the ATLAS detector and their direction, energy and momentum are then determined using a sequence of online and off-line calibration algorithms [196]. The procedure is illustrated in Figure 3.14.

The jet energy is determined from the energy of its constituents through a set of corrections based on the simulation, that account for various effects such as the position of the primary vertex of the event, the detector energy response, energy leakage outside the calorimeter and pileup. Discrepancies between the data and the simulation are accounted for by *in situ* jet energy scale (JES) and resolution (JER) corrections that are only applied in data in the last step of the calibration.

At the LHC, top quark pairs and hence b-jets are produced copiously by the strong interactions between the quarks and gluons of the colliding protons, mainly in gluon-gluon fusion (ggF) or quark antiquark annihilation events. The leading-order Feynman diagrams for such processes are illustrated in Figure 4.1. These processes have large inclusive cross-sections: $\sigma_{\sqrt{s}=7\,\mathrm{TeV}}(t\bar{t} \to X) =$

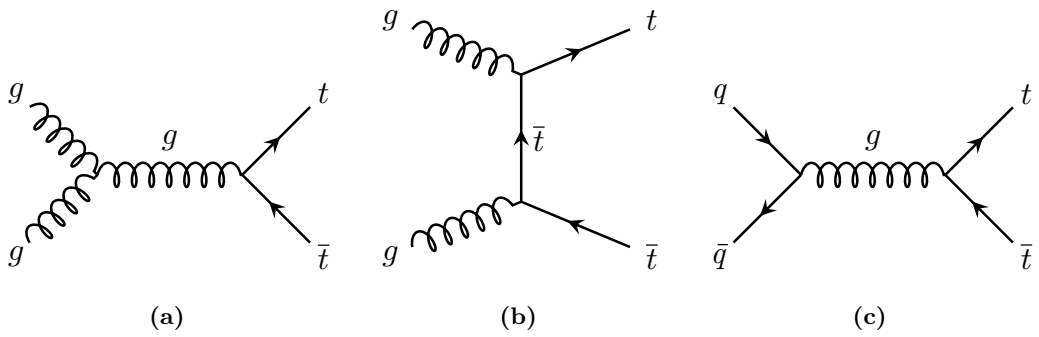


Figure 4.1: Main top-quark pair production processes: (a) gluon-gluon fusion (s-channel), (b) gluon-gluon fusion (t-channel) and (c) quark-antiquark annihilation.

177.3 pb, $\sigma_{\sqrt{s}=8\,\mathrm{TeV}}(t\bar{t} \to X) = 252.9$ pb and $\sigma_{\sqrt{s}=13\,\mathrm{TeV}}(t\bar{t} \to X) = 831.8$ pb [229–231], as shown in Figure 4.2. In particular, in the LHC Run 2, corresponding to 139 fb^{-1} of good-quality pp collision data recorded by ATLAS, this means that more than 100 million top-quark pairs have been produced and are available for analysis.

Due to their short lifetime $\tau_t \approx 5 \times 10^{-25}$ s [229], top quarks do not hadronize and instead decay via weak interaction to a W boson and a bottom quark: $\mathrm{BR}(t \to Wb) \approx 100\%$. As a result, $t\bar{t}$ decays can be classified in 3 channels based solely on how the two W bosons decay:

- **Hadronic events** (Figure 4.3a): the two W bosons decay hadronically producing light (u, d, s) and charm (c) quarks. The final state is composed of several light or charm-jets and two b-jets. The branching ratio of the hadronic channel [229] is:
$$\Gamma(t\bar{t} \to q\bar{q}bq\bar{q}\bar{b})/\Gamma(t\bar{t} \to W^+bW^-\bar{b}) \approx 45.7\%.$$
- **Lepton+jets events (also called semileptonic events)** (Figure 4.3b): one W boson decays hadronically into light- or charm-quarks. The other one decays leptonically into a charged lepton (respectively antilepton) $\ell = e, \mu, \tau$ and an antineutrino $\bar{\nu}_\ell$ (respectively neutrino) de-

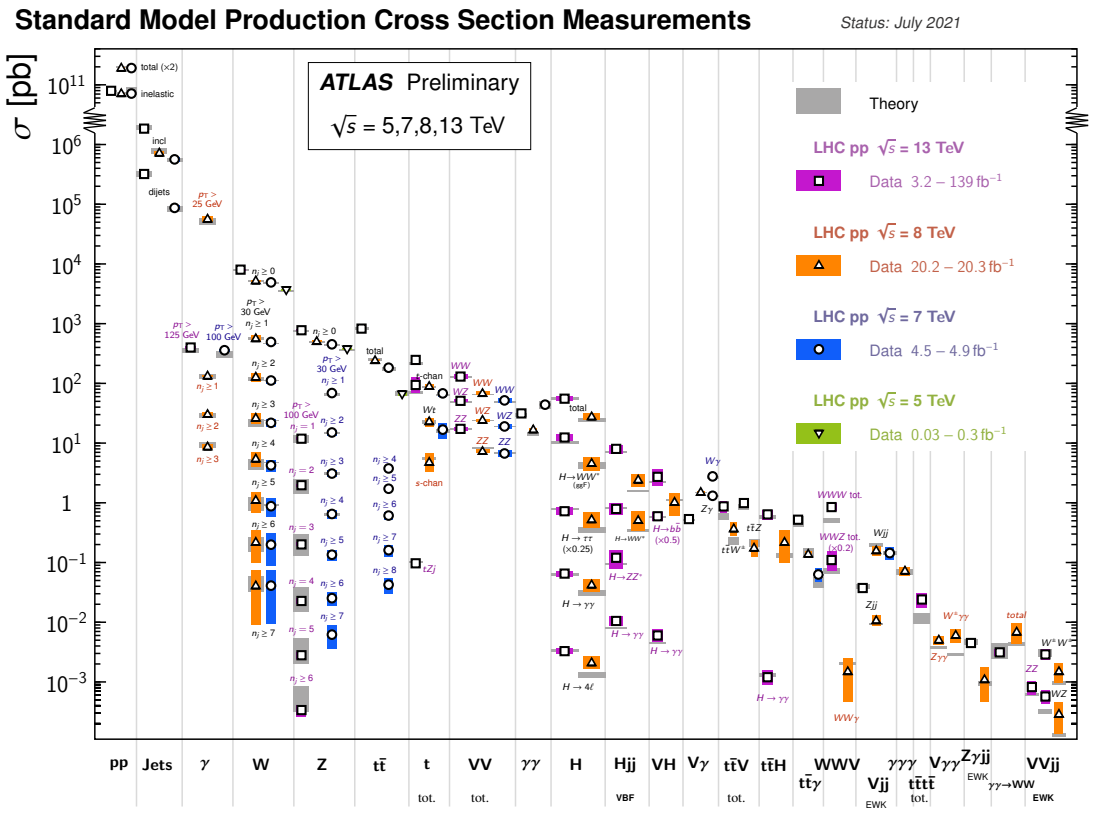


Figure 4.2: Standard Model total cross-section production at LHC for Run 1 and 2 $\sqrt{s} = 5, 7, 8, 13$ TeV (taken from Ref. [232]).

pending on the electric charge of the leptonic-decaying boson. The final state is composed of one charged lepton, several light-jets and two b-jets. The branching ratio of the lepton+jets events [229] is:

$$\Gamma(t\bar{t} \to \ell\nu b q\bar{q}\bar{b})/\Gamma(t\bar{t} \to W^+bW^-\bar{b}) \approx 43.8\%.$$

➢ **Leptonic events (also called dilepton events)** (Figure 4.3c): the two W bosons decay leptonically into two lepton-neutrino pairs. At leading-order accuracy in the strong coupling constant α_s, the only two jets in the event are those originated by the b-quark and antiquark. The branching ratio of the leptonic channel [229] is:

$$\Gamma(t\bar{t} \to \ell\nu b\ell'\nu'\bar{b})/\Gamma(t\bar{t} \to W^+bW^-\bar{b}) \approx 10.5\%.$$

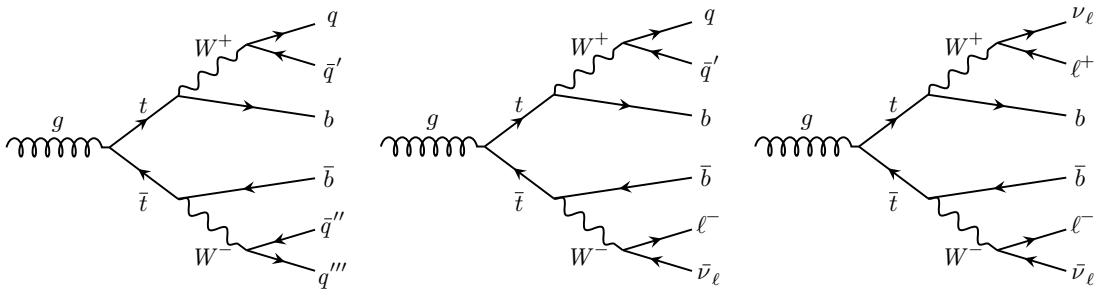


Figure 4.3: Examples of (a) hadronic, (b) lepton+jets and (c) leptonic top-quark pair decays with $\ell = e, \mu$, $U = u, c$ and $D = d, s$.

Using $t\bar{t}$ lepton+jets events for the b-JES calibration has several advantages:

- the top quark hadronic decay $t \to Wb \to q\bar{q}b$ (Figure 4.4) can be fully reconstructed and the invariant mass of the top quark can therefore be measured from the four-momenta of the three jets from the top quark decay.
- The presence of missing transverse energy, one light lepton, exactly two b-jets and a certain number of light jets is a strong signature. The lepton from the W boson decay allows efficient triggering of the events, and requirements on the presence of a lepton and on the missing transverse energy of the event largely reduce QCD backgrounds.
- The decay has a large branching ratio which ensures to have a large-enough sample for performing the b-JES calibration.

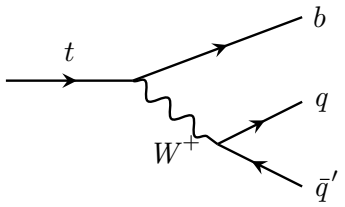
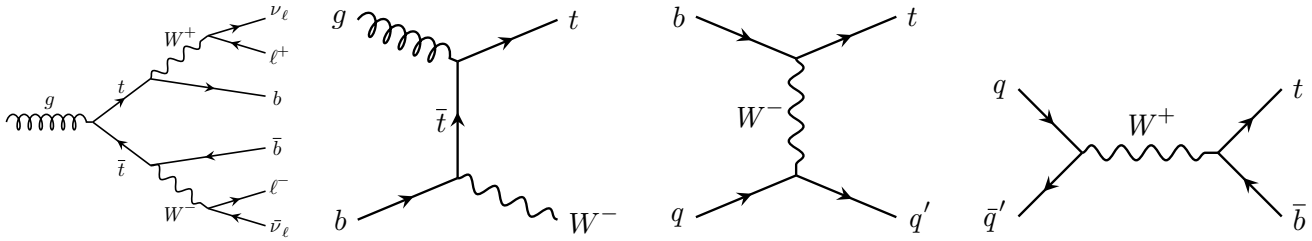


Figure 4.4: Top hadronic quark decay: $t \to Wb \to q\bar{q}b$.

Based on their cross-sections (Figure 4.2) and the final state selected by this study, the possible main sources of contamination for $t\bar{t}$ lepton+jets events are $t\bar{t}$ dilepton events (with one lepton not reconstructed, and additional jets from initial or final-state radiation) and single-top Wt events (with the W boson decaying leptonically, and an extra jet passing the b-tagging requirements), as presented in Figure 4.5a and 4.5b. Other single-top production mechanisms, such as s- and t-channels (Figures 4.5c and 4.5d), were simulated and found to contribute to the selected event yield by less than one per mille, since they either do not produce isolated leptons, or the hadronically-decaying top-quark candidate is reconstructed from a random combination of jets in the event and fails therefore the m_{qq} and m_{qqb} selection requirements described in the Section 4.3; they are therefore neglected. A possibly inaccurate estimation of the background, accounting for neglected contributions from W+jet and multi-jet events, was studied as detailed in Section 4.4.3.



(a) **(b)** **(c)** **(d)**

Figure 4.5: Main sources of background for $t\bar{t}$ lepton+jets events: (a) $t\bar{t}$ dilepton, (b) single top Wt, (c) single top t-channel and (d) single top s-channel.

The idea of the template method is to use the invariant mass distribution of a set of particles, including the one to be calibrated, and assuming that the others are properly calibrated, produced in the decay of a well known resonance with known mass and large production cross-section. The energy scale and resolution of the particle to calibrate are modified in the simulation until the invariant mass distributions in data and simulation agree. In the case of the b-JES calibration presented in this work, the invariant mass m_{qqb} of the b-jet and the two light jets identified as the products of a top quark hadronic decay is used, exploiting the fact that the top quark mass is rather well known (see Figure 4.6). For this measurement the top quark mass and its uncertainty are assumed to be $m_t^{\text{Nominal}} = 172.5 \pm 0.5$ GeV being compatible with the most recent combined

measurement $m_t = 172.44 \pm 0.48\,\text{GeV}$ from CMS [233] (independent of the ATLAS measurement) and $m_t = 172.69 \pm 0.48$ GeV from ATLAS [227]. Recently CMS measured a top quark mass of $m_t = 171.77 \pm 0.38$ GeV with $t\bar{t}$ lepton+jets events [234], however once combining all the CMS measurements, even though the new top mass value measured by CMS will be pulled down a bit, it should remain compatible with $m_t^{\text{Nominal}} = 172.5 \pm 0.5$ GeV.

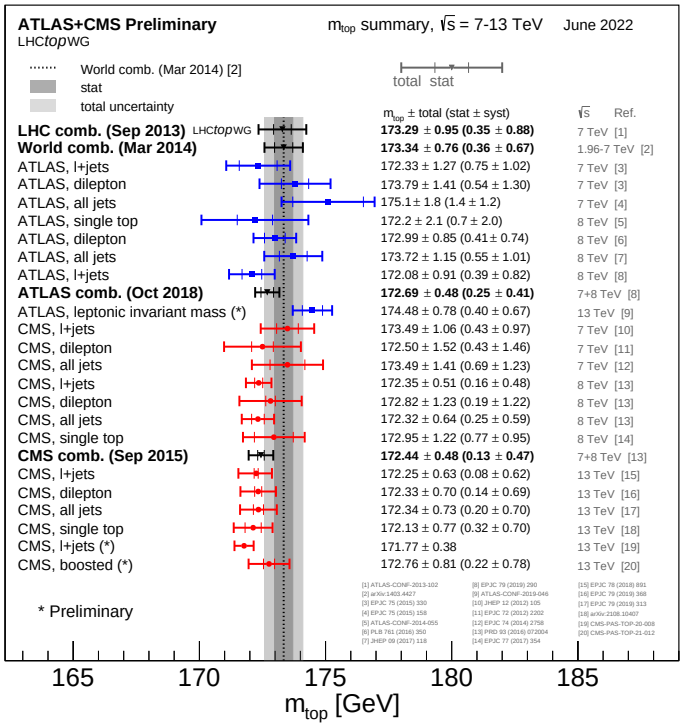


Figure 4.6: Top quark mass measurements (taken from Ref. [235]).

To reduce the impact on the results of the uncertainty in the light jet quark calibration, which might induce uncertainties in the m_{qqb} templates, the difference between m_{qqb} and the invariant mass m_{qq} of the two light quarks (which should be produced by a W boson decay) is used instead. Due to the imperfection in the calibration of the ATLAS detector and calibration procedure, the distribution in data of the reconstructed invariant mass difference $m_{qqb} - m_{qq}$ is different with respect to the distribution obtained with the MC simulation. The MC simulation is then used as a reference to correct the b-jets four-momentum components in data in the following way:

$$E_b^{\text{Data corrected}} = E_b^{\text{Data}}/(1+\alpha), \qquad p_{\text{T}b}^{\text{Data corrected}} = p_{\text{T}b}^{\text{Data}}/(1+\alpha). \tag{4.1.1}$$

The longitudinal component of the jet momentum is indirectly shifted by the same factor since the reconstructed jet direction is assumed not to be affected by the JES, and therefore the pseudorapidity:

$$\eta := -\ln\left[\tan\left(\frac{\theta}{2}\right)\right], \tag{4.1.2}$$

is not modified. Thus, $p_z = p_{\text{T}}/\tan(\theta)$ is rescaled as well when shifting the transverse momentum p_{T} of the b-tagged jet.

In order to determine the energy scale factor α, a chi-squared (χ^2) scan is performed by generating, with the help of the simulation, several $m_{qqb} - m_{qq}$ distributions, corresponding to different energy scale factors α_i (with $i \in [1, n]$)

$$E'^{\text{MC}}_b(\alpha_i) = (1+\alpha_i)E_b^{\text{MC}}, \qquad p'^{\text{MC}}_{\text{T}b}(\alpha_i) = (1+\alpha_i)p_{\text{T}b}^{\text{MC}}. \tag{4.1.3}$$

For each value α_i the chi-squared $\chi^2(\alpha_i)$ between the $m_{qqb} - m_{qq}$ MC and data distributions is calculated. The value α_{min} for which the $\chi^2(\alpha)$ curve has a minimum corresponds to the best

estimate of the energy shift to apply in order for the data and MC template to be statistically closest to each other and is taken as the nominal b-JES of the data. Statistical ($\delta\alpha^{\text{stat.}}$) and systematic ($\delta\alpha^{\text{syst.}}$) uncertainties on the value of α are then computed.

4.2 Data and simulation samples

To perform the b-JES measurement, the full Run 2 pp collision data set, collected at a center-of-mass energy $\sqrt{s} = 13$ TeV between 2015 and 2018, is used. The data were collected using the lowest-threshold unprescaled single-electron and single-muon triggers [236, 237], with minimum p_{T} thresholds varying from 20 to 26 GeV depending on the lepton flavour and the peak instantaneous luminosity during data taking. To increase the trigger efficiency for the muon channel, events are also collected with the missing transverse energy trigger if the reconstructed transverse momentum (p_{T}^{W}) of the leptonically decaying W boson ($t \to Wb \to \mu\nu b$) is greater than 150 GeV. Events are selected for analysis only if they are of good quality and if all the relevant detector components are known to have been in good operating condition [123]. The data correspond to an integrated luminosity $\mathcal{L} = 139\ \text{fb}^{-1}$ after trigger and good data quality requirements [238].

Signal and background templates are generated with the full Run 2 data taking conditions that spanned from 2015 to 2018 [239, 240]. Corrections to match the pileup distribution in data and the lepton and jet reconstruction and identification efficiencies as well as their energy scale and resolution were applied to the simulation.

A combination of nominal and alternative MC samples, produced with either a detailed full simulation (FS) [82] of the ATLAS detector response based on Geant4 [83], or with a faster simulation (AFII) that relies on a parameterization of the calorimeter response [241–244], is used to obtain the nominal results and estimate their systematic uncertainties. For the fast simulation some approximations and simplifications are performed such as relying on parameterized showers in the calorimeter instead of fully simulating the shower development through the ATLAS detector. The samples typically correspond to an equivalent luminosity which is at least ten times larger than that of the data. The choice of the nominal and alternative MC samples used to evaluate the modelling systematic uncertainties followed the recommendations provided by the ATLAS Physics Modelling Group [230] and Top Working Group [245].

Signal samples consist of simulated $t\bar{t}$ lepton+jets events. Background samples consist of simulated $t\bar{t}$ dilepton events and single-top Wt events. Following the Physics Modelling Group and Top Working Group recommendations, the nominal samples are generated at next-to-leading order (NLO) accuracy in QCD with the POWHEG v2 [79, 246–248] interfaced to PYTHIA 8.2 [74] for the modelling of the parton shower and of the underlying event. The diagram removal scheme [249] was used to remove $t\bar{t}$ events from the Wt simulation [250]. The final-state particles of the generated events are passed through the ATLAS detector full simulation. The top mass has been set to $m_t^{\text{Nominal}} = 172.5$ GeV [230, 251], and the h_{damp} parameter[1] controlling initial state radiation (ISR) has been set to $h_{\text{damp}} = \frac{3}{2} m_t = 258.75$ GeV. The NNPDF 3.0 parton distribution function (PDF) set is used [252] for the matrix element generation, while the NNPDF 2.3 PDFs are used for parton showering.

All simulated processes are normalized using the most precise theoretical predictions of their cross-sections. The $t\bar{t}$ samples are normalized using cross-sections calculated at next-to-next-to-leading order (NNLO) accuracy in QCD with all-order resummation of the next-to-next-to-leading logarithms (NNLL) [253], while single-top samples are normalized using cross-sections calculated at next-to-leading order (NLO) in QCD [254, 255].

1. The h_{damp} parameter controls the transverse momentum of the first additional emission beyond the leading-order Feynman diagram in the parton shower and therefore regulates the high-p_{T} emission against which the $t\bar{t}$ system recoils.

In addition to the hard-scattering process, each event is overlaid with additional inelastic pp collisions (pile-up) generated with PYTHIA 8.1 [73] using the ATLAS A3 set of tuned parameters [256] and the NNPDF23LO [257] parton distribution function (PDF) set. Simulated events are then reconstructed with the same algorithms as those applied to data. The simulation is corrected to match the pile-up distribution of the data, as well as the reconstruction and identification efficiencies and energy scale and resolution of leptons and jets measured with data control samples.

The table 4.1 summarizes the samples used. The nominal $t\bar{t}$ sample contains about 470 million lepton+jet and di-lepton $t\bar{t}$ simulated events, collectively denoted as non-fully-hadronic events. A dilepton filter is applied to separate the lepton+jet and di-lepton components and analyze them separately. The latter is then replaced with a dedicated di-lepton $t\bar{t}$ sample containing almost 320 million simulated events. To enhance the fraction of high-p_{T} events, two $p_{\mathrm{T}}(W)$-filtered $t\bar{t}$ non-fully-hadronic samples are also produced for $p_{\mathrm{T}}(W) \in [100, 200)$ GeV, for $p_{\mathrm{T}}(W) \geq 200$ GeV) and combined with the nominal $t\bar{t}$ samples using appropriate weights. Those two $p_{\mathrm{T}}(W)$-extension samples contain respectively 62 and 39 million events simulated. The nominal single-top Wt and $W\bar{t}$ samples contain in total about 80 million events, and at least one of the two W bosons decays leptonically.

For systematic uncertainty studies, the input files are obtained from samples produced with the same nominal generators but passed through the AFII simulation of the ATLAS detector response.

Modelling uncertainties are estimated by comparing the results obtained using the templates built with the nominal simulation samples to those obtained using the templates built with alternative simulation samples or alternative event weights in the nominal simulation samples. The nominal samples include internal alternative event weights that can be used to estimate modelling uncertainties related for instance to PDF uncertainties or missing higher orders in the matrix element calculations. In addition, the following alternative samples have been used:

- Matrix element uncertainty: alternative $t\bar{t}$ and single-top samples have been generated using the matrix-element generator MADGRAPH5_AMC@NLO v2.3.3 [81] interfaced to PYTHIA 8.2 for the parton shower.
- Parton shower uncertainty: alternative $t\bar{t}$ and single-top samples have been generated using the matrix-element generator POWHEG v2 interfaced to HERWIG 7 [77] for the parton shower.
- Top quark mass uncertainty: alternative samples with a generator-level top quark mass of 172.5 ± 0.5 GeV are used, corresponding to the uncertainty on the top quark mass measured by both ATLAS and CMS with Run 1 data [227, 233] as shown in Figure 4.6. Input files for single-top MC samples with variated top mass were not produced since those samples were not available and their effect on the calibration was estimated to be negligible anyway.

Sample informations	Process	Simulated events
POWHEG+PYTHIA 8 FS, Nominal	$t\bar{t}$ non all-hadronic	468 M
	$t\bar{t}$ non all-hadronic $p_T(W) \in [100, 200)$ GeV	62 M
	$t\bar{t}$ non all-hadronic $p_T(W) > 200$ GeV	39 M
	$t\bar{t}$ dilepton	319 M
	Single top Wt	78 M
POWHEG+PYTHIA 8, AFII, Nominal	$t\bar{t}$ non all-hadronic	231 M
	Single top Wt	39 M
POWHEG+PYTHIA 8, AFII, $m_t = 172$ GeV	$t\bar{t}$ lepton+jets	154 M
	$t\bar{t}$ dilepton	75 M
POWHEG+PYTHIA 8, AFII, $m_t = 173$ GeV	$t\bar{t}$ lepton+jets	155 M
	$t\bar{t}$ dilepton	75 M
POWHEG+PYTHIA 8, AFII, ISR Variation Up	$t\bar{t}$ lepton+jets	232 M
	$t\bar{t}$ dilepton	78 M
	Single top Wt	39 M
aMC@NLO+PYTHIA 8, AFII	$t\bar{t}$ lepton+jets	210 M
	$t\bar{t}$ dilepton	70 M
	Single top Wt	35 M
POWHEG+HERWIG 7, AFII	$t\bar{t}$ lepton+jets	191 M
	$t\bar{t}$ dilepton	98 M
	Single top Wt	39 M

Table 4.1: Characteristics of samples.

4.3 Event selection

The object selection (leptons, jets, missing transverse energy) follows closely that of the VH, $H \to b\bar{b}$ analysis described in Chapter 5. The general physics objects selection criteria, identifications, isolations and calibrations in ATLAS are presented in Chapter 3.

The event selection is inspired by the $t\bar{t}$ semi-leptonic event topology illustrated in Figure 4.7 and largely mutuated by the study of the b-tagging efficiency calibration using the same final state [258]. Some additional requirements are introduced as a consequence of the preselection applied in the derivations under study. The following selection criteria are applied:

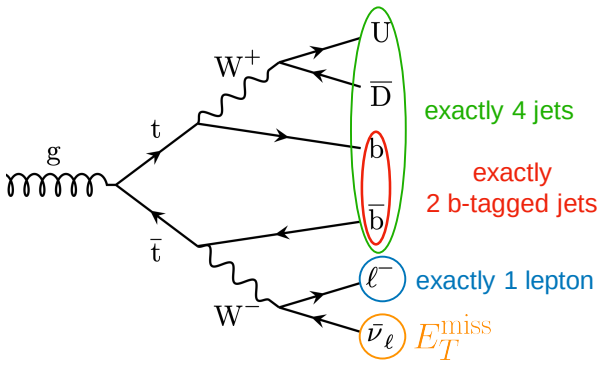


Figure 4.7: $t\bar{t} \to$ lepton+jets events
$\ell = e, \mu$, $U = u, c$ and $D = d, s$.

- **primary vertex**: at least one collision vertex candidate with at least two associated tracks is expected. Collision vertex candidates are identified from charged particle tracks with $p_{\text{T}} > 0.5$ GeV reconstructed in the inner detector [170]. Among the vertex candidates, the one with the largest scalar sum of the squared p_{T} of the associated tracks is considered as the primary (hard-scattering) vertex of the event.
- **Lepton selection**: exactly one *Tight* lepton (electron or muon) matched to the single-lepton trigger and have $p_{\text{T}}^{\text{lepton}} \geq 40$ GeV is required. Electron and muon candidates are required to satisfy impact parameter requirements both along the z direction (impact parameter < 0.5 mm)

and in the transverse plane (significance of the impact parameter < 5 for electrons and < 3 for muons). Events with additional *Loose* leptons are rejected *i.e.* leptons with $p_{\mathrm{T}} > 7$ GeV, passing *LHLoose* (electrons) or *Loose* (muons) identification requirements and *FCLoose* or *FixedCutLoose* isolation requirements respectively.

- Electron candidates are required to satisfy $|\eta| < 2.47$, excluding the small transition region between the barrel and the endcap of the electromagnetic calorimeter, $1.37 < |\eta| < 1.52$. To suppress background from semi-leptonic decays of hadrons or misidentification of other particles, electron candidates should pass an identification requirement using a likelihood-based method exploiting the properties of the reconstructed track and of the energy deposit in the calorimeter, and the quality of their matching. *Tight* electron candidates should have $p_{\mathrm{T}} > 40$ GeV and pass the *LHTight* identification requirements and tight calorimeter-based *FixedCutHighPtCaloOnly* isolation criteria. *Loose* electron candidates are required to have $p_{\mathrm{T}} > 7$ GeV and to pass the *LHLoose* identification criteria and a looser calorimeter-based *FCLoose* isolation requirement.
- Muon candidates are required to satisfy $|\eta| < 2.7$ and to pass identification and isolation requirements. *Tight* muon candidates should have $p_{\mathrm{T}} > 40$ GeV and pass the *Medium* identification requirements and tight track-based *FixedCutHighPtTrackOnly* isolation criteria. *Loose* muons candidates are required to have $p_{\mathrm{T}} > 7$ GeV and to pass the *Loose* identification criteria and a looser track-based *FixedCutLoose* isolation requirement.

➢ **Missing transverse energy and leptonic W boson reconstruction**:

- $E_{\mathrm{T}}^{\mathrm{miss}} \geq 30$ GeV, arising from the neutrino in the final state.
- $m_{\mathrm{T}}^{W} + E_{\mathrm{T}}^{\mathrm{miss}} \geq 60$ GeV. With $m_{\mathrm{T}}^{W} := \sqrt{2p_{\mathrm{T}}^{l} E_{\mathrm{T}}^{\mathrm{miss}} \left[1 - \cos(\Delta\phi(\boldsymbol{l}, \boldsymbol{\nu}))\right]}$ being the transverse mass of the W boson and $\Delta\phi(\boldsymbol{l}, \boldsymbol{\nu})$ the angle between the lepton and the neutrino in the transverse plane. Since electron fakes are more sensitive to $E_{\mathrm{T}}^{\mathrm{miss}}$ and muon fakes are more sensitive to m_{T}^{W}, both types of fake lepton candidates can be suppressed with this single selection requirement.
- $p_{\mathrm{T}}^{W} \geq 75$ GeV. This requirement was inherited from the analysis of the WH, $H \to b\bar{b}$ production [65], for which the simulated samples were produced. The loss in signal efficiency does not induce a decrease in the sensitivity of the measurement, which is limited by the systematic uncertainties.

➢ **Jet selection**: jets are reconstructed using the anti-k_t clustering algorithm [189] (*AntiKt4EMTopoJets*) with a radius parameter $R = 0.4$ from topological three-dimensional noise-suppressed clusters of calorimeter cells calibrated at the electromagnetic scale. Jets are required to have $p_{\mathrm{T}} > 20$ GeV and $|\eta| < 4.5$. Jet cleaning requirements [259] are applied to veto events with jet candidates arising from calorimeter noisy cells. To suppress jet candidates originating from pile-up, a jet-vertex-tagger [219] requirement is applied for jets with $|\eta| < 2.5$ and $p_{\mathrm{T}} < 120$ GeV, while the jet minimum p_{T} requirement is increased to 30 GeV for $|\eta| > 2.5$.

The jet selection is the following:

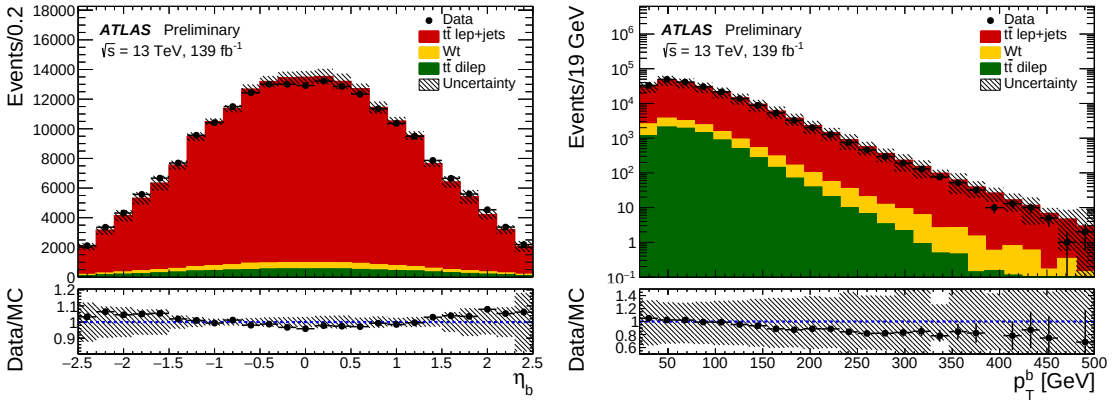
- Exactly 4 jets. In principle semileptonic $t\bar{t}$ events might have additional jets from ISR or FSR events. However, the samples used to produce the input files for the b-JES measurement were produced already including a 4 jets requirement. Moreover, allowing events with more jets can in principle accept also more background events such as hadronic $t\bar{t}$ events with a jet faking a lepton and misreconstructed jet energy leading to fake $E_{\mathrm{T}}^{\mathrm{miss}}$.
- Exactly two b-tagged jets. In the central region ($|\eta| < 2.5$) within the inner-detector acceptance, jets are identified as b-jets (b-tagged) with the MV2c10 tagging algorithm at the 70% efficiency working point with a selection tuned to achieve an average efficiency of 70% for b-jets in simulated $t\bar{t}$ events [207] which has a c-jet (light-jet) rejection respectively of 8.9 (300). Among the two b-tagged jets, the assignment to the hadronically- and leptonically-decaying top quarks is performed in the following way: the b-tagged jet that forms with the

two untagged jets the three-jet system with invariant mass m_{qqb} closer to the top-quark nominal mass are considered the product of the $t \to Wb \to q\bar{q}b$ decay chain, while the remaining b-tagged jet is considered to originate from the decay chain $t \to Wb \to \ell\nu b$ of the other top quark in the event. The b-tagged jets went through the exact same full calibration (shown in Figure 3.14) procedure as for light jets: light jets are fully calibrated. As a result the b-JES correction that is determined by this study is therefore a residual energy scale correction to be applied on top of the default JES calibration.

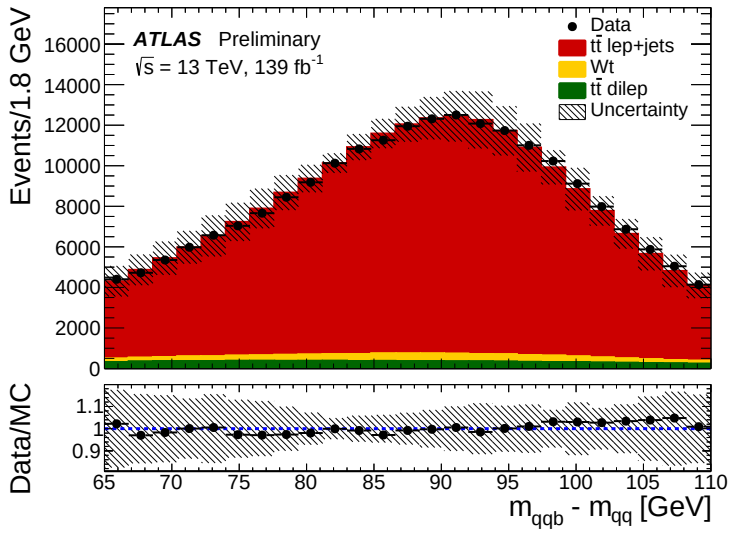
- Invariant mass of the two untagged jets close to the W boson mass, $60\text{GeV} \leq m_{qq} \leq 110\text{GeV}$.
- Invariant mass of the reconstructed hadronic top quark in the interval $130\text{ GeV} \leq m_{qqb} \leq 210\text{ GeV}$.

The number of events selected in data is 212 000 and the expected fractions of signal and background from the simulation are approximately 93% and 7% (4% from $t\bar{t}$ dilepton events and 3% from Wt production). According to the simulation, 97% of the selected b-tagged jets are produced by the hadronization of a b quark.

The Figure 4.8 shows a comparison of the data and the nominal FS POWHEG+PYTHIA 8 MC sample for different distributions. Events are required to pass all selections and to have an invariant mass difference $m_{qqb} - m_{qq} \in [65, 110]\,\text{GeV}$. In other words, all events that contributed to those histograms are taken into account when performing the b-JES scan using the $m_{qqb} - m_{qq}$ distribution. The uncertainty band around the MC prediction includes the theoretical and experimental uncertainty sources described in Section 4.4.3. Data and simulation are in agreement. A small trend as a function of the b-tagged jet kinematic variables and of the invariant mass difference $m_{qqb} - m_{qq}$ is observed, but it is within the uncertainties in the prediction.



(a) (b)



(c)

Figure 4.8: Data-simulation comparison for the following distributions: (a) pseudorapidity (η_b) and (b) transverse momentum (p_{T}^{b}) of the selected b-jet, and (c) $m_{qqb} - m_{qq}$ distributions. Data is compared to the nominal simulated event samples generated with POWHEG+PYTHIA8 and passed through a detailed simulation of the response of the ATLAS detector. Events are required to satisfy all selections and $m_{qqb} - m_{qq} \in [65, 110]$ GeV. The sum of signal and background in the simulation is scaled so that the data and simulation distributions have the same integral. The uncertainty band around the MC prediction includes the theoretical and experimental uncertainty sources described in Section 4.4.3. The MC distributions are scaled to data with the scale factor being extracted from the $m_{qqb} - m_{qq}$ distribution such as the integrals for data and MC for that distribution are equal (see Equation (4.4.2)).

4.4 In situ *b*-JES measurement with a template method

In this section the details of the b-JES measurement are provided, including how the χ^2 curves as a function of α are determined (Section 4.4.1), and how statistical (Section 4.4.2) and systematic (Section 4.4.3) uncertainties are derived.

As a reminder of the technique, several MC template distributions are generated by shifting the four momentum components of the b-tagged jets by a factor $(1+\alpha_i)$ (Equation (4.1.3)), α_i being called the energy scale factor. Shifting the b-tagged four-momentum induces a shift of the $m_{qqb} - m_{qq}$ distribution as shown in Figure 4.9: the position of the peak is moving towards higher masses as the b-jet energy scale factor increases. The $\chi^2(\alpha)$ curve, obtained by computing the different values $\chi^2(\alpha_i)$ between the shifted MC templates and the data, has a minimum denoted α_{min} which corresponds to the correction to apply in data given in Equation (4.1.1).

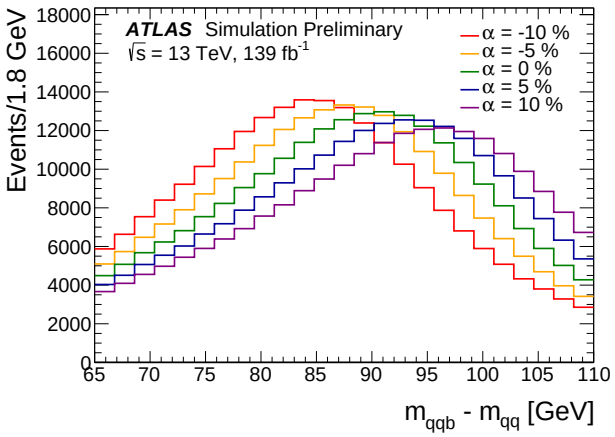


Figure 4.9: Example of templates for $m_{qqb} - m_{qq}$ distribution for $\alpha \in \{-10\%, -5\%, 0\%, +5\%, +10\%\}$ (Nominal FS, not actual step used for the χ^2 scans).

4.4.1 $\chi^2(\alpha)$ curves determination

The $\chi^2(\alpha)$ curves are constructed by building MC templates of the $m_{qqb} - m_{qq}$ variable for different values of α, normalizing those templates to the same integral as the data distribution, and then calculating the χ^2 between the MC and data templates:

$$\chi^2(\alpha_i) = \sum_{k \in \text{Bins}} \frac{(\text{Observed}_k - \text{Normalized Expected}_k(\alpha_i))^2}{\text{Observed}_k} = \sum_{k \in \text{Bins}} \frac{(N_k^{\text{Data}} - N_k'^{\text{MC}}(\alpha_i))^2}{N_k^{\text{Data}}}. \tag{4.4.1}$$

The χ^2 is computed without taking into account underflow and overflow bins. $N_k'^{\text{MC}}(\alpha_i)$ is the normalized content of the k^{th} bin of the MC distribution, obtained from the original content $N_k^{\text{MC}}(\alpha_i)$ of the distribution by:

$$N_k'^{\text{MC}}(\alpha_i) = \text{Factor} \times N_k^{\text{MC}}(\alpha_i) = \frac{\sum_{j \in \text{Bins}} N_j^{\text{Data}}}{\sum_{j \in \text{Bins}} N_j^{\text{MC}}(\alpha_i)} \times N_k^{\text{MC}}(\alpha_i). \tag{4.4.2}$$

The templates are built and the χ^2 is evaluated for various values of α separated by a step of 0.1% over the interval $[-10\%, +10\%]$, hence a total of 201 MC templates are generated for the nominal simulation and for every variation corresponding to a systematic uncertainty source, as described later in Section 4.4.3. The step and bounds of the scan have been chosen to have a relatively precise scan for all variations and such that scans are performed over a large enough interval to exclude the possibility to miss the minimum of the χ^2 curve. For the inclusive measurement 25 bins are used with $m_{qqb} - m_{qq} \in [65, 110]\,\text{GeV}$ as bounds for the distribution. For the differential measurement of the b-JES correction as a function of the b-tagged jet transverse momentum p_{T}^b, the templates and the χ^2 curve are computed separately for events with values of p_{T}^b belonging to different intervals. The templates created from simulated events are normalized to that of data separately in each p_{T}^b interval. The number of bins and bounds of the $m_{qqb} - m_{qq}$ templates for the measurement of the b-jet energy scale factor in each p_{T}^b interval are summarized in Table 4.3.

The χ^2 values obtained at fixed values of α are then interpolated to obtain a continuous curve in the following way. Let $(x_i, y_i)_{i \in [1,n]}$ be a set of n points with $x_1 < x_2 < \cdots < x_{n-1} < x_n$ and f a function that would describe that dataset. An interpolation is (strict definition) a curve f that passes throughout all the points of the dataset (x_i, y_i) *i.e.* such that $f(x_i) = y_i$: an interpolation can be smooth (*e.g.* cubic spline) or not (*e.g.* linear interpolation). A more general set of curves called

regression curves refers to functions that are smooth and close to dataset points $f(x_i) \approx y_i$ while at the same time less sensitive to outliers. Regression curves enable to describe the trends of a dataset and are less dependent to random fluctuations in the dataset (example a fit is a regression). Often the word interpolation refers to regression curves, but in this subsection interpolation will refer to its strict definition.

For smoothing the χ^2 curves, three methods have been considered:

- **Fit:** the problem is some $\chi^2(\alpha)$ curves do not exhibit a close to parabolic shape so fitting with a second order polynomial is not adequate, and it would be difficult to find an analytical function that would describe all χ^2 curves obtained.
- **Cubic splines interpolation:** A spline is a continuous interpolation over the whole dataset with the particularity of being piecewisely defined by polynomials. For a cubic spline, third order polynomials are used, and the curve obtained is $\mathcal{C}^2([x_1, x_n], \mathbb{R})$. A good property of splines is that if a point is removed from the dataset or added to it then the new spline obtained only differs from the original one in the vicinity of the point. So cubic splines are stable curves and in addition they are easy and fast to compute. On the other hand, cubic splines are interpolation curves, so they would induce an overfitting of the data and hence a strong sensitivity to fluctuations in the $\chi^2(\alpha)$ curves.
- **Penalized spline (P-spline):** it is a set of non-parametric regression models which depend only on a smoothing parameter $\lambda \geq 0$. Penalized splines are spline regression curves with penalization of their curvature which enables to reduce sensitivity to outlier points and fluctuations in the dataset. Once a value for the smoothing parameter λ is chosen, a dedicated algorithm finds the spline curve f_λ minimizing the penalized least-squares criterion or penalized residual sum of squares (PRSS) [260, 261]:

$$\text{PRSS}(f_\lambda) = \underbrace{\sum_{i=1}^{n}(y_i - f_\lambda(x_i))^2}_{\text{Least squares}} + \underbrace{\lambda \int_{x_1}^{x_n} \left[f''_\lambda(x)\right]^2 dx}_{\text{Penalize curvature}}. \tag{4.4.3}$$

  - The least squares term: ensures closeness between data and the regression curve: $y_i \approx f_\lambda(x_i)$.
  - The penalizing curvature term: reduces sensitivity to data fluctuations. There are two extreme cases for the smoothing parameter:
    - $\lambda = 0$: there is no restriction on the curvature of f_0, hence f_0 interpolates the dataset (as a cubic spline would do).
    - $\lambda = +\infty$: the curvature becomes impossible so $f''_{+\infty} = 0$ hence $f_{+\infty}$ is a linear regression of the dataset.

  As a result, λ is the trade-off between closeness to dataset and roughness of the curve (see Figure 4.11). For that reason it has been chosen to use a penalized spline to smooth the χ^2 curves (more precisely cubic penalized splines *i.e.* composed by third order polynomials as piecewise function).

  To determine such penalized spline curves the open-source C++ Splinter (SPLine INTERpolation) library [262] has been used. Further details can be found in Appendix A.

  An interesting feature of the penalized spline approach is that a generalization to multivariate functions is possible and already implemented in the Splinter library, thus making it possible in the future to use the template method to perform a two-dimensional simultaneous determination of the b-JES and b-JER corrections.

As explained previously, the smoothing parameter has to be chosen by the user, introducing some arbitrariness (a bit like one would have to choose a function to use before performing a fit). The smoothing parameter is called a hyper-parameter and it can be optimized instead of fixing that

parameter arbitrarily by "hand".

The question is how to determine the "best" smoothing parameter that would enable to prevent the penalized spline curve f_λ from overfitting or underfitting the original χ^2 curve. In the following that optimal smoothing parameter will be denoted $\lambda_{\text{opt.}}$.

A method exists the cross-validated residual sum of squares (CVRSS) (detailed in Appendix A) but was unsuccessful because the penalized curves obtained are overfitting the original χ^2 curves. It consists in minimizing a loss function CVRSS(λ) (Equation (A.1)) that depends on the smoothing parameter: the optimal parameter $\lambda_{\text{opt.}}$ being the value associated to the minimum.

A new and more adequate loss function CV(λ) had to be found. No solutions to this problem were proposed in the literature or after asking on different forums (see Appendix A) leading to an empirical search of a more general loss function. It was decided to focus the searches to the following form:

$$\text{CV}(\lambda) = \sum_{i=1}^{n} \left[\rho_1 \left(y_i - f_\lambda^{(-i)}(x_i) \right) + \rho_2 \left(f_\lambda(x_i) - f_\lambda^{(-i)}(x_i) \right) \right], \tag{4.4.4}$$

where $f_\lambda^{(-i)}$ is the penalized spline obtained when removing the i^{th} point (x_i, y_i) of the dataset. And ρ_1, ρ_2 are two even, positive and convex functions (to be chosen) since the loss function to minimize should be convex [261]. The minimum of the two functions $x \mapsto \rho_{1,2}(x)$ has to be in $x = 0$ and they must tend to infinity for $x \to \pm\infty$: $\lim\limits_{x \to \pm\infty} \rho_{1,2}(x) = +\infty$.

In Equation (4.4.4):

- ρ_1 ensures the closeness of the penalized spline to the data even when removing a point.
- ρ_2 ensures the stability of the penalized spline obtained when removing a point with respect to the original penalized spline. Hence ρ_2 plays a key role because it determines how stable the spline curve is to itself when a dataset point is removed: it ensures the smoothness of the penalized spline. This was the missing term in the CVRSS loss function Equation (A.1) which only compares closeness to the data.

As a result:

- For small values of λ, the spline will drastically change when removing a point: both ρ_1 and ρ_2 will have large values and CV(λ) will be large
- For intermediate values of λ, both ρ_1 and ρ_2 will be small because at the same time the spline is close enough to the data and the curve is smooth enough when removing a point CV(λ) is small.
- For large values of λ, the curve is stable when removing a point so ρ_2 is rather small but the curve does not fit anymore the data so ρ_1 is large hence CV(λ) is large.

As a consequence CV(λ) has a minimum $\lambda_{\text{opt.}}$ which should give a penalized spline curve with expected properties.

It remains therefore to choose the functions ρ_1 and ρ_2. The function ρ_2 must have faster variations than ρ_1 since the spline stability to outliers and its smoothness is more important than closeness to the data. As a result ρ_1 and ρ_2 (see Figure 4.10) are defined thanks to a modified Huber loss function (ρ) as follows:

$$\rho(x, c) = \begin{cases} (x/c)^2 & \text{if } |x| \leq c \\ 2(|x/c| - 0.5) & \text{if } |x| > c \end{cases} \tag{4.4.5}$$

$$c_1 = 1 \quad , \quad c_2 = c_1/20 \quad , \quad \rho_1(x) = \rho(x, c_1) \quad , \quad \rho_2(x) = \rho(x, c_2) \tag{4.4.6}$$

The Huber loss function is commonly used as a loss function [261]. Here ρ is "modified" because it is normalized such that $\rho(c, c) = 1$. The function $x \mapsto \rho(x, c)$ is continuous and its derivative is continuous (even for $x = c$). Basically, c determines the interval for which $x \mapsto \rho(x, c)$ is smaller or equal to 1. For $|x| > c$ then ρ starts to have large values.

For each χ^2 curves, the Migrad algorithm of the Minuit 2 [263, 264] library was used to find the minimum $\lambda_{\text{opt.}}$ of the loss function $\text{CV}(\lambda)$ defined in Equation (4.4.4), with ρ_1 and ρ_2 defined in Equation (4.4.6). This method will be called *optimized penalized spline* or OP-spline method. This empirical loss function has been tested and works both for already smooth and non-smooth original χ^2 curves. Hence, all results derived in Section 4.5 rely on the OP-spline method.

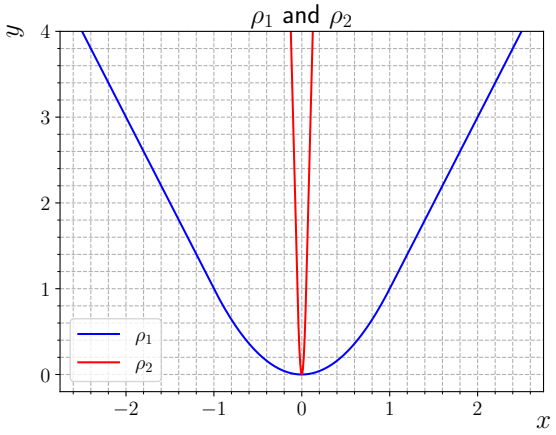


Figure 4.10: Modified Huber loss functions ρ_1 and ρ_2 defined in Equation (4.4.6).

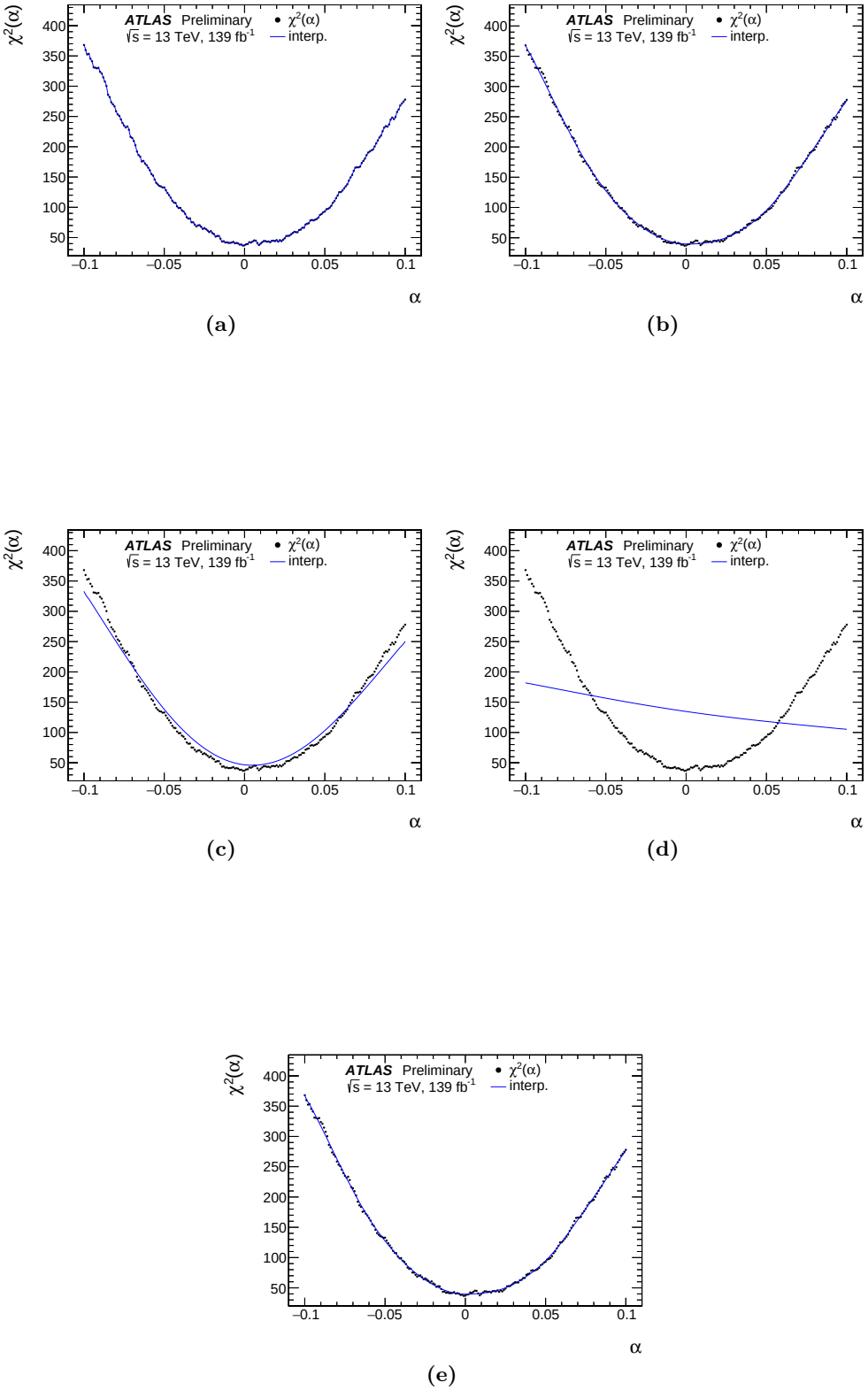


Figure 4.11: Comparison of the original versus smoothed χ^2 curve with penalized spline (P-spline) and different smoothing values(a) $\lambda = 0$, (b) $\lambda = 3000$, (c) $\lambda = 5 \times 10^5$ and (d) $\lambda = 1 \times 10^8$ or (e) optimized penalized spline method (OP-Spline) with the optimal smoothing value found to be $\lambda_{\text{opt.}} = 1584$ for a χ^2 curve associated to the POWHEG+PYTHIA 8 $m_t = 173$ GeV sample for $p_{\text{T}}^b \in [250, 500)$ GeV.

4.4.2 *b*-JES statistical uncertainty

Let's define $\chi^2_{\text{min}} := \chi^2(\alpha_{\text{min}})$, once the value α_{min} that minimizes the $\chi^2(\alpha)$ curve has been determined.

As a consequence of Wilks'theorem, the statistical uncertainty ($\delta\alpha^{\text{stat.}}$) on the value α_{min} can be determined by looking at the point on the left and right of α_{min} such that $\Delta\chi^2 = 1$ *i.e.* such that the χ^2 is increased by 1 with respect to χ^2_{min} as shown in Figure 4.12.

By taking as statistical uncertainty $\delta\alpha^{\text{stat.}} = \max(\delta\alpha_l, \delta\alpha_r)$ one obtains the so called 68.3% confidence-level interval on α (or in other word the $\pm 1\sigma$ uncertainty on α_{min}). The value of the

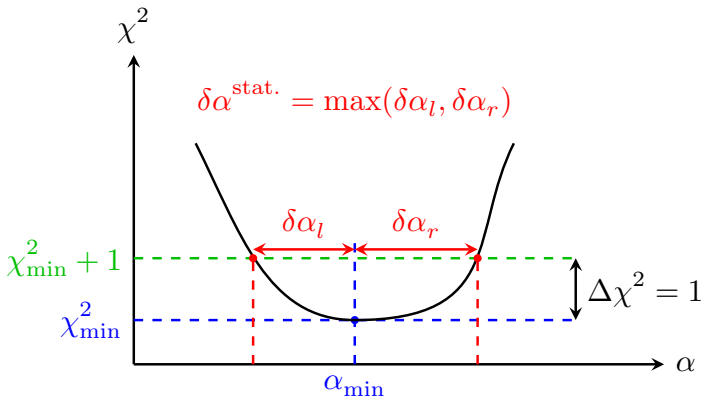


Figure 4.12: Scheme of χ^2 curve and statistical error determination.

χ^2 differences above the minimum ($\Delta\chi^2$) is indeed coming from the fact the χ^2 scan is performed with a unique parameter: the energy scale factor α. So $\chi^2 = \chi^2(\alpha)$ depends on one parameter only. The $\Delta\chi^2$ value to use to determine the 68.3% CL depends on the number of parameters used for the χ^2 scan. In case for instance of a χ^2 scan with two parameters such as the b-JES (α) and b-JER (a) is performed, then one would obtain a two-dimensional paraboloid $\chi^2 = \chi^2(\alpha, a)$ (see Ref. [224] for an example) and the 68.3% CL 2D contour in the (α, a) plane would be the locus of points for which $\Delta\chi^2 = 2.30$.

It can be generalized to several parameters by computing the $\Delta\chi^2$ value needed depending on the wanted confidence interval β and the number n of parameters of the χ^2 function using the following formula:

$$\Delta\chi^2(n) = F^{-1}_{\chi^2_n}(\beta) \tag{4.4.7}$$

Where $F^{-1}_{\chi^2_n}$ is called the chi-square quantile and corresponds to the inverse of the cumulative χ^2 distribution for n degrees of freedom [265, 266].

4.4.3 *b*-JES systematic uncertainties and total uncertainty

Various sources of systematic uncertainties have been taken into account. For most of the sources of uncertainties, positive and negative variations are considered. Others are estimated as two-point systematic uncertainties from the comparison between the nominal and alternative results. In total, 286 variations have been considered, as listed below.

➢ **Experimental systematic uncertainties**. They consist of uncertainties in the detector response that can affect the shape and position of the $m_{qqb} - m_{qq}$ distribution. The main sources of uncertainties are related to the jet reconstruction and selection. Additional uncertainties are related to the b-tagging efficiency. Uncertainties on electrons, muons and $E_{\text{T}}^{\text{miss}}$ are not taken

into account since they do not affect the $m_{qqb} - m_{qq}$ distribution and have a negligible effect on the templates since the MC templates are normalized to the data distribution and therefore uncertainties in the efficiencies do not modify the extracted energy scales.

Concerning the uncertainties related to the jet reconstruction, selection and calibration, the so-called jet category reduction (CR) is used. It is a decomposition of uncertainties into 36 eigenvectors assumed to be uncorrelated [196] which are listed in Table 4.2. For each eigenvector a positive and a negative variation are considered. For each variation, the nominal full-simulation

Uncertainty type	Type	Contributions
Detector NPs	Up/Down	JET_CR_JET_EffectiveNP_Detector1, JET_CR_JET_EffectiveNP_Detector2
Mixed NPs	Up/Down	JET_CR_JET_EffectiveNP_Mixed1, JET_CR_JET_EffectiveNP_Mixed2, JET_CR_JET_EffectiveNP_Mixed3
Modelling NPs	Up/Down	JET_CR_JET_EffectiveNP_Modelling1, JET_CR_JET_EffectiveNP_Modelling2, JET_CR_JET_EffectiveNP_Modelling3, JET_CR_JET_EffectiveNP_Modelling4
Statistical NPs	Up/Down	JET_CR_JET_EffectiveNP_Statistical1, JET_CR_JET_EffectiveNP_Statistical2, JET_CR_JET_EffectiveNP_Statistical3, JET_CR_JET_EffectiveNP_Statistical4, JET_CR_JET_EffectiveNP_Statistical5, JET_CR_JET_EffectiveNP_Statistical6
η-Intercalibration (Modelling, NonClosure and TotalStat)	Up/Down	JET_CR_JET_EtaIntercalibration_Modelling, JET_CR_JET_EtaIntercalibration_NonClosure_highE, JET_CR_JET_EtaIntercalibration_NonClosure_negEta, JET_CR_JET_EtaIntercalibration_NonClosure_posEta, JET_CR_JET_EtaIntercalibration_TotalStat
Flavor uncertainty (Composition and Response)	Up/Down	JET_CR_JET_Flavor_Composition, JET_CR_JET_Flavor_Response
JER	Up/Down	JET_CR_JET_JER_DataVsMC, JET_CR_JET_JER_EffectiveNP_1, JET_CR_JET_JER_EffectiveNP_2, JET_CR_JET_JER_EffectiveNP_3, JET_CR_JET_JER_EffectiveNP_4, JET_CR_JET_JER_EffectiveNP_5, JET_CR_JET_JER_EffectiveNP_6, JET_CR_JET_JER_EffectiveNP_7restTerm
Pile-up (OffsetMu, OffsetNPV, PtTerm and RhoTopology)	Up/Down	JET_CR_JET_Pileup_OffsetMu, JET_CR_JET_Pileup_OffsetNPV, JET_CR_JET_Pileup_PtTerm, JET_CR_JET_Pileup_RhoTopology
PunchThrough	Up/Down	JET_CR_JET_PunchThrough_MC16
SingleParticleHighPt	Up/Down	JET_CR_JET_SingleParticle_HighPt

Table 4.2: List of jet-related sources of systematic uncertainties.

MC samples include alternative event weights or alternative events corresponding to the modified kinematic properties of the various particle candidates, associated to $\pm 1\sigma$ variations of each eigenvector. Such alternative weights or events are used to calculate the alternative $m_{qqb} - m_{qq}$ templates, while the reference ones are those from the full-simulation MC samples with the nominal event weights. The eigenvectors can be summarized as follows [193, 196, 225]:

- **Jet energy scale effective uncertainties**: the multijet p_{T}-balance (MJB) technique is used to calibrate in-situ the jet energy. This calibration uses momentum conservation in dijet (or multijet) events and related uncertainties are derived as a function of η and p_{T} to account for mis-modelling of physics, detector and event topology effects on the momentum balance of the multijet system. Those uncertainties are divided in several components:
  - Detector nuisance parameters: uncertainties related to the response of the detector with respect to the jet energy.
  - Mixed nuisance parameters: uncertainties coming from various modelling sources such as the JES uncertainty due to close-by jets in the recoil system or the jet fragmentation modelling uncertainty.
  - Modelling nuisance parameters: physics modelling of events uncertainty such as the modelling of the angle between the leading jet and the recoil system.
  - Statistical nuisance parameters: statistical uncertainties related to the MJB in-situ calibration.
- **η-intercalibration**: well-measured jets in the central region of the detector ($|\eta_{\mathrm{det}}| < 0.8$) are used to derive a residual calibration for jets in the forward region ($0.8 < |\eta_{\mathrm{det}}| < 4.5$) again with the MJB method.
  - Modelling: physics mismodelling of the envelope of the MC, pile-up and event topology variations.
  - Non-closure: this uncertainty covers the observed non-closure of the calibration method in

the $2.0 < |\eta_{\text{det}}| < 2.6$ region when comparing the response in data with that in MC samples (POWHEG+PYTHIA 8) after applying the derived η-intercalibration.
  - TotalStat: covers the statistical uncertainties of the MC-to-data response ratio.
- **Flavor uncertainties** (flavor response and flavor composition): account for uncertainties in the jet response and simulated jet composition of light-quark, b-quark, and gluon-initiated jets.
- **Jet energy resolution uncertainties** (denoted JER), it includes *DataVsMC* and *EffectiveNP* components. Quantify the uncertainty in the resolution of the jet energy calibration.
- **Pile-up related uncertainties**:
  - μ offset (*OffsetMu*): uncertainty on the modelling of the number of collisions (μ) per bunch crossing in MC simulation.
  - N_{PV} offset (*OffsetNPV*): uncertainty on the modelling of the number of reconstructed primary vertices (N_{PV}) in MC simulation.
  - p_{T} dependence (*PtTerm*): p_{T} dependence uncertainty on the jet variables that are sensitive to pileup.
  - ρ topology (*RhoTopology*). Per event, the pileup contribution is calculated and subtracted using a median p_{T} density (ρ) of jets in the η-ϕ-plane. The ρ topology systematics cover the uncertainty on the per-event p_{T} density (ρ) in MC simulation.
- **Punch-through**: uncertainty in the global sequential calibration (GSC) punch-through correction. The punch-through correction, also called n_{segments} correction, reduces the tails of the response distribution caused by high-p_{T} jets that are not fully contained in the calorimeter. This correction is derived as a function of the jet energy because it is more correlated with the energy escaping the calorimeter.
- **Single-particle response** (*SingleParticleHighPt*): high-p_{T} jet uncertainty on the calorimeter response for single-particle (*i.e.* single isolated particles) and test-beam measurements.

Following the recommendations of the JET/ETmiss group, JER-related systematic uncertainties are applied to all jet candidates, including b-jet ones (true flavor), while JES-related uncertainties are applied to all jets but the b-jets, since the calibration is designed with the purpose of replacing those uncertainties for b-jets (see Figure 4.13). For similar reasons the dedicated

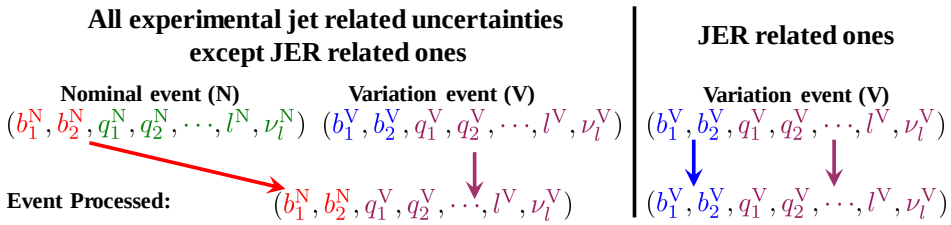


Figure 4.13: Replacement of b-jets for experimental jet systematic uncertainties.

b-JES-response related uncertainties are not included in the list of the experimental systematic uncertainties.

The additional experimental uncertainties that were considered are:

- ***b*-tagging efficiency uncertainties** [267, 268]. Flavour-tagging efficiencies for b-jets and mistag rates for light and c-jets depend on jet momentum and pseudorapidity. A total of 170 variations, corresponding to up and down variations of 85 independent eigenvectors, were considered to estimate those uncertainties (FT_EFF_Eigen: 90 variations for b-jets, 40 for c-jets and 40 for light-jets. No reduction of the number of those NPs has been performed). Those b-tagging efficiencies and mistag rates uncertainties are determined from measurements in data and are propagated to the b-JES measurement. The alternative templates are built using alter-

native event weights corresponding to b-tagging data/MC efficiency scale factors changed by $\pm 1\sigma$.

➢ **Theoretical and modelling uncertainties**. Various sources of uncertainties have been evaluated following the recommendations of References [230, 245]:

- **Initial (ISR) and final (FSR) state radiation of quarks or gluons** (see Figure 4.14). There are 4 variations in total. Alternative templates have been determined using either alternative event weights in the nominal full-simulation samples or alternative AFII samples corresponding to increased or decreased ISR or FSR. For increased ISR (1 variation), the h_{damp} parameter was doubled (from 1.5 to 3 times the top-quark mass) while simultaneously dividing by a factor of two the renormalization and factorization scales, and increasing by 1σ the showering parameter Var3c [250, 269] that corresponds to the value of α_s used for ISR. For decreased ISR (1 variation), the renormalization and factorization scales were multiplied by a factor 2 and the showering parameter Var3c was decreased by 1σ. For increased or decreased FSR (2 variations), the showering parameter Var2 [250, 269], corresponding to the value of α_s used for FSR, was varied by $\pm 1\sigma$.

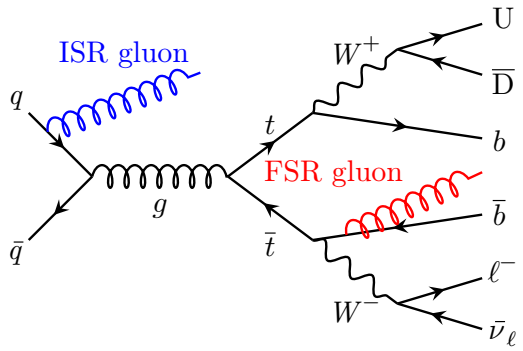


Figure 4.14: Scheme of initial state (ISR) and final state radiations (FSR).

- **Uncertainties on the hard process renormalization μ_R and factorization μ_F scales** (4 variations). Alternative templates have been determined using, in the nominal samples, alternative event weights corresponding to scales $\mu_R = \mu_0/2$ or $\mu_R = 2\mu_0$ (and similarly for μ_F), where μ_0 is the nominal value of the renormalization and factorization scales.
- **PDFs uncertainties** (30 variations). The nominal FS samples include alternative weights corresponding to the PDF set PDF4LHC15_nlo_30 (ID = 90900) and to 30 alternative PDF sets (IDs between 90901 and 90930) [270, 271] representing its uncertainties. Reference templates are built with the nominal PDF4LHC15_nlo_30 set and alternative templates are built using the 30 alternative PDF sets.
- **Uncertainties in the matrix element computations** (1 variation). The b-JES uncertainty was estimated by comparing the values obtained using templates from AFII MC samples produced either with the nominal (POWHEG) or alternative matrix-element generators (MADGRAPH5_AMC@NLO), both interfaced to the same program (PYTHIA 8) for parton showering.
- **Uncertainties in the modelling of the parton shower** (1 variation). The b-JES uncertainty was estimated by comparing the values obtained using templates from AFII MC samples produced the nominal matrix-element generator (POWHEG) interfaced either to the nominal (PYTHIA 8) or to an alternative (HERWIG 7.0) generator for parton shower simulation.
- **Uncertainties in the top quark mass** (2 variations). Reference templates are built with the nominal AFII samples generated with a nominal top quark mass $m_t^{\text{Nominal}} = 172.5$ GeV while alternative templates are built with AFII samples generated assuming a top quark mass of 172.0 or 173.0 GeV.

- **Other background uncertainties** (2 variations). Neglected contributions, mainly from W+jet and multi-jet processes, arising from random combinations of jets in the event, should have a non-peaking $m_{qqb} - m_{qq}$ distribution and a yield similar to or smaller than that of $t\bar{t}$ dilepton events [272]. Therefore, to account for a possible inaccurate estimation of the non-peaking background, the b-JES was measured using alternative templates that were created removing or doubling the contribution from $t\bar{t}$ dilepton events, which have a relatively flat $m_{qqb} - m_{qq}$ distribution in the range under study, as shown in Figure 4.8c. Such uncertainty was found to be small.

The table 4.1 summarizes the characteristics of the samples used for evaluating the different systematic uncertainties.

For each uncorrelated source of uncertainty, alternative $m_{qqb} - m_{qq}$ templates have been produced after varying the corresponding parameter (e.g. the top-quark mass) by $\pm 1\sigma$, where σ is its uncertainty. The new templates are normalized to the data and used to calculate new χ^2 curves $\chi^2_{\pm 1\sigma}(\alpha)$, and the minima of the new curves, $\alpha_{\min,\pm 1\sigma}$, are found. The assigned uncertainty is then:

$$\delta\alpha_{\pm}^{\text{syst.}} = \alpha_{\min,\pm 1\sigma} - \alpha_{\min}. \tag{4.4.8}$$

A few uncertainties, such as the *two-point generator systematic uncertainties* for which there are no $\pm 1\sigma$ variations available, are fully symmetrized. They are estimated by repeating the b-JES measurement using an alternative ensemble of templates obtained with an alternative set of simulations (*e.g.* corresponding to a different parton shower model). Few other uncertainties found to be asymmetric are as well fully symmetrized. In such cases the systematic uncertainty is obtained from the alternative value of the b-JES parameter $\alpha_{\min,\text{alt}}$ and the nominal one:

$$\delta\alpha_{\pm}^{\text{syst.}} = \pm \left| \alpha_{\min,\text{alt}} - \alpha_{\min} \right|. \tag{4.4.9}$$

As all individual systematic uncertainties are assumed to be uncorrelated, the total systematic uncertainties ($\delta\alpha_{\pm}^{\text{tot. syst.}}$) are obtained as the quadratic sum of individual positive and negative systematic errors:

$$\delta\alpha_{+}^{\text{tot. syst.}} = + \sqrt{\sum_{\substack{v \in \text{Variations} \\ \delta\alpha^{\text{syst.}}(v) > 0}} \delta\alpha^{\text{syst.}}(v)^2}, \qquad \delta\alpha_{-}^{\text{tot. syst.}} = - \sqrt{\sum_{\substack{v \in \text{Variations} \\ \delta\alpha^{\text{syst.}}(v) < 0}} \delta\alpha^{\text{syst.}}(v)^2}. \tag{4.4.10}$$

Finally, to obtain the total error the usual quadrature sum is applied:

$$\delta\alpha_{\pm}^{\text{tot.}} = \pm\sqrt{(\delta\alpha^{\text{stat.}})^2 + (\delta\alpha_{\pm}^{\text{tot. syst.}})^2}, \tag{4.4.11}$$

$\delta\alpha^{\text{stat.}}$ being the statistical error measured for the full simulated nominal sample and $\delta\alpha_{\pm}^{\text{tot. syst.}}$ being the total positive and negative systematic errors defined in Equation (4.4.10).

4.5 Results of the *b*-JES measurement

This section contains all b-JES results obtained with the template method.

As the template method has never been used to calibrate b-jets, tests of the procedure have been performed generating pseudo-data, *i.e.* using the nominal full simulated MC sample and shifting the b-jets energy scale with a known value α_{true} to test if one could retrieve it accurately using the template method. These studies are presented in Section 4.5.1. Then, inclusive and p_{T}^b-differential b-JES measurements in data have been performed. They are shown in Sections 4.5.2 and 4.5.3, respectively. All χ^2 curves presented in the pseudo-data test and data measurements are obtained using the optimized penalized spline (OP-spline) regression.

4.5.1 Template method validation with pseudo-data

To test the accuracy of the template method, a shifted MC $m_{qqb} - m_{qq}$ distribution is generated using the nominal POWHEG+PYTHIA 8 sample with a known b-JES value (denoted α_{true}). This α_{true}-shifted $m_{qqb} - m_{qq}$ MC distribution will be called pseudo-data in the following.

For the $m_{qqb} - m_{qq}$ distribution, the histograms used in the χ^2 scan have 25 bins between 65 GeV and 110GeV. The number of bins is reasonably low in order to suppress large statistical fluctuations in each bin, which is of high importance for the template method since it relies on shifting events from one bin to another when changing the energy of the b-tagged jets. The bounds of the $m_{qqb} - m_{qq}$ distribution have been chosen such that the expected value $m_t - m_W$ is approximately in the middle of the distribution and such that the first and last bin of the distribution have roughly the same number of events so that the left and right part of the peak are "balanced".

In order to not bias the test, the value α_{true} has been purposely chosen to be equal to $\alpha_{\text{true}} = 0.55\%$ so that α_{true} is not a point of the $\chi^2(\alpha)$ scan. Indeed, α_{true} is right in the middle of two successive points ($\alpha = 0.5\%$ and $\alpha = 0.6\%$) of the scan since it is performed between $[-10\%, +10\%]$ with a step of 0.1%.

Once the pseudo-data distribution is obtained, statistically equivalent distributions are generated as explained below. In total $N_{\text{seeds}} = 30$ seeds are used, which results in 30 statistically equivalent $m_{qqb} - m_{qq}$ distributions. Examples of the statistically equivalent distributions are shown in Figure 4.15. The generation of the distributions is as follows. Let N_i^{original} be the content of i^{th} bin of the original pseudo-data distribution and $N_i(s)$ be the bin content of the statistically equivalent distribution obtained using a random generator that requires a seed: s. Each bin content N_i^{original} follows a Poisson distribution whose mean value μ is equal to $\mu = N_i^{\text{original}}$. As a result $N_i(s)$ can be computed from N_i^{original} using a random number obtained from the Poisson distribution of mean N_i^{original}:

$$N_i(s) = \text{Poisson}(N_i^{\text{original}}, s). \tag{4.5.1}$$

Since several $m_{qqb} - m_{qq}$ statistically equivalent distributions are generated, there are several χ^2

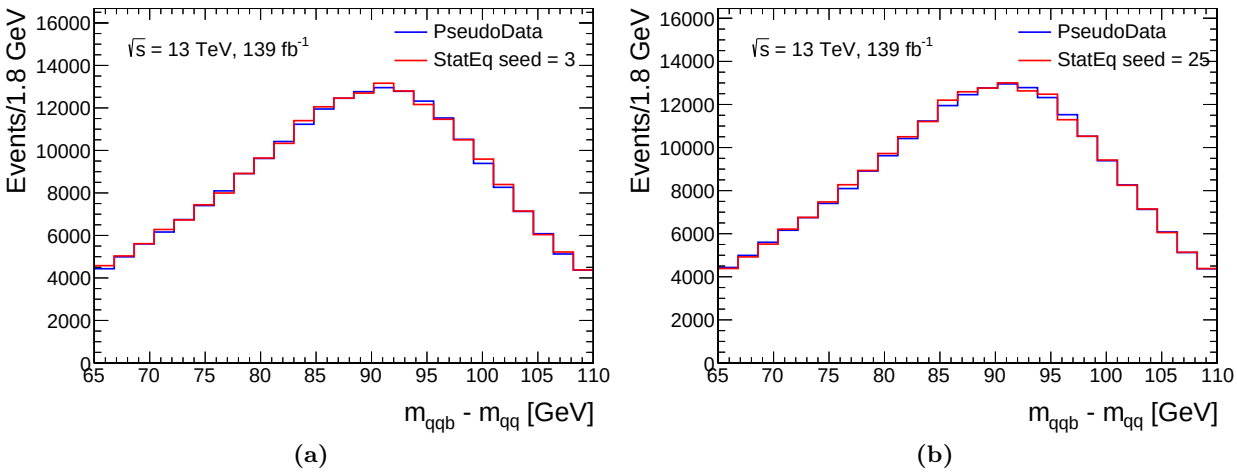


Figure 4.15: Original pseudo-data distribution and two statistically equivalent distributions ($\alpha_{\text{true}} = 0.55\%$) for the (a) seed = 3 and (b) seed = 25.

curves, each one being the comparison between the MC templates and the pseudo-data distribution associated to one of the seed s. The minimum of such χ^2 curve and the statistical uncertainty associated to that minimum are respectively denoted $\alpha_{\text{min}}(s)$ and $\delta\alpha^{\text{stat.}}(s)$.

The average result α_{min} is deduced from the individual values $\alpha_{\text{min}}(s)$ by computing the mean:

$$\alpha_{\text{min}} = \frac{1}{N_{\text{seeds}}} \sum_{s \in \text{Seeds}} \alpha_{\text{min}}(s). \tag{4.5.2}$$

Moreover since all the statistically equivalent pseudo-data distributions are created from the same MC shifted distribution, the values $\alpha_{\text{min}}(s)$ are assumed to be fully correlated in order to not underestimate the statistical uncertainty. With such an assumption, the statistical uncertainty $\delta\alpha^{\text{stat.}}$ on α_{min} is also the mean of the individual $\delta\alpha^{\text{stat.}}(s)$:

$$\delta\alpha^{\text{stat.}} = \frac{1}{N_{\text{seeds}}} \sum_{s \in \text{Seeds}} \delta\alpha^{\text{stat.}}(s). \tag{4.5.3}$$

The equation (4.5.3) comes from the fact that if a variable X is the average of several measurements X_i with uncertainties denoted σ_{X_i} :

$$X := \frac{1}{N} \sum_{i=1}^{N} X_i. \tag{4.5.4}$$

Then, as explained in the section "2.4.3 Linear functions of random variables" in reference [273], the uncertainty (σ_X) on the variable X is given by:

$$\sigma_X^2 = V(X) = \frac{1}{N^2} \sum_{i=1}^{N} \sigma_{X_i}^2 + \frac{2}{N} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \rho(X_i, X_j) \sigma_{X_i} \sigma_{X_j}, \tag{4.5.5}$$

where $\rho(X_i, X_j)$ is the correlation between the measurements X_i and X_j. Assuming full correlation between all measurements (*i.e.* $\rho(X_i, X_j) = 1$), then the equation (4.5.5) can be rewritten as follows:

$$\sigma_X = \frac{1}{N} \sum_{i=1}^{N} \sigma_{X_i}, \tag{4.5.6}$$

which is exactly the formula (4.5.3) where $X = \alpha_{\text{min}}$, $X_i = \alpha_{\text{min}}(i)$ and $\sigma_{X_i} = \delta\alpha^{\text{stat.}}(i)$.

The Figure 4.16 presents the χ^2 curves, the b-JES value found and its statistical and systematic uncertainties for the pseudo-data test. In Figure 4.16b the individual values $\alpha_{\text{min}}(s)$ (blue points) associated to each minimum of the χ^2 curves are plotted with their statistical uncertainty $\delta\alpha^{\text{stat.}}(s)$ (blue error bars). There is also the average value α_{min} (red line) defined in Equation (4.5.2) and its statistical uncertainty $\delta\alpha^{\text{stat.}}$ (gray band) from Equation (4.5.3). Finally, the expected value $\alpha_{\text{true}} = 0.55\%$ (green line) is also shown on that figure. For the Figure 4.16c, the systematic uncertainties are ranked from the highest one to the lowest one, they are grouped and summed in quadrature by categories such as for example PDF set uncertainties. The measured value for this test is $\alpha_{\text{min}} \pm \delta\alpha^{\text{stat.}} \pm \delta\alpha^{\text{syst.}} = 0.57\% \pm 0.08\%^{+2.50\%}_{-2.49\%}$ which is compatible with the expected value $\alpha_{\text{true}} = 0.55\%$. This result shows that the b-JES measurement will be dominated by the systematic uncertainties, with an expected sensitivity around 2.5%. The three leading systematic uncertainties as expected are those related to the theoretical modelling of the parton shower and of the matrix element calculation (POWHEG+HERWIG 7 and AMC@NLO+PYTHIA 8) and to the uncertainty in the top mass as they affect the shape and position of the $m_{qqb} - m_{qq}$ distribution.

It has been checked that even for the leading systematic uncertainties the χ^2 curves are well-behaved. Such χ^2 curves will be shown for the data measurement in the coming sections.

A potential bias from the method was investigated by performing closure tests using pseudo-datasets obtained from the nominal simulation with different value α_{true} ranging from -4% to 4%. For the ten tests performed the difference between the measured and generated values of α was below 0.03%, smaller than the expected statistical precision of the measurement (around 0.08% for the inclusive measurement with a generated value of $\alpha_{\text{true}} = 0.55\%$) and significantly smaller than the expected systematic uncertainty (2.5%).

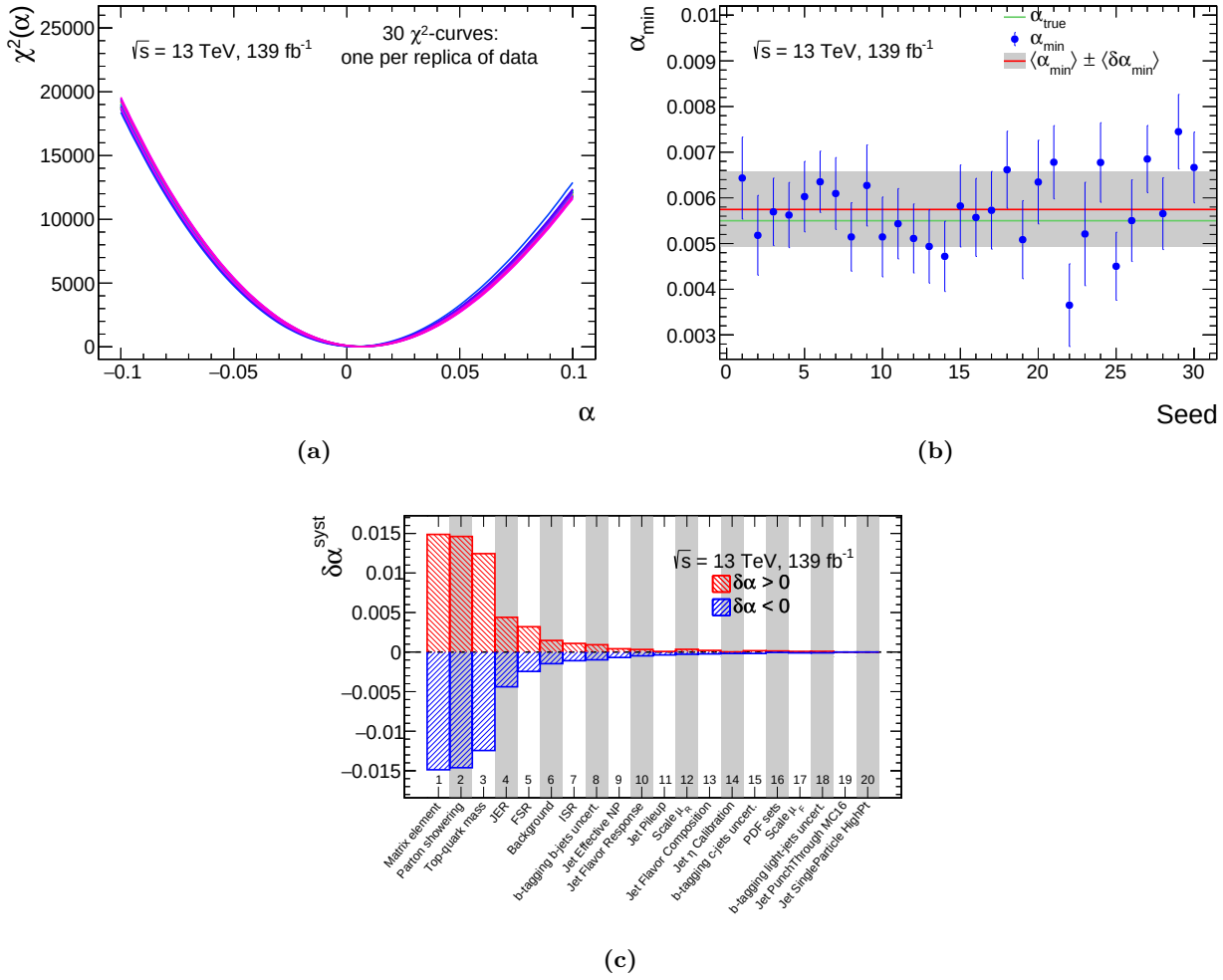


Figure 4.16: Inclusive b-tagged jet energy scale measurement in pseudo-data for $\alpha_{\text{true}} = 0.55\%$: (a) χ^2 curves for the 30 replica of the data (shown in different colors), (b) best-fit value and statistical error $\alpha_{\text{min}} \pm \delta\alpha^{\text{stat.}}$ (blue points with error bars) for each replica, as well as their average (red line with grey error band) and the expected value α_{true} (green line), (c) impact $\delta\alpha^{\text{syst.}}$ of the systematic uncertainties on the measured value of α_{min}.

4.5.2 Inclusive measurement in data

For the inclusive b-JES measurement in data, the same binning and the same distribution bounds are used as for the pseudo-data test presented in Section 4.5.1.

Here the data distribution is computed using the nominal b-jet energy. As for the pseudo-data test, statistically equivalent distributions to data are generated and α_{min} and $\delta\alpha^{\text{stat.}}$ are obtained by averaging over the results corresponding to the individual seeds as described in the previous section.

The reason to use several statistically equivalent distributions even for a measurement in data is because, as explained in Equations (4.4.1) and (4.4.2), the MC templates are normalized to data when computing the χ^2 values. It enables to only take into account differences in shapes and positions between the data and MC templates. However, there is a statistical uncertainty on that normalization and unwanted fluctuation in bins of the data distribution could bias the results. In order to increase the robustness with respect to that normalization uncertainty and the fluctuations in the bins, α_{min} and $\delta\alpha^{\text{stat.}}$ are obtained by averaging the statistically equivalent distributions results as described in Equations (4.5.2) and (4.5.3).

The results of the b-JES inclusive measurement in data are presented in Figure 4.17. The energy

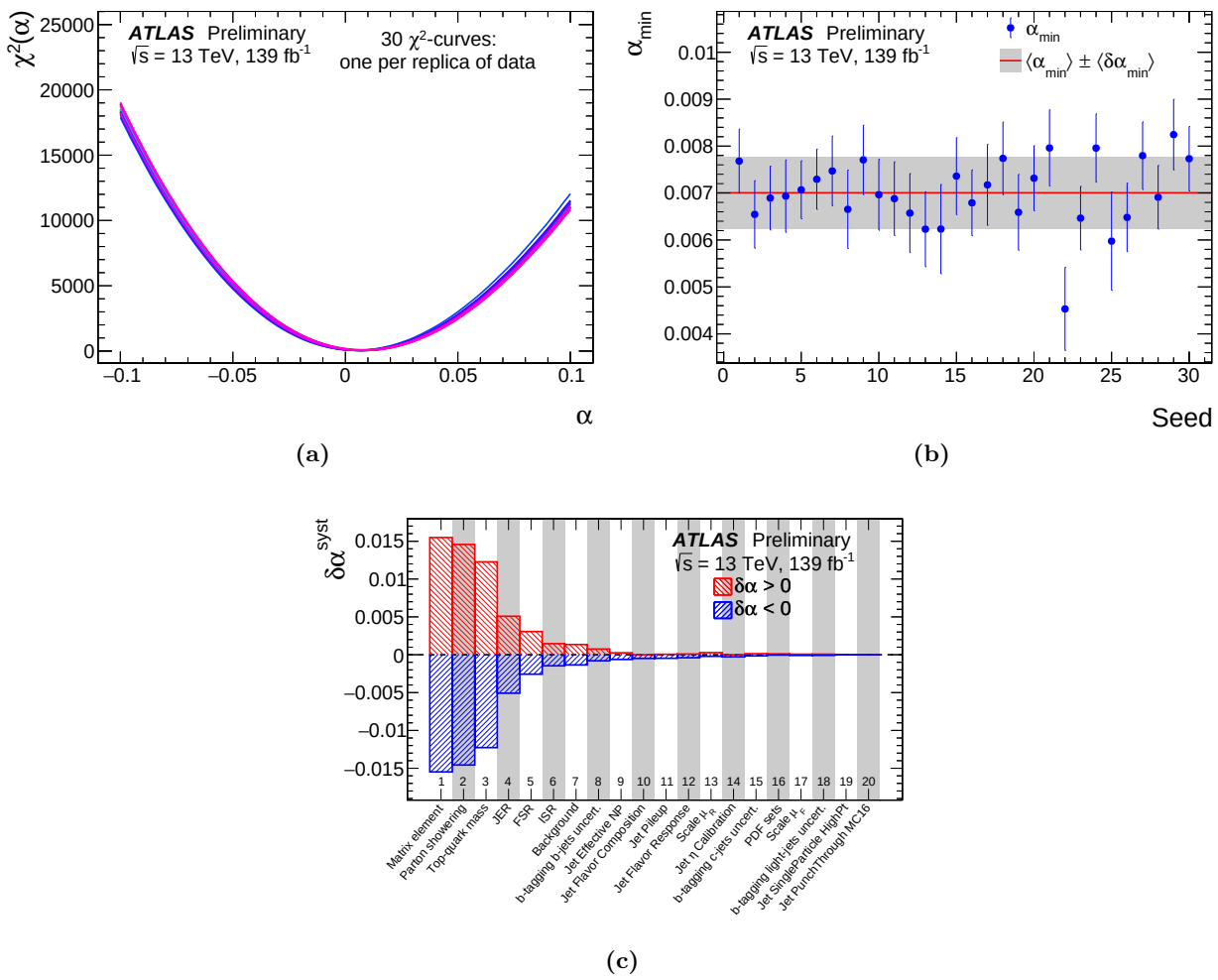


Figure 4.17: Inclusive b-tagged jet energy scale measurement in data: (a) χ^2 curves for the 30 replica of the data (shown in different colors), (b) best-fit value and statistical error $\alpha_{\text{min}} \pm \delta\alpha^{\text{stat.}}$ (blue points with error bars) for each replica, as well as their average (red line with grey error band), (c) impact $\delta\alpha^{\text{syst.}}$ of the systematic uncertainties on the measured value of α_{min}.

scale factor measured in data is $\alpha_{\text{min}} \pm \delta\alpha^{\text{stat.}} \pm \delta\alpha^{\text{syst.}} = 0.70\% \pm 0.08\%^{+2.54\%}_{-2.53\%}$.

The first thing to notice is that the statistical and systematic uncertainties are close to the expected ones obtained with the pseudo-data test. Moreover, the ranking of the systematic uncertainties is similar to that obtained with pseudo-data, with the two modelling uncertainties and the top mass uncertainty being the three leading ones.

Referring to Equation (4.1.1), a value of $\alpha_{\text{min}} > 0$ implies that the measured energy and transverse momentum of b-tagged jet in data are higher than predicted by the simulation. Therefore, the b-JES correction in data will decrease those quantities. Equation (4.1.1) can be written in a simpler way:

$$E_b^{\text{Data corrected}} = (1 + \alpha_{\text{correction}})E_b^{\text{Data}} \qquad p_{\text{T}b}^{\text{Data corrected}} = (1 + \alpha_{\text{correction}})p_{\text{T}b}^{\text{Data}}, \tag{4.5.7}$$

with $\alpha_{\text{correction}}$ being linked to α_{min} by identification in Equations (4.1.1) and (4.5.7):

$$\alpha_{\text{correction}} = \frac{1}{1 + \alpha_{\text{min}}} - 1. \tag{4.5.8}$$

The Figure 4.18 shows a comparison of the $m_{qqb} - m_{qq}$ distribution in data versus the nominal FS POWHEG+PYTHIA 8 before and after the b-JES correction is applied to data. Before correction in Figure 4.18a, a lack (respectively an excess) of events in data at low (respectively at high) $m_{qqb} - m_{qq}$ values is observed though this difference is compatible within uncertainties. Hence, to correct the data the b-JES correction needs to shift events towards lower values which is coherent with a positive value of α_{min}.

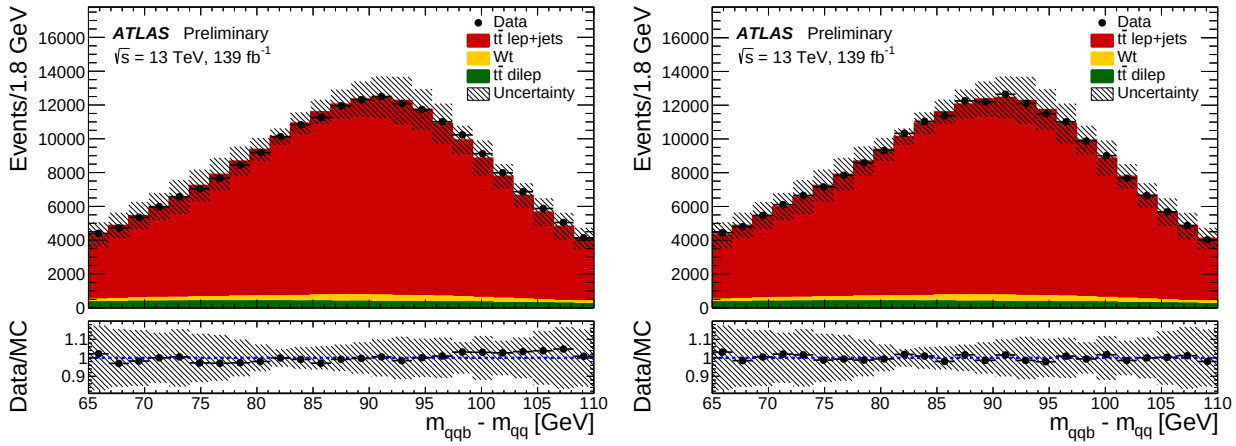


Figure 4.18: Distribution of $m_{qqb} - m_{qq}$ in data (black dots with error bars) compared to the nominal simulated event samples generated with POWHEG+PYTHIA 8 and passed through a detailed simulation of the response of the ATLAS detector, (a) before and (b) after applying b-JES correction to the data. The uncertainty band around the MC prediction includes the theoretical and experimental uncertainty sources described in Section 4.4.2.

Figure 4.19 shows the measured b-JES between the data and the nominal full or the fast POWHEG+PYTHIA 8 simulations.

Using templates built with the nominal FS POWHEG+PYTHIA 8 gives $\alpha_{\text{min}}(\text{FS}) \pm \delta\alpha^{\text{stat.}} = 0.70\% \pm 0.08\%$, while with the templates built from the AFII POWHEG+PYTHIA 8 samples $\alpha_{\text{min}}(\text{AFII}) \pm \delta\alpha^{\text{stat.}} = -0.17\% \pm 0.08\%$. This difference was also observed in pseudo-data, and is the result of differences for the nominal $m_{qqb} - m_{qq}$ distributions between the full and fast detector simulations. Looking at the value of the minimum χ^2_{min} of each individual χ^2 curve, the nominal FS sample has better agreement with data ($\chi^2_{\text{reduced}}(\text{FS}) \approx 2.0$) than the AFII nominal sample ($\chi^2_{\text{reduced}}(\text{AFII}) \approx 2.2$), as expected from a more realistic simulation of the detector response. It should be noted that the systematic uncertainty on α_{min} largely covers this difference.

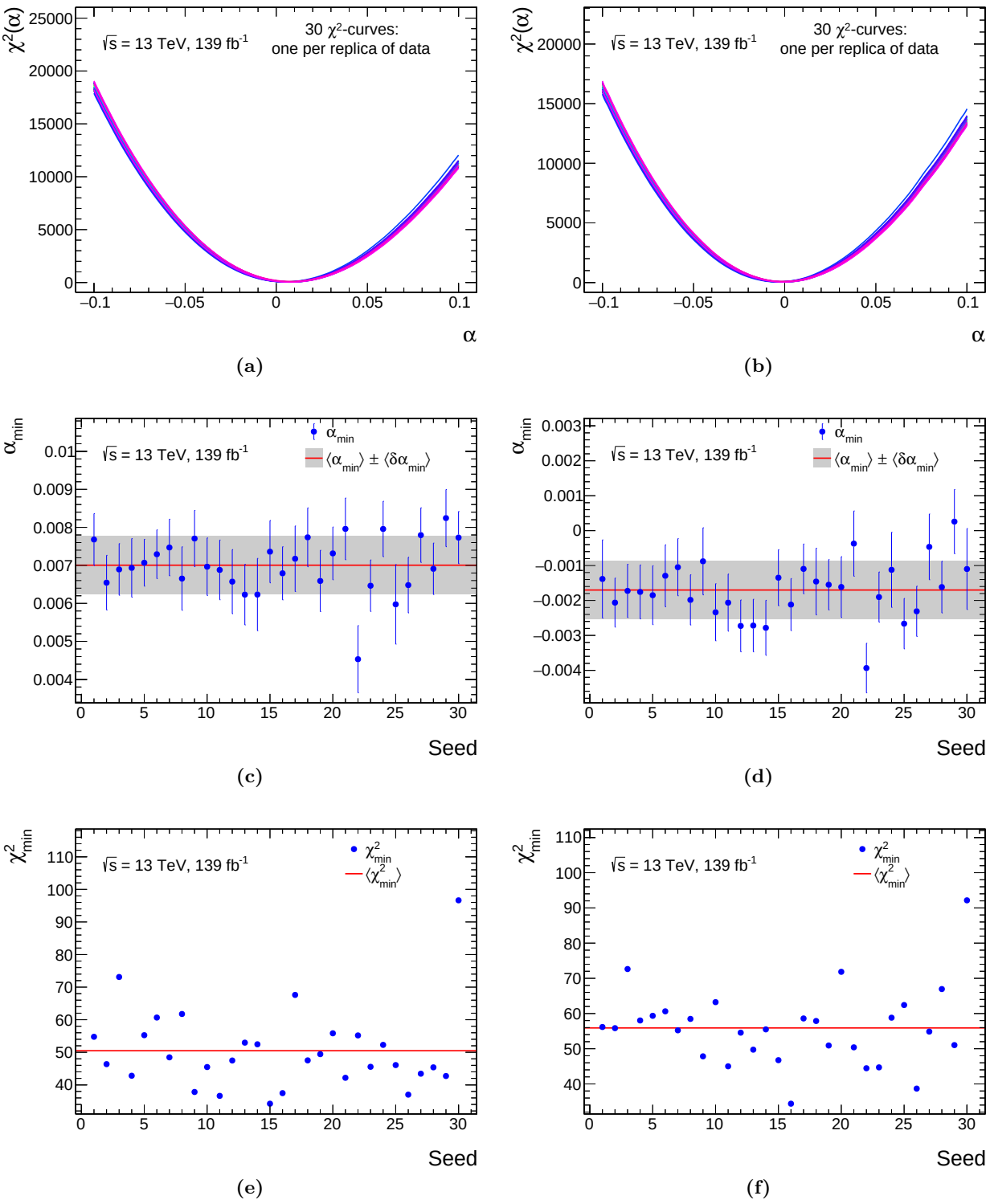


Figure 4.19: Comparison of the nominal (POWHEG+PYTHIA 8) (a), (c) and (e) full simulated versus (b), (d) and (f) fast simulated results: (a) and (b) χ^2 curves, (c) and (d) α_{min} and $\delta\alpha^{\text{stat.}}$ and (e) and (f) χ^2_{min} values.

4.5.3 Differential p_{T} measurement in data

The same procedure described in the previous Sections has been applied to perform a differential b-JES measurement. Since the size of the selected sample is not enough for a two-dimensional differential measurement of the b-JES factor $\alpha(p_{\mathrm{T}}^{b}, \eta^{b})$ as a function of the transverse momentum (p_{T}^{b}) and pseudorapidity (η^{b}) of the b-tagged jet, a simpler one-dimensional differential measurement $\alpha(p_{\mathrm{T}}^{b})$ as a function only of p_{T}^{b} was performed.

The range $p_{\mathrm{T}}^{b} \in [20\ \mathrm{GeV}, 30\ \mathrm{GeV})$ was excluded from the measurement because for those values of transverse momentum the $m_{qqb} - m_{qq}$ distribution is not narrow enough and tends to be "flat" resulting in the template method not being precise enough to have reasonable uncertainties. Moreover, it was tested that for $p_{\mathrm{T}}^{b} > 500$ GeV no $t\bar{t}$ lepton+jets events would pass the selection. As a result the phase space was divided in p_{T}^{b}-bins covering the range between 30 GeV and 500 GeV.

As p_{T}^{b} increases, it was observed that the position of the $m_{qqb} - m_{qq}$ peak is shifted towards higher values both in data and MC. A reason that could explain such behaviour is that at low p_{T}^{b}, some particles emitted inside the b-jet are not contained in the $R = 0.4$ cone around the jet axis so not all energy deposits are taken into account for the reconstructed jet which would lead to b-jet having a lower reconstructed energy and transverse momentum than the original b-hadron that initiated the jet. Hence, the reconstructed m_{qqb} mass would be lower than it should be. On the contrary as p_{T}^{b} increases, emitted particles inside the b-jet tend to be more and more collimated resulting in the reconstructed m_{qqb} mass to have higher values than at low p_{T}^{b}. As a result as p_{T}^{b} increases it would cause the reconstructed $m_{qqb} - m_{qq}$ value to increase as well.

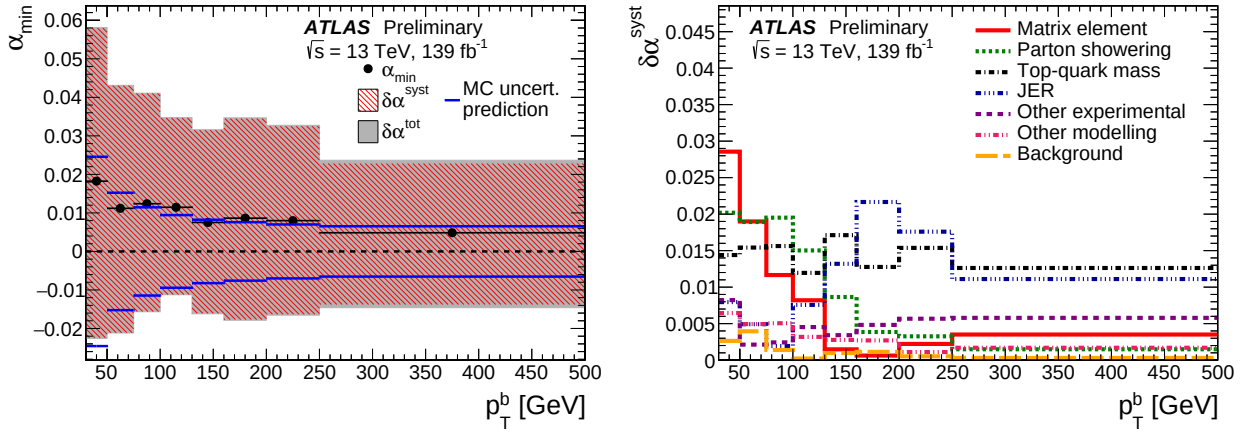
In addition, as p_{T}^{b} increases the number of selected event decreases rapidly. To tackle this reduction in statistics and shift of the $m_{qqb} - m_{qq}$ distribution it was decided to use a variable number of bins and different bounds for the $m_{qqb} - m_{qq}$ histograms. The Table 4.3 details the bounds and number of bins used depending on the p_{T}^{b}-region under study. The phase space was divided in eight p_{T}^{b}-regions:

p_{T}^{b} regions	Number of bins for the $m_{qqb} - m_{qq}$ distribution	Bounds for the $m_{qqb} - m_{qq}$ distribution
$[30, 50)$ GeV	25	$[65, 110]$ GeV
$[50, 75)$ GeV	25	$[65, 110]$ GeV
$[75, 100)$ GeV	20	$[70, 110]$ GeV
$[100, 130)$ GeV	20	$[75, 110]$ GeV
$[130, 160)$ GeV	20	$[75, 110]$ GeV
$[160, 200)$ GeV	20	$[80, 110]$ GeV
$[200, 250)$ GeV	15	$[80, 110]$ GeV
$[250, 500)$ GeV	10	$[80, 110]$ GeV

Table 4.3: $m_{qqb} - m_{qq}$ distribution: number of bins and bounds depending on the p_{T}^{b}-region.

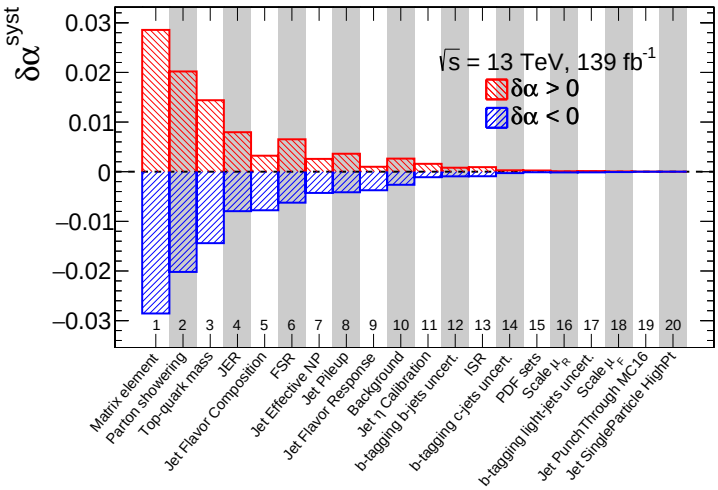
The results obtained in the different p_{T}^{b}-regions are shown in Figures 4.20 and 4.21. In particular the Figure 4.20a summarizes the measurement in data of $\alpha_{\mathrm{min}}(p_{\mathrm{T}}^{b})$ with the total systematic uncertainty and the total uncertainty. The results in this figure show values of the b-JES in agreement with the inclusive value of the order of 0.7%, with a trend that seems to imply a decrease of the b-JES with p_{T}, from around 1.8% to 0.5%. The total uncertainty also decreases as a function of p_{T}^{b} from 4% to 2%. A comparison to the currently assigned b-JES uncertainty prediction derived from MC simulation in ATLAS is shown. In order to reduce the b-JES measurement uncertainty, one would need improvement for the theoretical uncertainties concerning the modelling of events (especially parton showering and matrix element computation) and as well an increased precision

on the top-quark mass measurement. It is worth mentioning the matrix element uncertainty is quite conservative. The samples are known to overestimate that uncertainty [230] but they were recommended by the Physics Modelling group to be used for the ATLAS Run 2 results. Moreover, there are possible circular dependencies taking into account both top mass uncertainties and $t\bar{t}$ process modelling uncertainties that could result in some overestimation. That interplay could be improved in the future. The Figure 4.20b presents the contribution to the b-JES uncertainty, as a function of p_{T}^{b}, of the various sources of systematic uncertainties, grouped in seven categories: matrix element, parton showering, top-quark mass, other modelling uncertainties (ISR and FSR, hard process renormalization and factorization scales, PDFs), background normalization, JER uncertainties and the other experimental uncertainties. The matrix element and parton showering uncertainties dominate at low p_{T}^{b} ($\approx 2-3\%$) and then rapidly decrease with increasing p_{T}^{b}, while the JER uncertainties follow the opposite trend. The uncertainty from top-quark mass is approximately constant around 1.5%. The other modelling uncertainties, the other experimental uncertainties and the background uncertainty are small (below 0.8%) everywhere. Finally, Figure 4.22 demonstrates the impact of the b-JES correction once applied on data. Obviously as the correction is less important in high p_{T}^{b}-regions, the improvement is less visible but in general an improvement is obtained especially in low p_{T}^{b}-regions.

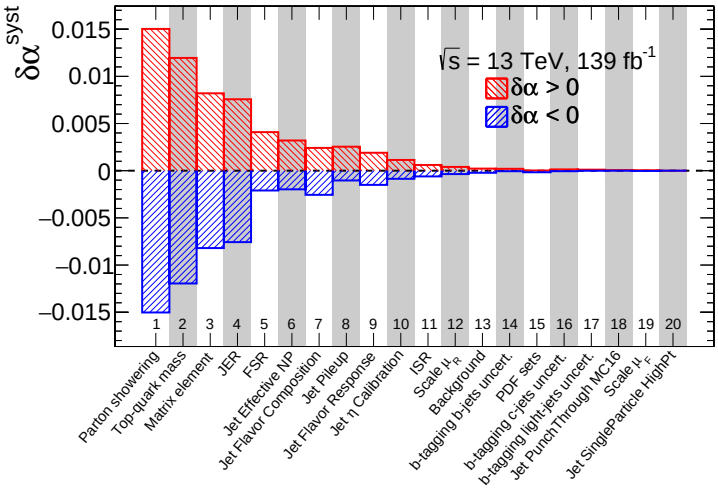


(a) **(b)**

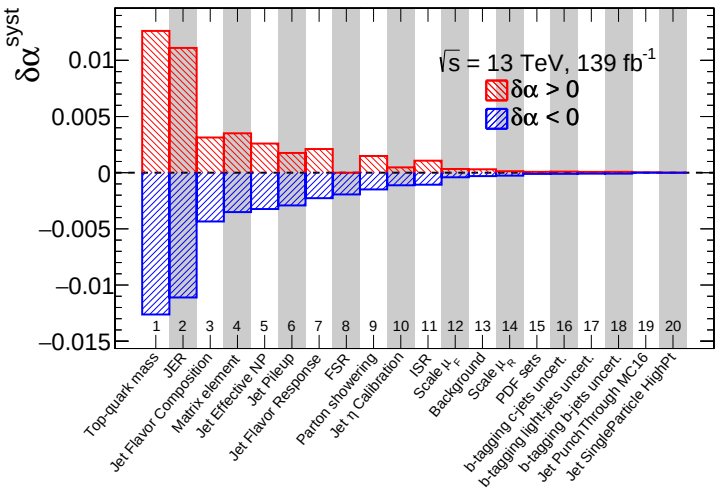
Figure 4.20: (a) Measured b-tagged jet energy scale as a function of the jet transverse momentum. The black dots correspond to the best-fit value of the b-JES while the hatched (solid) red (gray) area corresponds to the systematic (total) uncertainty. The blue histogram shows the sum in quadrature of the b-jet specific uncertainties derived from the comparison of the response in different generators, and all other jet energy scale uncertainties including those from the *in situ* calibration of light jets. (b) Contribution to the b-JES uncertainty as a function of p_{T}^{b} from the systematic uncertainties, grouped in seven categories.



(a) $p_{\mathrm{T}}^{b} \in [30, 50)$ GeV

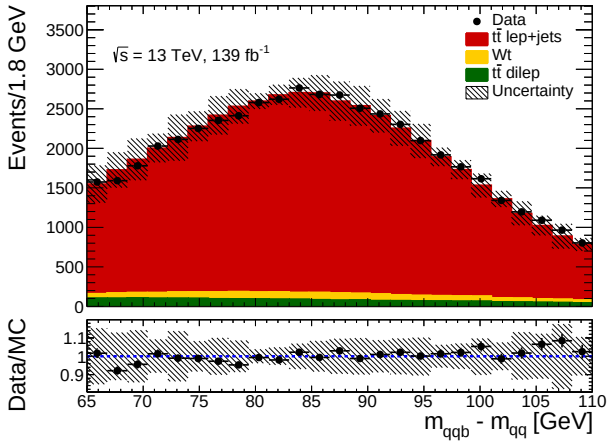


(b) $p_{\mathrm{T}}^{b} \in [100, 130)$ GeV

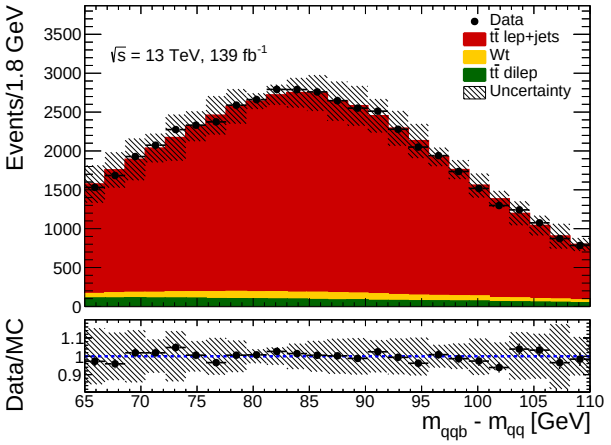


(c) $p_{\mathrm{T}}^{b} \in [250, 500)$ GeV

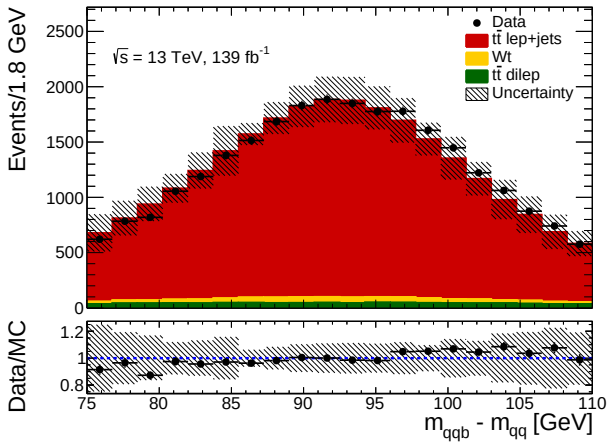
Figure 4.21: Impact $\delta\alpha^{\mathrm{syst.}}$ of the systematic uncertainties on the measured value of α_{min} for the differential p_{T}^{b}-measurement in data for 3 p_{T}^{b}-regions: (a) $[30, 50)$ GeV, (b) $[100, 130)$ GeV and (c) $[250, 500)$ GeV.



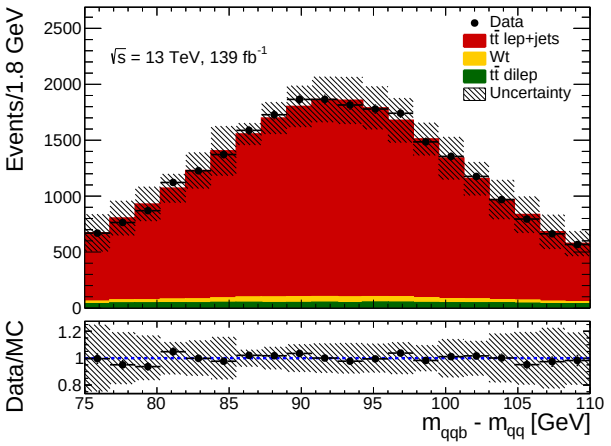
(a) Without b-JES correction for $p_{\mathrm{T}}^{b} \in [30, 50)$ GeV



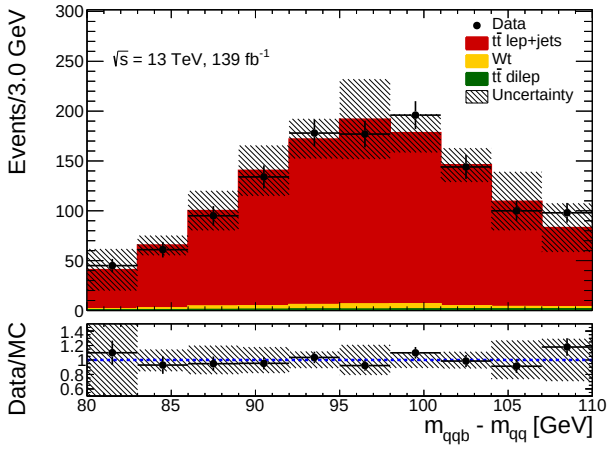
(b) With b-JES correction for $p_{\mathrm{T}}^{b} \in [30, 50)$ GeV



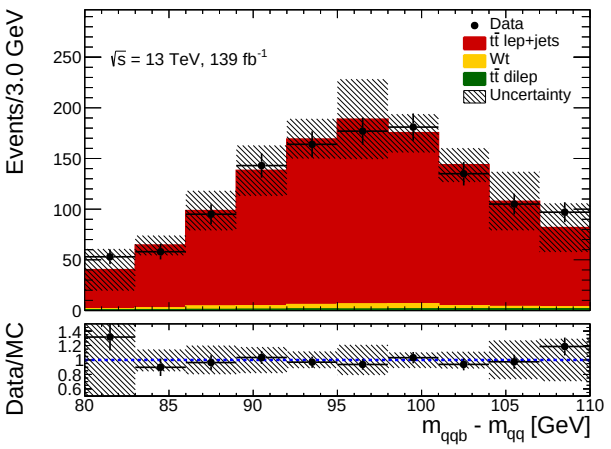
(c) Without b-JES correction for $p_{\mathrm{T}}^{b} \in [100, 130)$ GeV



(d) With b-JES correction for $p_{\mathrm{T}}^{b} \in [100, 130)$ GeV



(e) Without b-JES correction for $p_{\mathrm{T}}^{b} \in [250, 500)$ GeV



(f) With b-JES correction for $p_{\mathrm{T}}^{b} \in [250, 500)$ GeV

Figure 4.22: Data-nominal simulation (full simulated POWHEG+PYTHIA 8) comparison for the $m_{qqb} - m_{qq}$ distributions before and after applying b-JES correction to the data for 3 p_{T}^{b}-regions: $[30, 50)$, $[100, 130)$ and $[250, 500)$ GeV.

4.6 Conclusion

The b-JES is an *in situ* correction that shifts the transverse momentum and the energy of b-jets in data in order to compensate the imperfect calibration of the ATLAS detector and the full jet calibration procedure.

This chapter presents a novel method in ATLAS to determine this correction with lepton+jets events from $t\bar{t} \to \ell\nu q\bar{q}b\bar{q}$ by reconstructing the invariant mass difference $m_{qqb} - m_{qq}$ of the top hadronic decay ($t \to Wb \to qqb$). Templates of this variable for different values of the b-JES are built using MC simulations, and χ^2 curves between the $m_{qqb} - m_{qq}$ data distribution and the MC templates ones are then calculated. A smoothing technique of the χ^2 curves relying on penalized splines is used and results are averaged over individual χ^2 curves obtained from the comparison of MC templates and statistically equivalent distributions generated from the original data distribution. This approach enables reducing the sensitivity of the results with respect to statistical fluctuations in the original distribution.

The results are obtained for jets reconstructed with the anti-k_t clustering algorithm (AntiKt4EMTopoJets) with a radius parameter $R = 0.4$ from topological three-dimensional noise-suppressed clusters of calorimeter cells.

Prior to the measurement in data, tests using pseudo-data have been carried out to check that the method yields unbiased and accurate results.

The inclusive measurement in data found the b-JES correction is of the order of 0.7% with a total uncertainty of the order of 2.5%. A one-dimensional differential measurement of the b-JES correction with respect to the jet transverse momentum was also performed, in 8 bins of variable width between 30 and 500 GeV. The measurement is consistent with the inclusive measurement and seems to indicate a decrease of the b-JES with p_{T}.

After the b-JES correction is applied in data, the agreement between $m_{qqb} - m_{qq}$ corrected data distribution and the nominal fully simulated MC prediction is improved, as expected.

For the future, the b-JES measurement presented in this chapter could benefit from several improvements and developments.

The study could be repeated with the latest ATLAS recommendations in terms of b-tagging and jet reconstruction algorithms using the DL1r tagger and PFlow jets instead of the MV2c10 tagger and EMTopo jets. Thanks to the higher c- and light-jet rejections of the DL1r tagger, the event selection would be even purer in $t\bar{t}$ lepton+jets events. Moreover, the PFlow jets have several improvements compared to EMTopo jets, in particular in terms of energy resolution which should reflect on the $m_{qqb} - m_{qq}$ template distributions used in this measurement by reducing the peak width hence increasing the precision on the energy scale factor measured.

The measurement could also be performed in 2-dimension as a function of the transverse momentum and the pseudorapidity of b-jets dividing phase space with a decreasing number of bins and increasing bin width in high p_{T}- and high $|\eta|$-regions.

The measurement relies on the top quark mass and associated top quark mass uncertainties. It is one of the leading uncertainty in all p_{T}^{b}-regions, instead a simultaneous b-JES and top quark mass measurement, along the lines of the two following studies [226, 227], could be performed in order to discard top quark mass uncertainties.

The matrix element uncertainties are known to have been overestimated due to a quite conservative approach that was recommended for the results of Run 2. This uncertainty could be reduced in the future. There is also a possible interplay between top quark uncertainties and matrix element uncertainties in the sample used.

Another possibility would be to apply the b-JES factor to simulation in order to match data instead of applying the correction to data. In that case, the b-JES values derived for each type of simulation (parton showering, matrix element computation uncertainty...) could be considered as MC-to-MC corrections rather than uncertainties.

Finally, such b-JES measurement could be, one day, potentially performed again with the latest ATLAS recommendations at that time and integrated in the ATLAS workflow. It would be of interest for instance for the $VH, H \rightarrow b\bar{b}$ analysis (described in Chapter 5) as it would reduce discrepancies between data and simulations, and the b-JES uncertainties would be obtained from data rather than estimated in simulations.

Chapter 5

Higgs decay to a pair of b- or c-quarks in association with a vector boson

5.1 Introduction to $VH \to b\bar{b}/c\bar{c}$ events and Higgs decay topologies . . . 157
5.2 Data and simulation samples . . . 161
5.3 $VH, H \to b\bar{b}/c\bar{c}$ event reconstructions and selections . . . 165
5.3.1 Overview of the $VH, H \to b\bar{b}/c\bar{c}$ event combination . . . 169
5.3.2 Common selections for the resolved and boosted topologies . . . 173
5.3.3 Specific selections for the resolved topology . . . 175
5.3.4 Specific selections for the boosted topology . . . 177
5.3.5 Specific selections for the 0-lepton channel . . . 178
5.3.6 Specific selections for the 1-lepton channel . . . 179
5.3.7 Specific selections for the 2-lepton channel . . . 179
5.3.8 Specific $VH, H \to c\bar{c}$ selections . . . 180
5.4 Jet mass scale calibration for small-R jets . . . 182
5.5 Flavour tagging methods . . . 185
5.6 Multivariate discriminants in the $VH, H \to b\bar{b}/c\bar{c}$ analyses . . . 189
5.7 Fit formalism . . . 195
5.7.1 Hypothesis testing, p-value and significance . . . 195
5.7.2 Likelihood function, signal strength and observed significance . . . 196
5.7.3 Asimov dataset and expected significance . . . 199
5.7.4 Pulls, impacts of nuisance parameters and correlations in fit . . . 200
5.7.5 Breakdown of uncertainties . . . 201
5.7.6 Asimov significance . . . 201
5.7.7 Transformation D . . . 202
5.8 Background normalizations and systematic uncertainties . . . 203
5.8.1 Background normalizations and modelling uncertainties . . . 203
5.8.2 Experimental uncertainties . . . 207
5.9 Results from the previous publications . . . 209
5.9.1 $VH, H \to b\bar{b}$ resolved and boosted combination . . . 210
5.9.1.1 Combination strategy . . . 210
5.9.1.2 STXS division of phase space . . . 211
5.9.1.3 Results of the STXS measurement . . . 212
5.9.1.4 EFT interpretation . . . 219
5.9.1.5 Comparison with the CMS latest results . . . 222
5.9.2 $VH, H \to b\bar{b}/c\bar{c}$ combination . . . 224
5.9.2.1 $VH, H \to c\bar{c}$ only likelihood fit . . . 224
5.9.2.2 Combination of $VH, H \to b\bar{b}$ and $VH, H \to c\bar{c}$ . . . 227
5.9.2.3 Comparison with CMS latest results . . . 228
5.9.3 $VH, H \to b\bar{b}/c\bar{c}$ cross-check analyses . . . 229
5.9.3.1 Dijet invariant mass cross-check . . . 229

5.9.3.2 Diboson cross-check 230
5.10 Expected results for the Legacy Run 2 analysis 235
5.10.1 Expected $VH, H \to b\bar{b}$ results 235
5.10.2 Expected $VH, H \to c\bar{c}$ results 242
5.11 Future improvements and prospects for the $VH, H \to b\bar{b}/c\bar{c}$ analysis 246
5.11.1 Run 3 of the LHC 246
5.11.2 High luminosity phase of the LHC 249
5.11.3 Future Circular Collider 251
5.12 Conclusion 252

As highlighted in the Introduction of this thesis, I have been deeply involved in the ATLAS $VH, H \to b\bar{b}/c\bar{c}$ analysis team. The group is targeting a Legacy Run 2 publication for early 2023 and aims, for the first time, for a combination of the three following analyses: the $VH, H \to b\bar{b}$ resolved, the $VH, H \to b\bar{b}$ boosted and the $VH, H \to c\bar{c}$ (resolved) one. As a consequence, the analysis is currently still blinded, so only the expected results for the Legacy publication can be provided in this thesis. The current chapter and the next Chapter 6 are both dedicated to this analysis. I would like to emphasize that both of them have been mostly written before any internal documentation had even started. It was a major challenge to summarize everything in a comprehensive way, produce all the content (tables, schemes, figures) needed given how complex this analysis is and considering the number of significant changes that occured with respect to the previous publications. I also needed to take into account the constant changes over time of the analysis to keep this thesis up-to-date based on the progress and decisions of the group.

The current chapter provides a general presentation of the $VH, H \to b\bar{b}/c\bar{c}$ analysis. Since my work has been focused on $VH, H \to b\bar{b}$ studies, a greater importance is given to the $VH, H \to b\bar{b}$ related part of the analysis compared to $VH, H \to c\bar{c}$ which is more briefly described. A study that I have performed on the jet mass scale calibration is also presented in this chapter, more details are provided hereafter. On the other hand the next chapter is focusing on multivariate studies that I performed in the $VH, H \to b\bar{b}$ 0- and 1-lepton channels. The multivariate approach is an essential part of the analysis as it enhances signal from background discrimination and thus increases sensitivity to $H \to b\bar{b}/c\bar{c}$ decays which is the ultimate goal to probe the SM predictions and perform for instance STXS measurements and EFT interpretations.

In addition to the physics related studies that I performed, I was also strongly involved from a technical point of view in the analysis. As a result of my work, since a bit more that one year now (October 2021), I have been appointed to by the ATLAS $VH(b\bar{b}/c\bar{c})$ group as responsible for the analysis framework. That code plays a leading role in our workflow since it is producing the inputs needed for fits and multivariate analysis studies: it is thus used on a daily basis by several persons. In that context, I was part of the $VH(b\bar{b}/c\bar{c})$ coordination team that takes the important decisions concerning the analysis. I have been one of the main developer and code reviewer of the analysis during the 2 last years of my PhD. I also took care of the documentation of the code both for beginners but also for advance usage of the framework. I developed several important features such as the submission of jobs to the grid (recent development) which should allow our group getting more quickly results. I also provided in many occasion advice to other persons of the group for coding features, debugging and so on. More details about this important contribution can be found in the Introduction of this thesis.

This chapter is organized as follows. First an introduction to the general context of the Legacy Run 2 $VH, H \to b\bar{b}/c\bar{c}$ analysis is provided in Section 5.1 and previous publication results are briefly highlighted. Signal and background events to be considered are introduced as well as the resolved and boosted Higgs decay topologies. The Section 5.2 details the dataset, and corresponding signal and background simulation samples used for such measurements emphasizing the settings for the generation of the nominal Monte Carlo event simulations. The Section 5.3 fully describes

the event reconstructions and selections for the $VH, H \to b\bar{b}$ resolved and boosted analyses as well as the $VH, H \to c\bar{c}$ analysis. In Section 5.4 is presented a study that I carried out about the jet mass scale calibration (JMS) of small-radius jets. ATLAS had, for the first time, provided such mass calibration for small-R jets so the $VH, H \to b\bar{b}/c\bar{c}$ analysis team had to test and decide whether or not to include such jet corrections for the Legacy Run 2 publication. In Section 5.5, flavour tagging techniques to avoid reducing the amount of simulated background events with mistagged light and c-jets (respectively b-jets) after b-tagging (c-tagging) requirements is discussed. The Section 5.6 is dedicated to the overall presentation of the multivariate approach developed for the $VH, H \to b\bar{b}/c\bar{c}$ analysis which results in a significant increase in sensitivity to $H \to b\bar{b}$ and $H \to c\bar{c}$ decays with respect to the previous ATLAS publications. The following Section 5.7 introduces the fit formalism and the fit model of the $VH, H \to b\bar{b}/c\bar{c}$ analysis while the Section 5.8 documents all the systematic uncertainties taken into account for the future publication and as well how the normalization of backgrounds are derived in the different regions of phase space. The Section 5.9 summarizes all important recent and past results obtained by the ATLAS $VH, H \to b\bar{b}/c\bar{c}$ analysis team and, when relevant, a comparison to the CMS corresponding latest results is shown. On the other hand, the expected results for the incoming Legacy Run 2 publication are provided in Section 5.10. The Section 5.11 goes beyond the scope of the Legacy Run 2 analysis and of this thesis by shedding light on future possible and expected improvements that could or should be achievable for different timescales: at the end of the Run 3 of the LHC, for the high-luminosity phase of the LHC and with a possible Future Circular Collider. Finally, the Section 5.12 contains some concluding remarks.

5.1 Introduction to $VH \to b\bar{b}/c\bar{c}$ events and Higgs decay topologies

Although the Higgs boson decays mainly to a pair of b-quarks, the measurement of the $H \to b\bar{b}$ decay properties with ggF, VBF, $t\bar{t}H$ and $b\bar{b}H$ production modes is experimentally challenging due to large contamination of QCD multi-jet background. Indeed, that decay was observed in 2018 by ATLAS and CMS [3, 4] thanks to the so-called VH mechanism which allows to partially circumvent those difficulties by targeting an associated production of the Higgs boson with an electroweak vector boson $V = W$ or Z [274]. For the leptonic decays of the vector bosons, the multi-jet background can be significantly suppressed and the lepton signature, especially in decays to electrons or muons, allows for efficient triggering. The VH leptonic decays provide the best inclusive sensitivity to $H \to b\bar{b}$ and $H \to c\bar{c}$ decays, and VH production mode to date. Such $VH, H \to b\bar{b}/c\bar{c}$ decays are classified in three categories based on the number (0, 1, 2) of charged leptons ($l = e, \mu$), as illustrated in Figure 5.1, corresponding respectively to the following vector boson decays: $Z \to \nu\bar{\nu}$, $W \to l\nu$ and $Z \to l^+l^-$. Taus are not considered as they quickly decay before reaching the ATLAS detector as explained in Section 3.4. All the measurements are performed as a function of the transverse momentum of the vector boson (p_{T}^V) which, for signal events, is correlated with the Higgs boson transverse momentum (p_{T}^H) and is reconstructed with a better resolution.

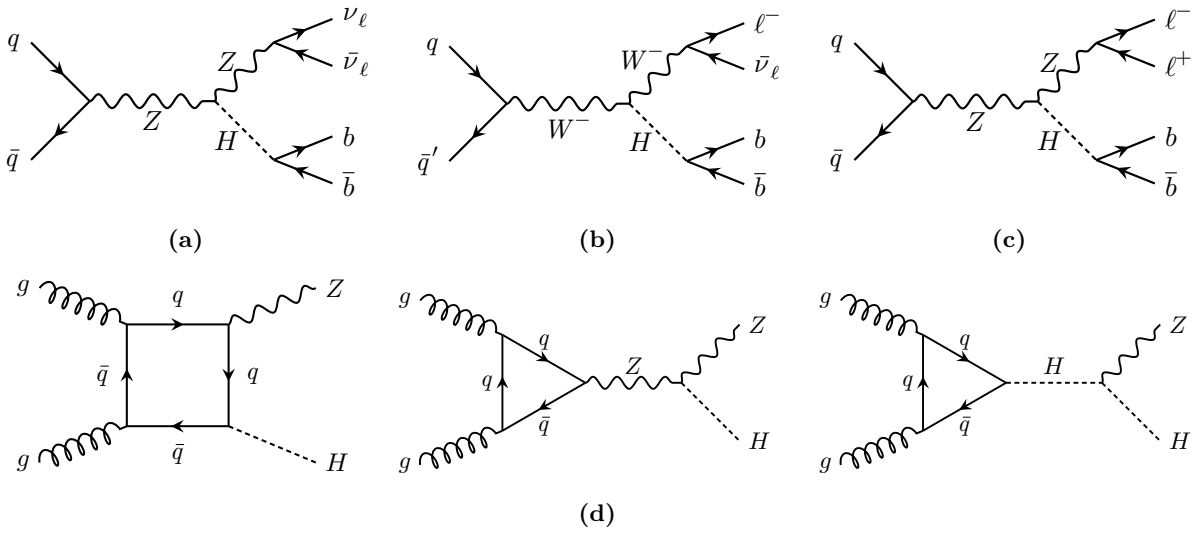


Figure 5.1: Leading order Feynman diagrams of the VH production followed by a $H \to b\bar{b}$ decay for the (a) 0-lepton, (b) 1-lepton and (c) 2-lepton channels ($l = e, \mu$). For the ZH production, the gluon initiated ($gg \to ZH$, (d)) process is also contributing, its cross-section is approximately a factor 6 times smaller than the quark initiated ($qq \to ZH$, (a) and (c)) one [53, 65, 90].

As a result, the first observation of $H \to b\bar{b}$ decays was reported in 2018 using pp collisions of the Run 1 data set collected at 7 TeV and 8 TeV, combined with a partial Run 2 data set collected at 13 TeV both by the ATLAS [3] and CMS collaborations [4].

Since the observation of $H \to b\bar{b}$ decay and VH mechanism, the experimental focus has shifted towards precision measurements of the Higgs boson production kinematics. In its latest results, the ATLAS collaboration has performed inclusive and differential cross-section measurements both with the simplified template cross-section (STXS) [65, 275] and the fiducial [276] approaches. The results have also been interpreted in terms of the Wilson coefficients of BSM operators of an effective field theory Lagrangian [5, 65, 275]. Another important milestone has been achieved combining the $VH, H \to b\bar{b}$ and $VH, H \to c\bar{c}$ analyses [6] which confirmed with a 95% confidence level that the Higgs-charm coupling is weaker than the Higgs-bottom coupling. Finally, a combination for different regimes of the transverse momentum of the vector boson (p_{T}^{V}) was performed [5] in order to reconstruct the Higgs boson candidate (targeting $H \to b\bar{b}$ decays) either as two individual jets (referred to as resolved topology) or as a single large-radius jet (referred to as boosted topology) as schematized in Figure 5.2. All the results mentioned above will be discussed extensively in this chapter.

The ATLAS $VH(b\bar{b}/c\bar{c})$ analysis team is now working on a *Legacy Run 2 publication* in which I have been deeply involved (as already highlighted at the beginning of this chapter), aiming at combining for the first time the $VH, H \to b\bar{b}$ resolved, the $VH, H \to b\bar{b}$ boosted and the $VH, H \to c\bar{c}$ analyses. This paper will serve as a reference measurement for the coming years. It includes all the latest ATLAS developments, updates and recommendations in terms of reconstruction, identification and isolation of objects, modelling of events and recommended sources of uncertainties. The updates and development with respect to the previous papers will be emphasized in this chapter.

For $H \to b\bar{b}/c\bar{c}$ decays, the angular distance between the two b-hadrons (or c-hadrons) [190, 191], as explained in Section 3.6.1, can be approximated by:

$$\Delta R(b, b) \approx \frac{2m_H}{p_{\mathrm{T}}^{H}}. \tag{5.1.1}$$

The Equation (5.1.1) implies that the $H \to b\bar{b}/c\bar{c}$ decays can be reconstructed in ATLAS with the boosted topology for Higgs transverse momentum (p_{T}^{H}) of the order of few hundreds of giga

electron-volts. For instance for $p_{\mathrm{T}}^{H} = 250, 300$ and 400 GeV the angular distances between the two b-jets (or c-jets) are respectively $\Delta R(b,b) \approx 1, 0.83$ and 0.63. The Figure 5.3 presents events display of candidate $VH, H \to b\bar{b}$ decays reconstructed with the resolved or boosted topologies for the three lepton channels.

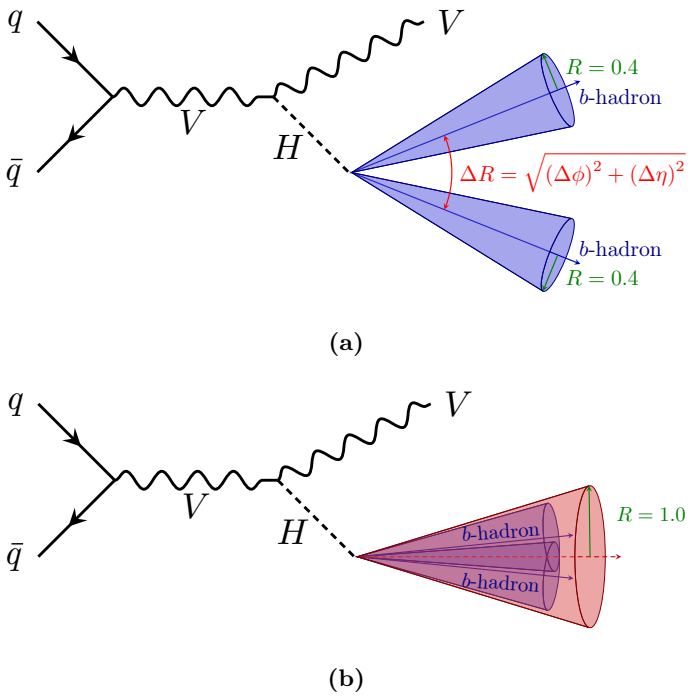
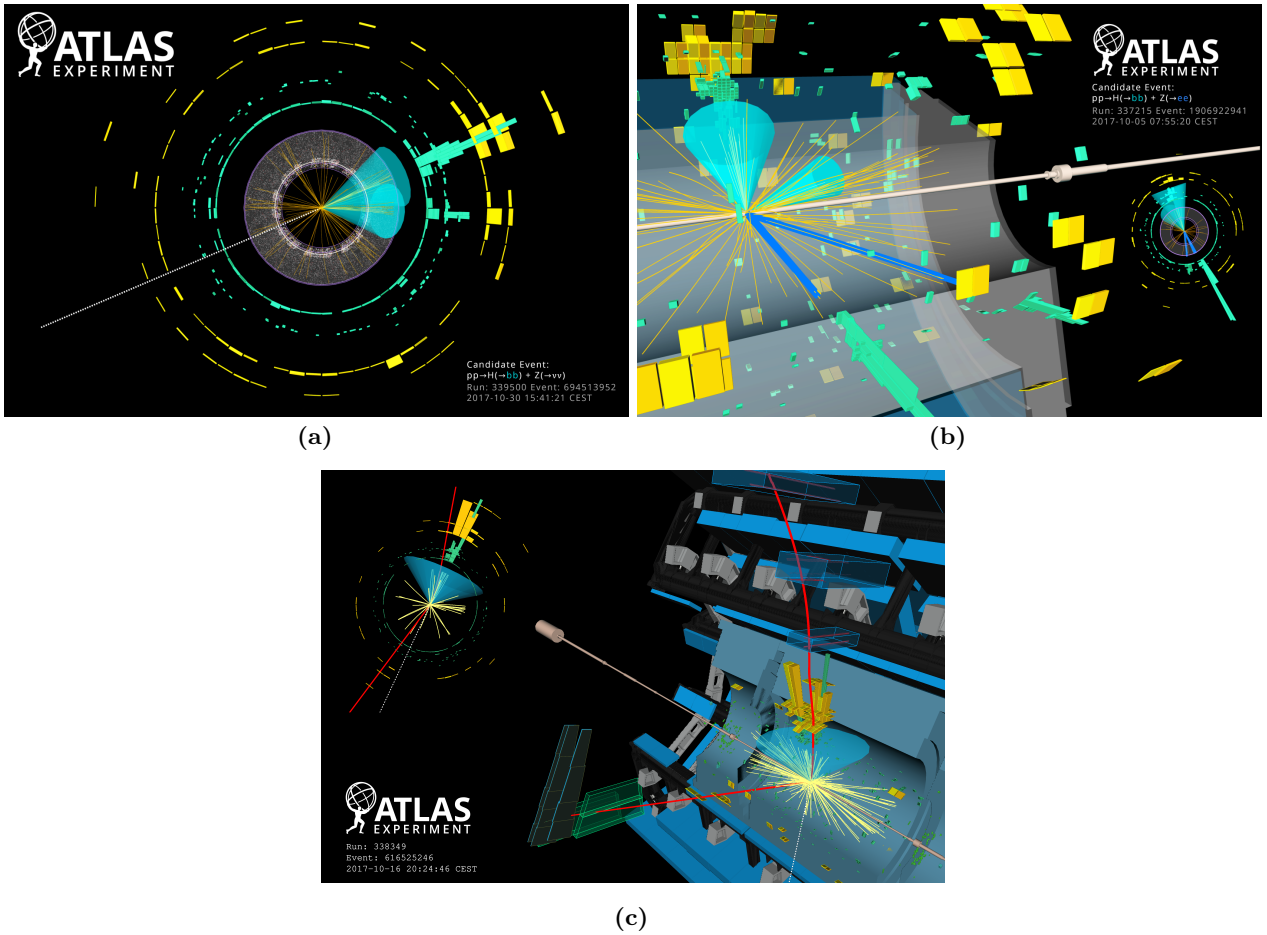


Figure 5.2: $VH, H \to b\bar{b}$ (a) resolved and (b) boosted topologies. For the resolved topology, events are reconstructed with the anti-k_t $R = 0.4$ clustering algorithm, and the b-tagging criteria are required for those small-R jets. While for the boosted topology, events are reconstructed with the anti-k_t $R = 1.0$ algorithm, and b-tagging criteria are required for the VR track-jets that are ghost-associated to the large-R jet.

Finally, the background processes to take into account for the three channels are presented in Figure 5.4, they can be divided in six categories: $t\bar{t}$, single top, W + jets, Z + jets, diboson and multi-jet events. Depending on the channel, their relative contributions can vary. Those background sources can be split in two categories:

- **the non-resonant backgrounds:** for example $t\bar{t}$ events are a non-resonant background when the two b-jets are correctly identified. Since each b-jet is emitted by one of the top quark, their invariant mass distribution (m_{bb}) is not expected to peak around a certain value and their kinematic properties ($\Delta R(b,b), p_{\mathrm{T}} \ldots$) are not expected to mimic the $H \to b\bar{b}$ two body disintegration. Those non-resonant background sources are hence reducible by means of event selections. Moreover, the machine learning algorithms used in these analyses can further discriminate those events from signal ones taking into account correlations between event kinematic properties. However, those events can still reduce sensitivity to signal mimicking its signature (*i.e.* its characteristics) due to stochastic effects or due to particle mis-identifications hence they can fulfil event selections and may be wrongly considered as signal-like by machine learning algorithms. As a result, the $t\bar{t}$ events which have a large cross-section are still one of the dominant background in the 1-lepton channel.
- **The resonant backgrounds**: the two b-jets in that case are produced by the decay of a particle such as $Z \to b\bar{b}$. A fraction of those events is an "irreducible" background because intrinsically they share several kinematic characteristics with $H \to b\bar{b}$ decays and are hence signal-like disintegrations. For $Z \to b\bar{b}$ decays, the upper tail of invariant mass distribution of the b-jet system (m_{bb}) can mimic the $H \to b\bar{b}$ signal. Event selections and machine learning algorithms



(a) **(b)**

(c)

Figure 5.3: ATLAS event display of (a) $ZH \to \nu\nu b\bar{b}$ and (b) $ZH \to \ell\ell b\bar{b}$ ($l = e$) candidate events reconstructed with the resolved topology. The b-jets (blue cones) are reconstructed with small-R jets ($R = 0.4$), their associated energy deposition in the electromagnetic and hadronic calorimeter are respectively schematized by the green and yellow bars. For the $ZH \to \nu\nu b\bar{b}$ candidate event, the neutrinos leave the detector unseen, and are reconstructed through the missing transverse energy (dashed line). For the $ZH \to \ell\ell b\bar{b}$ candidate event, the 2 electrons produced by the Z boson are represented by the blue tracks with a large energy deposit in the electromagnetic calorimeter. (c) A boosted $WH \to \ell\nu b\bar{b}$ candidate event is reconstructed with a single large-R jet ($R = 1.0$). A two-prong structure is visible within the large-R jet with energy deposits in both the electromagnetic and hadronic calorimeters. In addition, a so-called *muon-in-jet* (upper red track), which created hits in the barrel muon chambers, can be found within the large-R jet. It comes from a semi-leptonic decay of one of the two b-hadrons contained in the large-R jet. On the other hand, the isolated muon (lower red track) produced by the W boson decay, created hits in the endcap-muon chambers. The associated neutrino is reconstructed through the missing transverse energy (taken from Refs. [275, 277]).

can of course reduce their contribution, but their discrimination from signal is in general more complicated.

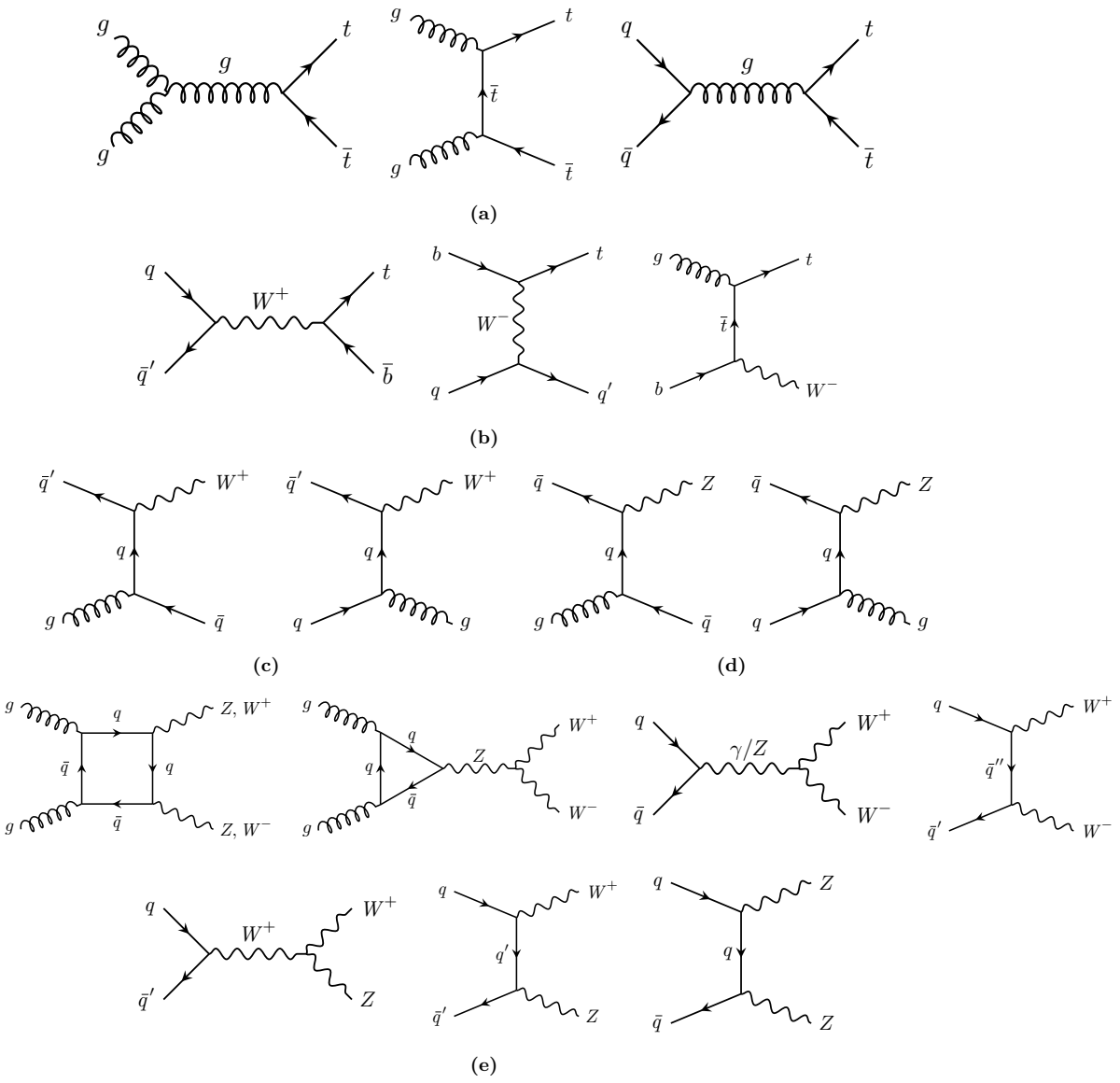


Figure 5.4: Main $VH, H \to b\bar{b}/c\bar{c}$ background processes: respectively (a) gluon and quark induced $t\bar{t}$ events, (b) s-channel, t-channel and Wt single top events, (c) gluon and quark induced W+jets events, (d) gluon and quark induced Z+jets events and (e) gluon and quark induced diboson events. Multi-jet background is not shown. It is reduced to the percent level for the 1-lepton channel while its impact was found to be negligible for the 0- and 2-lepton channel as described in References [3, 65].

5.2 Data and simulation samples

The data were collected in pp collisions at $\sqrt{s} = 13$ TeV during the Run 2 of the LHC. The data sample corresponds to an integrated luminosity of 139 fb^{-1}, reported per year in Table 5.1, after requiring that all detector subsystems were operating normally and recording high-quality data [123]. The uncertainty in the combined 2015-2018 integrated luminosity is 1.7% [278], obtained using the LUCID-2 detector [279] for the primary luminosity measurements.

Collision events considered were recorded with a combination of triggers selecting events with high missing transverse momentum or with a high-p_{T} lepton(s), depending on the channel. More

details about the trigger selections are provided in Section 5.3.

Year	LHC delivered	ATLAS recorded	ATLAS good quality
2015	4.2 fb^{-1}	3.9 fb^{-1}	3.2 fb^{-1}
2016	38.5 fb^{-1}	35.6 fb^{-1}	32.9 fb^{-1}
2017	50.2 fb^{-1}	46.9 fb^{-1}	44.3 fb^{-1}
2018	63.3 fb^{-1}	60.6 fb^{-1}	58.5 fb^{-1}
Total	156.2 fb^{-1}	147.0 fb^{-1}	138.9 fb^{-1}

Table 5.1: Integrated luminosity per year of data-taking (numbers reported come from Refs. [280, 281]).

The properties of signal and background processes are studied with Monte Carlo event simulation relying on the ATLAS detector simulation [82] which is based on GEANT4 [83]. A summary of all the signal and background processes with the corresponding generators used for the nominal samples is shown in Table 5.2. The associated leading Feynman diagrams for those processes can be found in Figures 5.1 and 5.4. They are all normalized using the most precise theoretical predictions currently available of their cross-sections. In addition to the hard scatter simulated process, each event is overlaid with pile-up collisions generated with PYTHIA8.1 [73] using the ATLAS A3 set of tuned parameters [256] and the NNPDF23LO [257] parton distribution function (PDF) set. Those additional pp collisions are generated following the average profile of the number of interactions (μ) per bunch crossing measured in data. Data and simulated events go through the exact same set of algorithms.

Process	ME generator	ME PDF	PS and Hadronisation	UE model tune	Cross-section order
Signal ($m_H = 125$ GeV, $b\bar{b}$ and $c\bar{c}$ branching fraction set to 58.2% and 2.89%)					
$qq \to W(\to \ell\nu)H(\to b\bar{b}/c\bar{c})$	POWHEG BOX v2 [79] + GOSAM [283] + MINLO [284, 285]	NNPDF3.0NLO$^{(\star)}$ [252]	PYTHIA 8.212 [74]	AZNLO [282]	NNLO(QCD)+ NLO(EW) [286–292]
$qq \to Z(\to \nu\nu/\ell\ell)H(\to b\bar{b}/c\bar{c})$	POWHEG BOX v2 + GOSAM + MINLO	NNPDF3.0NLO$^{(\star)}$	PYTHIA 8.212	AZNLO	NNLO(QCD)$^{(\dagger)}$+ NLO(EW)
$gg \to Z(\to \nu\nu/\ell\ell)H(\to b\bar{b}/c\bar{c})$	POWHEG BOX v2	NNPDF3.0NLO$^{(\star)}$	PYTHIA 8.212	AZNLO	NLO+ NLL [293–297]
Top quark ($m_t = 172.5$ GeV)					
$t\bar{t}$	POWHEG BOX v2 [79, 246]	NNPDF3.0NLO	PYTHIA 8.230	A14 [269]	NNLO+NNLL [298]
s-channel	POWHEG BOX v2 [79, 247]	NNPDF3.0NLO	PYTHIA 8.230	A14	NLO [299]
t-channel	POWHEG BOX v2 [79, 247]	NNPDF3.0NLO	PYTHIA 8.230	A14	NLO [300]
Wt	POWHEG BOX v2 [79, 248]	NNPDF3.0NLO	PYTHIA 8.230	A14	Approximate NNLO [301]
Vector boson + jets					
$W \to \ell\nu$	SHERPA 2.2.1 [75, 76, 302, 303]	NNPDF3.0NNLO	SHERPA 2.2.1 [304, 305]	Default	NNLO [306]
$Z/\gamma^* \to \ell\ell$	SHERPA 2.2.1	NNPDF3.0NNLO	SHERPA 2.2.1	Default	NNLO
$Z \to \nu\nu$	SHERPA 2.2.1	NNPDF3.0NNLO	SHERPA 2.2.1	Default	NNLO
Diboson					
$qq \to WW$	SHERPA 2.2.1	NNPDF3.0NNLO	SHERPA 2.2.1	Default	NLO
$qq \to WZ$	SHERPA 2.2.1	NNPDF3.0NNLO	SHERPA 2.2.1	Default	NLO
$qq \to ZZ$	SHERPA 2.2.1	NNPDF3.0NNLO	SHERPA 2.2.1	Default	NLO
$gg \to VV$	SHERPA 2.2.2	NNPDF3.0NNLO	SHERPA 2.2.2	Default	NLO

Table 5.2: Signal and background processes with the corresponding generators used for the nominal samples. If not specified, the order of the cross-section calculation refers to the expansion in the strong coupling constant (α_s). ($\star$) The events were generated using the first PDF in the NNPDF3.0NLO set and subsequently reweighted to the PDF4LHC15NLO set [271] using the internal algorithm in POWHEG BOX v2. ($\dagger$) The NNLO(QCD)+NLO(EW) cross-section calculation for the $pp \to ZH$ process already includes the $gg \to ZH$ contribution. The $qq \to ZH$ process is normalized using the cross-section for the $pp \to ZH$ process, after subtracting the $gg \to ZH$ contribution. An additional scale factor is applied to the $qq \to VH$ processes as a function of the transverse momentum of the vector boson, to account for electroweak (EW) corrections at NLO. This makes use of the VH differential cross-section computed with HAWK [307, 308] (slightly changed table taken from Ref. [65]).

The simulation of the different nominal samples is performed with the settings described below. For all samples, except those generated with SHERPA , the EVTGEN V1.2.0 program [309] was

used to describe the decays of b- and c-hadrons.

➢ **The signal samples**: are generated assuming a mass of the Higgs boson of $m_H = 125$ GeV. The difference between the generated value of the mass and the measured value from ATLAS and CMS combination measurement (125.09 ± 0.24 GeV) [88] has a negligible impact on the acceptance and the reconstructed distributions given the experimental resolution of the ATLAS detector.

- **The qq-initiated processes** are simulated with up to one additional parton at next-to-leading-order (NLO) accuracy in QCD using the POWHEG BOX v2 [79] and the GOSAM [283] matrix element (ME) generator with the MiNLO (Multiscale Improved NLO) [284, 285] procedure applied.
- **The gg-initiated processes** are simulated at leading order (LO) in QCD with POWHEG BOX v2. The cross-section for those processes was calculated at NLO in QCD including soft gluon re-summation up to next-to-leading logarithms (NLL) [293–297].

For all signal processes (WH and ZH), the matrix element (ME) computations are performed with the NNPDF3.0NLO PDF set and are afterwards reweighted to the PDF4LHC15NLO PDF set [271]. Events are then interfaced to PYTHIA 8.212 [74] for the parton showering (PS) and hadronization with the AZNLO [282] set of tune for the underlying events (UE). The total inclusive cross-sections were calculated at next-to-next-to-leading-order (NNLO) QCD and NLO electroweak (EW) [286–292] accuracy, including photon contributions calculated with HAWK [307, 308].

➢ **The top samples (top pair and single top samples)**: are generated with the POWHEG BOX v2, assuming a top mass $m_t = 172.5$ GeV, and interfaced with PYTHIA 8.230 [73] for PS with the ATLAS A14 [269] set of tune for the UE. The NNPDF3.0NLO [252] PDF set is used for the ME computations. The top background is one of the leading background of this analysis and samples have the following characteristics:

- **for $t\bar{t}$ samples**, the event generator uses matrix elements derived at NLO accuracy, and the inclusive process cross-section is obtained from resummation at NNLO and next-to-next-to-leading logarithm (NNLL) computations [298]. The $t\bar{t}$ samples as mentioned in the Chapter 4 are split into *non-all hadronic* and *dilepton* samples in order to increase the number of simulated events available using dedicated event filtering to have a coherent combination. Several p_{T}^{V}-slices are combined for the *non-all hadronic* sample to enhance statistics especially in high-p_{T} regions. Finally, in addition to the sample used in Chapter 4, new sliced samples as a function of $E_{\mathrm{T}}^{\mathrm{miss}}$ have been produced and are used in the 0- and 1-lepton channels.
- **For the single top s- and t-channels**, the cross-section normalization are estimated from resummed calculations at NLO [299, 300].
- **The single top Wt channel** normalization is estimated from approximate NNLO calculations of the cross-section [301]. At higher QCD orders, the single top Wt events can overlap with the leading order $t\bar{t}$ events. In order to account for those ambiguities and avoid interferences effects, the diagram subtraction (DS) scheme [249] is used: the Wt cross-section is corrected with a subtraction term in order to remove the double counting of some events when summing $t\bar{t}$ and single top Wt contributions. As that correction is only performed at cross-section level, an interference term between those two samples remains. In the previous rounds of the analysis, the diagram removal (DR) scheme [249] was used instead: this procedure consists in removing doubly resonant diagrams for the Wt amplitude computations. The DS samples are now preferred because a better agreement between the data and the simulations was found compared to the DR samples. $E_{\mathrm{T}}^{\mathrm{miss}}$-filtered Wt samples have been produced and are currently used in the 0- and 1-lepton channels to increase MC statistics. For the 2-lepton channel, a specific Wt dilepton sample has been produced and is combined with the inclusive Wt sample.

➢ **The V + jets samples ($V = W, Z$)**: are simulated with the SHERPA 2.2.1 [75, 76, 302, 303]

ME generator using OPENLOOPS [75, 302, 303] with NLO accuracy corrections up to two additional jets, and at LO for up to four additional jets. The SHERPA 2.2.1 PS model is used with its default set of tune for the UE. The NNPDF3.0NNLO [252] PDF set is used for ME computations. The cross-section for those backgrounds are obtained from an NNLO fixed-order estimate [306]. Since $V+\text{jets}$ background is also one of the leading contamination, those samples are also composed of p_{T}^{V}-extensions and are in addition completed with different slices in $\max(H_{\text{T}}, p_{\text{T}}^{V})$, H_{T} being the scalar p_{T}-sum of all parton-level jets with $p_{\text{T}} > 20\,\text{GeV}$. On top of that, those samples have been generated with flavour filters in order to ensure statistics for heavy flavour components. As a result b-filtered, b-vetoed+c-filtered and b-vetoed+c-vetoed samples are available selecting events which respectively have at least one b-hadron, no b-hadron and at least one c-hadron, and no b-hadron and c-hadron in the event.

- **The diboson samples**: event generation, parton showering and underlying event modelling are all performed with SHERPA 2.2.1 or 2.2.2 with their default set of tune for the UE. The NNPDF3.0NNLO PDF set is used for ME computations. The dominant qq-initiated processes containing zero or one additional parton are calculated at NLO accuracy in QCD in the ME, while two or three additional partons are included at LO accuracy in QCD. For the subdominant gg-initiated processes zero or one additional parton are calculated at LO accuracy in QCD in the ME.
- **The QCD multi-jet (MJ) background**: despite not being a source of genuine missing transverse momentum, misreconstruction of this background can lead in a small fraction of events to have a significantly large, fake, missing transverse momentum when the energy of some of the reconstructed objects is not well calibrated. Since this background has a large cross-section, it needs to be taken into account. Thanks to angular event selections, this background is negligible for the 0- and 2-lepton channels while it is reduced to the percent level in the 1-lepton channel [3, 65], and it is furthermore greatly reduced in high-p_{T}^{V} regions. The multi-jet events passing the selections are involving fake leptons that are either jets wrongly identified as electromagnetic showers or electrons/muons produced in non-prompt weak decays in jets. In both cases the fake-lepton or lepton-in-jet is expected to be non-isolated. Indeed, the number of MJ events passing the selection criteria is very limited resulting in a lack of statistics. Modelling this background with simulations is hence impossible: its contribution needs to be estimated using a data-driven technique.

  The MJ background is estimated using an enriched control region of phase space (denoted MJ CR) inverting the isolation criterion usually required for the analysis (criterion detailed in Section 5.3). The b-tagging requirement is also loosened to increase statistics accepting as well events with only 1 b-tagged jet. The MJ CR is orthogonal to the signal region (SR) *i.e.* there is no overlap between them, and still enables to select kinematically similar events as for the SR. It is assumed that the MJ shape is not changing between the MJ CR and SR and shape uncertainties are considered for the MJ background as described in Section 5.8.1. A template fit of the transverse mass of the W boson (m_{T}^{W}, quantity defined in Section 4.3) is then performed in order to obtain the MJ yield in the SR:

  - in the MJ CR, the shape of the m_{T}^{W} MJ background is obtained from a fit of that distribution subtracting from data the shape of the other backgrounds (top, diboson and $V+\text{jets}$) templates which are referred to as the EW component.
  - Then the m_{T}^{W} MJ shape obtained in the MJ CR is used in the SR and the template fit of m_{T}^{W} is repeated in the SR with the EW and MJ components shapes: it enables to obtain the MJ yield.

  The m_{T}^{W} mass for MJ yield extraction, as shown in Figure 5.5, is used because it was found to be the most discriminating variable between the MJ and EW components. Then the shape of the MJ background for any desired distributions (including the ones of interest for which the significance is extracted) are obtained similarly using the shape of the distribution from the MJ CR in the SR, and the yield obtained previously thanks to the m_{T}^{W} template fit. The fits are

performed simultaneously dividing the phase space according to the lepton flavour (e/μ), the jet multiplicity (2/3-jets regions) and the p_{T}^{V}-regions.

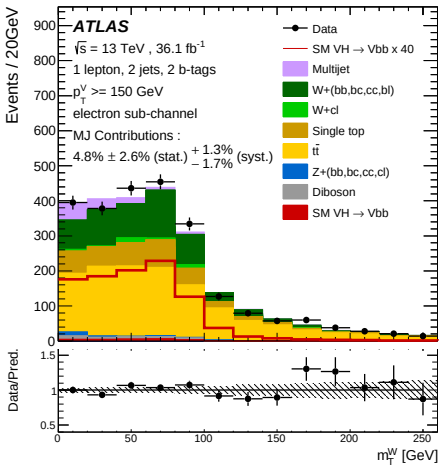


Figure 5.5: Multi-jet (MJ) estimate in the 1-lepton 2 b-tagged 2-jets category with the template fit using the m_{T}^{W} distribution in the electron sub-channel in the signal region for 2015+2016 ATLAS data-taking (taken from Ref. [310]).

SHERPA 2.2.11 samples are currently under study by the ATLAS $VH(b\bar{b}/c\bar{c})$ analysis team and could be used for the Legacy Run 2 publication as a replacement for the SHERPA 2.2.1 and 2.2.2 samples. These new samples come with an increased number of simulated events and updated modelling tunes.

As a result compared to the previous published results, an increase of the number of simulated events is obtained thanks to further extensions into sliced or filtered samples[1] which should allow obtaining smoother distributions and reduce statistical MC uncertainties in particular for low-statistic regions of phase space such as high-p_{T} and/or heavy-flavour regions.

5.3 $VH, H \to b\bar{b}/c\bar{c}$ event reconstructions and selections

The reconstruction and the identification of objects, the definition of electron and muon quality criteria and isolation working points, the flavour tagging of jets and so on is detailed in the Chapter 3. Below is a quick summary of the object characteristics used for the $VH, H \to b\bar{b}/c\bar{c}$ analyses.

- **Primary vertex (PV) and tracks reconstruction** (Section 3.1): the primary vertices are reconstructed with a 3D imaging technique. The PV of interest is selected as the one with the highest $\sum p_{\mathrm{T}}^{2}$ of all associated tracks. Tracks are reconstructed by means of a pattern search and fitting procedure.
- **Electrons** (Section 3.2): are reconstructed from topological clusters of energy deposits in the electromagnetic calorimeter and matched to a track in the inner detector. Quality and isolation criteria provided by ATLAS can be chosen depending on the events and signal-to-background ratio targeted by an analysis.
- **Muons** (Section 3.3): are reconstructed thanks to measurements from the muon spectrometer. Quality and isolation criteria provided by ATLAS can be chosen depending on the events and signal-to-background ratio targeted by an analysis.

1. In the previous rounds of the analysis, sliced samples were already used. For the Legacy Run 2 publication more extensions have been requested which should be beneficial to reduce statistic uncertainties.

- **Taus** (Section 3.4): decay quickly either leptonically or hadronically. In the former case, the visible decay products are reconstructed as electron or muon candidates. In the latter, hadrons can be clustered into jets that are then identified as hadronically-decaying tau leptons by a RNN algorithm which comes with increased performances in identification and rejection of fake tau-jets compared to the previous BDT algorithm.
- **Jets** (Section 3.6): small-R jets are built from topo-clusters calibrated at the electromagnetic scale and are reconstructed with the anti-k_t $R = 0.4$ clustering algorithm while large-R jets are reconstructed with $R = 1.0$ from topo-clusters calibrated at the local hadronic scale. The VR track-jets used in this analysis, whose radius R can vary between $0.02 < R < 0.4$ and have $\rho = 30$ GeV (see Equation (3.6.12)), are ghost-associated to the large-R jets. The large-R jets are b-tagged thanks to the VR track-jets which are preferred to small-R jets because of the better spatial resolution of tracks with respect to the calorimeter clusters. Moreover, the VR track-jet minimal and maximal radius, and ρ parameters have been optimized for $H \to b\bar{b}$ double b-tagging in boosted decays over a large p_{T}^{H}-spectrum as shown in Figure 3.26.
  - For small-R jets, a new algorithm called Particle Flow (PFlow) exploiting both information from the calorimeter and tracker has been developed by ATLAS compared to the former EMTopo jet reconstruction algorithm which only takes into account information from the calorimeter. The PFlow jets have several improvements in particular a better reconstruction efficiency, jet energy and angular resolutions are expected with as well a higher suppression of pile-up contamination compared to EMTopo-jets.
  - The large-R jets are reconstructed from the EM topological clusters using only calorimeter information. New reconstruction algorithms similar to the PFlow algorithm but for large-R jets, such as Track-CaloClusters (TCC, [311]) or Unified Flow Objects (UFO, [312]) jets have been developed by ATLAS but for TCC jets a significant decrease of signal yields was found while the main background yields remained unchanged, and the UFO jet could be used for analysis at the moment since their calibrations is not yet finalized.
  - A new ANN flavour tagging algorithm, called DL1r, has been developed and comes with an increased rejection for light- and c-jets for b-tagging compared to the former MV2c10 BDT algorithm both when tagging small-R jets and VR track-jets as shown in Figure 3.29. The increase in performance is obtained thanks to the use, by the DL1r tagger, of the low level RNNIP tagger output variables which exploits correlations between tracks leading to an increased discrimination between b-, c- and light-jets. To be noted that flavour tagging can only be performed within the tracker acceptance *i.e.* for jets with $|\eta| < 2.5$ as track related information are primordial for discrimination.
  - A jet cleaning procedure is applied to each event to avoid local pile-up fluctuations that can induce spurious jets that would appear when applying pile-up corrections. This further decrease of pile-up noise is obtained using a jet vertex tagger (JVT).
- **Missing transverse energy** (Section 3.8): the $E_{\mathrm{T}}^{\mathrm{miss}}$ is reconstructed from the negative sum of all objects hard and soft ones measured in the calorimeter. It accounts for undetected particles in the ATLAS detector such as neutrinos exploiting the conservation of momentum in the transverse plane. The $E_{\mathrm{T}}^{\mathrm{miss}}$ resolution is expected to be improved thanks to the PFlow jets. The track-based $p_{\mathrm{T}}^{\mathrm{miss}}$ corresponds to the transverse missing momentum which is based on the same principle but only using charged particle tracks momenta measured in the tracker.

An overlap removal procedure (Section 3.7) is applied before applying event selections in order to lift ambiguities between reconstructed objects associated to the same signal in the detector.

The Section 5.3.1 provides a quick overview of the combination of $VH, H \to b\bar{b}$ and $VH, H \to c\bar{c}$ events while other sections are presenting the different selections. Since the $VH, H \to b\bar{b}$ resolved and $VH, H \to c\bar{c}$ analyses have been harmonized as much as possible a comparison of the resolved and boosted selections is first presented while the $VH, H \to c\bar{c}$ specific selections are discussed in a dedicated section highlighting the differences in selection with respect to the $VH, H \to b\bar{b}$

resolved case.

Table 5.3 summarizes the selections applied to the resolved and boosted topologies for the $VH, H \to b\bar{b}$ analyses. Some selections discussed in the Sections below may not be the final ones for the Legacy Run 2 publication as the analysis is still being optimized.

Selection	Topology	0-lepton	1-lepton		2-lepton	
			e sub-channel	μ sub-channel	e sub-channel	μ sub-channel
Trigger	R & B	$E_{\mathrm{T}}^{\mathrm{miss}}$	Single lepton	$E_{\mathrm{T}}^{\mathrm{miss}}$	Single lepton	$E_{\mathrm{T}}^{\mathrm{miss}}$
Leptons	R & B	0 *loose* leptons	Exactly 1 *tight* lepton 0 additional *loose* leptons $p_{\mathrm{T}} > 27$ GeV	$p_{\mathrm{T}} > 25$ GeV	Exactly 2 *loose* leptons Same flavour for both leptons $p_{\mathrm{T}} > 27$ GeV for the leading lepton –	Opposite charges
$m_{\ell\ell}$	R	–	–	–	$81\text{ GeV} < m_{\ell\ell} < 101\text{ GeV}$	
	B	–	–	–	$66\text{ GeV} < m_{\ell\ell} < 116\text{ GeV}$	
$E_{\mathrm{T}}^{\mathrm{miss}}$	R	> 150 GeV	> 30 GeV	–	–	
	B	> 400 GeV	> 50 GeV	–	–	
p_{T}^{V}	R	> 150 GeV			> 75 GeV	
	B	> 400 GeV				
Small-R jets	R	$p_{\mathrm{T}} > 20$ GeV for $\lvert\eta\rvert < 2.5$ $p_{\mathrm{T}} > 30$ GeV for $2.5 < \lvert\eta\rvert < 4.5$				
	B	$p_{\mathrm{T}} > 30$ GeV				
Large-R jets	B	$p_{\mathrm{T}} > 250$ GeV and $\lvert\eta\rvert < 2.0$				
Track-jets	B	At least 2 track-jets with $p_{\mathrm{T}} > 10$ GeV, $\lvert\eta\rvert < 2.5$ matched to the leading large-R jet				
b-tagging	R	Exactly 2 b-tagged small-R jets (DL1r, 70% WP) and $p_{\mathrm{T}} > 45$ GeV for the leading b-tagged jet				
	B	Exactly 2 b-tagged track-jets (DL1r, 85% WP) out of the 3 leading track-jets matched to the leading large-R jet				
m_{bb}	R	> 50 GeV				
m_{J}	B	> 50 GeV				
Jet categories	R	***2 categories:*** exactly 2 ***or*** 3 small-R jets			***2 categories:*** 2 ***or*** ≥ 3 small-R jets	
	B	***2 categories:*** at least 1 large-R jet and 0 ***or*** ≥ 1 b-tagged track-jet not matched to the leading large-R jet				
$\min[\Delta\phi(\boldsymbol{E}_{\mathbf{T}}^{\mathbf{miss}}, \text{small-R jets})]$	R	$> 20^{\circ}$ (2-jets), $> 30^{\circ}$ (3-jets)	–			
	B	$> 30^{\circ}$	–			
$\Delta\phi(\boldsymbol{E}_{\mathbf{T}}^{\mathbf{miss}}, \boldsymbol{H}_{\mathbf{cand}})$	R & B	$> 120^{\circ}$	–			
$\Delta\phi(\boldsymbol{E}_{\mathbf{T}}^{\mathbf{miss}}, \boldsymbol{p}_{\mathrm{T}}^{\mathrm{miss}})$	R & B	$< 90^{\circ}$	–			
$\Delta\phi(\boldsymbol{b_1}, \boldsymbol{b_2})$	R	$< 140^{\circ}$	–			
H_{T}	R	> 120 GeV (2-jets), > 150 GeV (3-jets)	–			
Signal regions	R	***2 regions (2 b-tagged):*** 2-jets ***and*** 3-jets regions passing the $\Delta R(b,b)$-signal selection			***2 regions (2 b-tagged):*** 2-jets ***and*** $\geq$ 3-jets regions passing the $\Delta R(b,b)$-signal selection	
	B	***2 regions (2 b-tagged no add. b-tagged track-jet):*** 0 (HP-region) ***and*** ≥ 1 add. small-R jets (LP-region)			***1 region (2 b-tagged $\geq$ 0 add. b-tagged track-jet):*** ≥ 0 add. small-R jets region (no split)	

Table 5.3: Summary of the $VH, H \to b\bar{b}$ event selection and categorization in the 0-, 1- and 2-lepton channels for the resolved and boosted topologies respectively denoted by R and B in the table. The 2 small-R jets ($b\bar{b}$) system or the large-R jet (J) Higgs candidate is sometimes denoted H_{cand}.

5.3.1 Overview of the $VH, H \to b\bar{b}/c\bar{c}$ event combination

This section provides a brief overview of the combination of the $VH, H \to b\bar{b}$ and $VH, H \to c\bar{c}$ events. In the next sections, the details will be provided about the event selections, the strict definition of the signal (SR) and control (CR) regions, b- and c-tagging working points.

The aim of the Legacy Run 2 publication is to combine the $VH, H \to b\bar{b}$ resolved, the $VH, H \to b\bar{b}$ boosted and the $VH, H \to c\bar{c}$ (resolved) events: the phase space needs to be divided in 3 orthogonal (*i.e.* non overlapping) regions for those measurements. This division is possible thanks to the use of b- and c-taggers and based on event topology criterions so that events can only be selected by a single analysis and are not double counted. The $VH, H \to c\bar{c}$ signal and control regions phase space has significantly changed, compared to the previous published paper [6], and is now harmonized with the $VH, H \to b\bar{b}$ resolved definitions.

As explained in Section 3.6.5, the DL1r b- and c-taggers output 3 probabilities p_b, p_c and p_{light} for jets corresponding to their respective probabilities to be a b-, c- and light-jet. The discriminants used for b- and c-tagging can afterwards be computed as:

$$\mathcal{D}_{\text{DL1r}_b} = \log\left(\frac{p_b}{f_c \times p_c + (1 - f_c) \times p_{\text{light}}}\right) \tag{5.3.1}$$

$$\mathcal{D}_{\text{DL1r}_c} = \log\left(\frac{p_c}{f_b \times p_b + (1 - f_b) \times p_{\text{light}}}\right), \tag{5.3.2}$$

with $f_c = 0.018$ (respectively $f_b = 0.3$) the fraction of c-jets (b-jets) in the background sample used for the DL1r b-tagger (c-tagger) training.

For the $VH, H \to b\bar{b}/c\bar{c}$ analysis, a *2D pseudo-continuous b/c-tagging* strategy, represented in Figure 5.6, is used in order to ensure that each small-R jet falls into a unique category either being a b-tagged or a c-tagged or an untagged jet. Jets whose $\mathcal{D}_{\text{DL1r}_b}$ discriminant passes the 70% or 60% b-tagging efficiency discriminant threshold are respectively classified as loose and tight b-tagged jets or referred to as 70% b-tagged and 60% b-tagged jets and are denoted "B". Non b-tagged jets are similarly checked to be loose (denoted "L") or tight (denoted "T") c-tagged. The remaining jets are untagged (also referred to as non-tagged and are denoted "N"). Scale factors are derived for each bin using $t\bar{t}$ reference samples to correct for differences in tagging efficiencies between data and simulation.

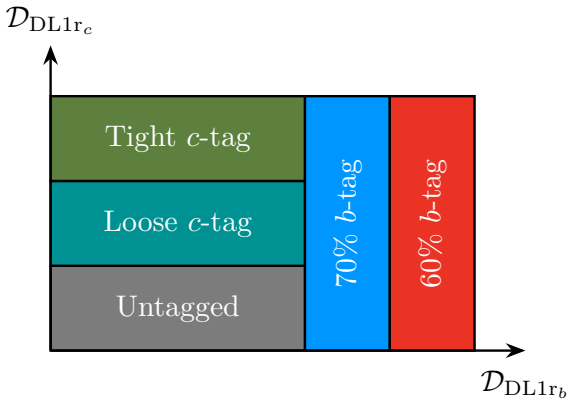


Figure 5.6: Small-R jets pseudo-continuous 2D b/c-tagger implementation to obtain an orthogonal division of phase space. Jets passing the 60% or the 70% efficiency b-tagging requirements are b-tagged (denoted B) and are respectively referred to as 60% (or tight) and 70% (or loose) b-tagged jets. Jets failing b-tagging requirements are tested to be c-tagged with tight and loose requirements and, if so, are respectively denoted T and L. The remaining jets are untagged jets, they are also referred to as non-tagged jets and are denoted N.

Both for the resolved and boosted $VH, H \to b\bar{b}$ analyses, exactly 2 b-tagged jets in the event are required:

- **for the resolved topology**, the *only 2 b-tagged* requirement concerns all small-R jets in the event. Control regions are defined by means of continuous low and high ΔR versus p_{T}^{V} cuts while the signal region is the central region between those 2 side-bands. The signal and control regions are split into 2 categories: the 2-jets and 3-jets events ($\geq$ 3-jets events in the 2-lepton channel), the 3-jets (or ($\geq$ 3-jets) category allowing to recover VH decays with initial (ISR) or final (FSR) state radiations. The two b-tagged jets are constituting the Higgs candidate system. In the 3-jets category, events with a third tight c-tagged jet are vetoed in the 0- and 1-lepton channels as a "BBT" top control region (*i.e.* 2 b-tagged jets and 1 tight c-tagged jet) which is highly pure in $t\bar{t}$ events is foreseen to be used for the $VH, H \to c\bar{c}$ analysis. Such c-veto has a minimal impact on the $VH, H \to b\bar{b}$ signal yields.
- **For the boosted topology**, the leading large-R jet of the event *i.e.* the one with the highest transverse momentum is required to have exactly 2 of its 3 leading ghost-associated VR track-jets to be b-tagged. If such criterion is satisfied then the leading large-R jet constitutes the Higgs candidate otherwise the event is rejected. If an additional track-jet (*i.e.* a track not ghost-associated to the leading large-R jet) passes b-tagging requirements, then the event is classified in the top control region otherwise the event enters the signal region. The signal region is further split in two categories, the high purity (HP SR) and low purity (LP SR) signal regions corresponding respectively to events having no additional small-R jet or at least one additional small-R jet, the additional small-R jets being jets not matched to the leading large-R jet *i.e.* $\Delta R(\text{small-R jet, leading large-R jet}) > 1.0$.

The $VH, H \to c\bar{c}$ analysis only exploits the resolved topology and is requiring that at least two small-R jets in the event are c-tagged, with at least one of them being tightly c-tagged. The signal and control regions are defined with the same ΔR cut as for the $VH, H \to b\bar{b}$ resolved topology and events are classified as well in the 2- and 3-jets categories. As a result the Higgs candidate system can either be composed of 2 tightly c-tagged jets (TT signal region), or one tightly c-tagged and one loosely c-tagged jet (TL signal region) as shown in Figure 5.7a: the TT and TL categories are kept separated and are not merged. For the 3-jets category, jets are first ordered by their tagging priority (T > N > B) *i.e.* tight c-tagged jets have the highest priority while b-tagged ones have the lowest priority. Jets within the same tagging priority category are then p_{T}-ordered *i.e.* the leading jet in that category has the highest priority. The two jets in the event with the highest overall priority are forming the Higgs-candidate system. As a consequence, in the 3-jets category the TT (respectively TL) regions gather events corresponding to the following flavour tagging triplets TTT, TTL and TTN (respectively TLL and TLN). The TTB and TLB flavour tagging triplets are not considered because a b-veto is applied in signal regions for the additional jet. Extending the signal region to the tight-untagged (TN) region is currently under study as it gives a 7% sensitivity increase when performing statistic only fits (*i.e.* taking only into account statistical uncertainties in the fit). In the 3-jets category, the TNB region (abusively referred to as CNB) is used as a top control region which is shared also with the $VH, H \to b\bar{b}$ analysis in order to constrain in particular the *top(lq)* component with *lq* being the true flavour of the 2 c-tagged jets with *q* standing for c- or light-jet and l for light-jet. Finally, the untagged regions (NN) for the 2-jets and 3-jets categories are enriched in $V + ll$ components and are used as control regions to normalize the $V + ll$ contribution.

In the previous $VH, H \to c\bar{c}$ publication [6], for the 3-jets category, the Higgs candidate system was chosen as the two leading jets of the events (*i.e.* jets with the highest p_{T}) and then only those two jets were required to satisfy the c-tagging requirements without any consideration for the third jet. The selection of jets as the Higgs candidate system has changed because previously the c-tagging efficiency was relatively low and mistag rates was high while the improved performances of the 2D b/c-tagger allow now for prioritizing jet flavour tagging over their transverse momentum.

The orthogonality between the 3 analyses is ensured in the following way:

- **$VH, H \to b\bar{b}$ resolved and boosted orthogonality**: following two similar studies performed respectively in the 0- and 1-lepton channels concerning the combination of resolved and boosted events (study presented in Section 6.2), it was decided to keep the strategy used for the previous combination of those topologies [5]: events are reconstructed with the resolved topology if the reconstructed transverse momentum of the vector boson $p_{\mathrm{T}}^{V} < 400$ GeV while they are reconstructed with the boosted topology if $p_{\mathrm{T}}^{V} > 400$ GeV. This simple to implement combination strategy provides a sensitivity close to that of significantly more complex approaches.
- **Resolved $VH, H \to b\bar{b}$ and $VH, H \to c\bar{c}$ orthogonality**: the Figure 5.7a presents the regions of interest for the $VH, H \to b\bar{b}$ and $VH, H \to c\bar{c}$ analyses: the 2 b-tagged (BB) events are used for the $VH, H \to b\bar{b}$ analysis while the 1 and 2 c-tagged regions are used for $VH, H \to c\bar{c}$: the 2 analyses are by construction fully orthogonal.
- **Boosted $VH, H \to b\bar{b}$ and resolved $VH, H \to c\bar{c}$ orthogonality**: a remaining overlap, shown in Figure 5.7b exists between the boosted $VH, H \to b\bar{b}$ analysis and the resolved $VH, H \to c\bar{c}$ analysis as VR track-jets are used for b-tagging for the boosted analysis instead of small-R jets. The orthogonal division of the phase space has not yet been decided, there are several options such as either vetoing events with the leading large-R jet passing the $VH, H \to b\bar{b}$ b-tagging requirements or a more "drastic" solution would be to only perform the $VH, H \to c\bar{c}$ measurement below $400\,\mathrm{GeV}$ as for $p_{\mathrm{T}}^{V} > 400\,\mathrm{GeV}$ not much $VH, H \to c\bar{c}$ signal events are expected, and the sensitivity is expected to be low especially because the Higgs candidate is only reconstructed with the resolved topology for that analysis. The latest option is thus foreseen to be selected.

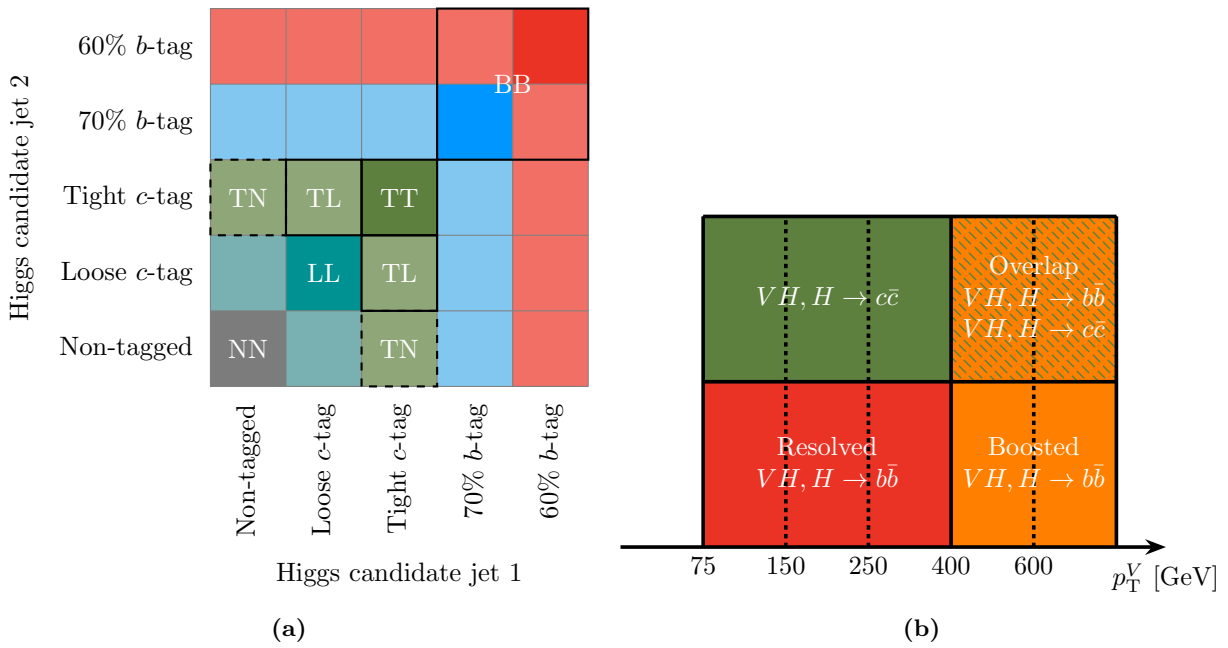


Figure 5.7: Combination of the $VH, H \to b\bar{b}$ resolved, $VH, H \to b\bar{b}$ boosted and $VH, H \to c\bar{c}$ analyses. (a) Resolved $VH, H \to b\bar{b}$ and $VH, H \to c\bar{c}$ Higgs jet candidates tagging requirements, their respective signal region are represented in solid black lines. Exactly two b-tagged jets (BB region) are required for the $VH, H \to b\bar{b}$ resolved analysis. While the $VH, H \to c\bar{c}$ analysis is requiring at least two small-R jets in the event to be c-tagged, with at least one of them being tightly c-tagged. The TT and TL regions are considered separately as signal regions for $VH, H \to c\bar{c}$, and an extension of the signal region to the 1 c-tagged TN region (represented by dashed line) is currently under study. (b) Presentation of the region of phase space of the 3 analyses. By construction of the resolved $VH, H \to b\bar{b}$ and $VH, H \to c\bar{c}$ signal and control regions and the use of the 2D b/c-tagger, the resolved $VH, H \to b\bar{b}$ and $VH, H \to c\bar{c}$ measurements are orthogonal. Thanks to the kinematic division of phase space, the $VH, H \to b\bar{b}$ resolved and boosted measurements are orthogonal as the former is performed for $p_{\mathrm{T}}^{V} < 400$ GeV while the latter for $p_{\mathrm{T}}^{V} > 400$ GeV. The remaining overlap between the boosted $VH, H \to b\bar{b}$ and the $VH, H \to c\bar{c}$ measurement needs to be lifted but at the moment the strategy has not been finalized yet.

5.3.2 Common selections for the resolved and boosted topologies

Electrons should fulfill a *loose* selection, they are required:

- to have $p_{\mathrm{T}} > 7$ GeV and $|\eta| < 2.47$,
- to have a small longitudinal impact parameter $|\Delta z_0| \sin(\theta) < 0.5$ mm and a transverse impact parameter significance $(d_0/\sigma(d_0))^{\mathrm{BL}} < 5$ (evaluated considering information of the B-Layer (IBL)) in order to reject contamination from pile-up.
- To meet the *LooseLH* electron identification criteria.
- To fulfill a recently developed loose isolation working point (WP): *Loose_VarRad* [313] (similar to the *Loose* isolation WP in Table 3.1) requiring $E_{\mathrm{T}}^{\mathrm{cone20}}/p_{\mathrm{T}} < 0.20$ and $p_{\mathrm{T,TTVA}}^{\mathrm{cone30}}/p_{\mathrm{T}} < 0.15$, with "TTVA" standing for track-to-vertex association indicating that the track was used in the vertex fit.

In the 1-lepton channel, *tight* electrons are selected using the *TightLH* electron identification criterion and the *HighPtCaloOnly* calorimeter-based isolation (presented in Table 3.1) in addition to the *Loose_VarRad* isolation.

Muons are also required to pass a *loose* selection, they should:

- be within the acceptance of the muon spectrometer $|\eta| < 2.7$,
- have $p_{\mathrm{T}} > 7$ GeV,
- have a small longitudinal impact parameter $|\Delta z_0| \sin(\theta) < 0.5$ mm and a transverse impact parameter significance $(d_0/\sigma(d_0))^{\mathrm{BL}} < 3$ is required in order to reject contamination from pile-up and cosmic rays.
- Satisfy the *Loose* quality criterion.
- Pass the *Loose_VarRad* isolation requirement.

In the 1-lepton channel, *tight* muons fulfill the *Medium* quality criterion have $|\eta| < 2.5$, and the stricter *HighPtTrackOnly* track isolation requirement in addition to the *Loose_VarRad* isolation.

Hadronically decaying τ-leptons are required to:

- have a $p_{\mathrm{T}} > 20$ GeV, $|\eta| < 2.5$, and be outside of the transition region between the barrel and endcap electromagnetic calorimeters $1.37 < |\eta| < 1.52$.
- Meet the *Loose* RNN tau identification quality criterion which has approximately a 10% τ-identification efficiency increase and a similar jet misidentification with respect to the previously used *Medium* BDT quality criterion. The reconstructed τ-leptons are not directly used in the event selection (they are not considered as jets), but they are taken into account in the missing transverse momentum computation and are also used to avoid double-counting hadronic τ-leptons as other objects. As a result τ-jets are better identified and are less misidentified as jets, events will have less extra jets leading to their migration to lower multiplicity jet regions especially in the 0- and 1- lepton-channels where semileptonic $t\bar{t}$ events are expected to be affected by that change as shown in Figure 5.8. Due to that switch, either a veto of events containing at least one τ-jet or using the number of τ-jets as input variable of the multivariate discriminant used in the analysis is being studied to mitigate the increase of $t\bar{t}$ background. On the other hand, the $WH, H \to b\bar{b}$ signal contribution in the 0-lepton channel is found to be slightly increased by 5% as presented in Figure 5.9. Indeed, the WH events represent around 15% to 20% of signal events in the 0-lepton channel because the lepton from the W-decay can be misidentified as a jet (in particular τ-leptons) or be undetected due to transition regions in the detector for instance. Moreover, with a better τ-identification such $WH \to \tau\nu b\bar{b}$ events are reconstructed with less extra-jets and thus mimic the $ZH \to \nu\nu b\bar{b}$ signature.

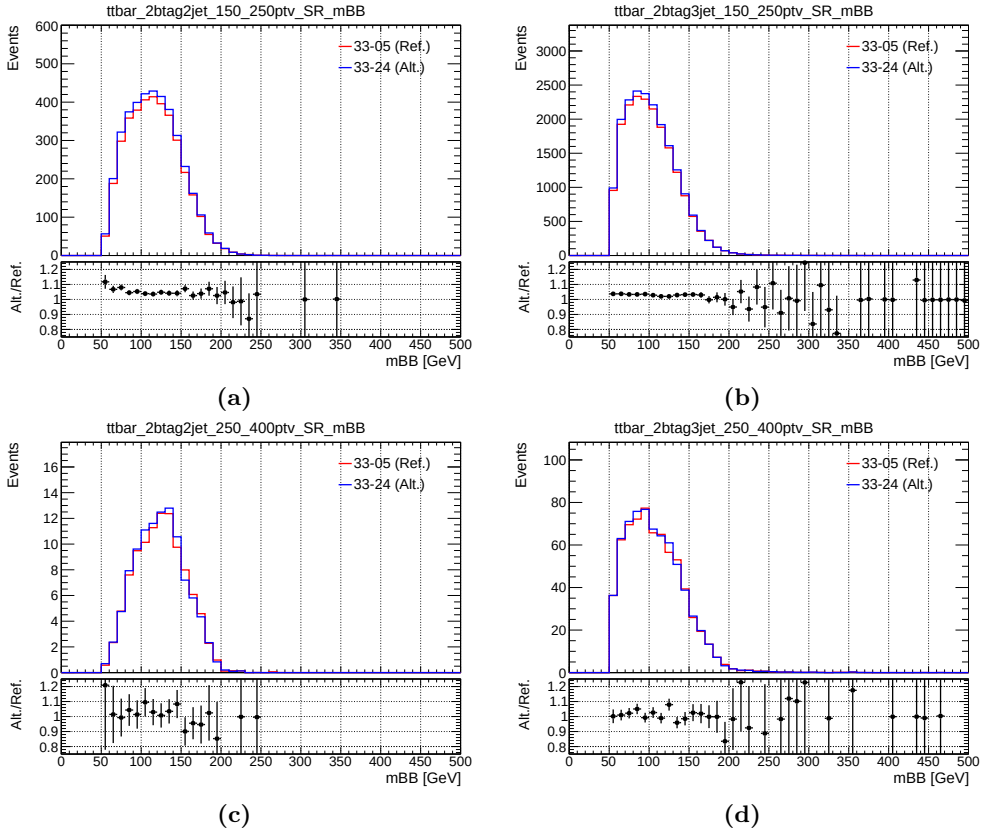


Figure 5.8: Comparison of the RNN *Loose* WP (blue ("33-24") histograms) versus the BDT *Medium* WP (red ("33-05") histograms) on the reconstructed invariant mass (m_{bb}) of the two b-tagged small-R jets misidentified as Higgs candidate for the $t\bar{t}$ background in the 1-lepton channel as predicted by the nominal simulation. The Figures (a) and (c) present the m_{bb} distribution for the 2 b-tagged 2-jets SR while (b) and (d) present the 2 b-tagged 3-jets SR. The distributions are plotted (a) and (b) for the $[150, 250]$ GeV p_{T}^{V}-region, and (c) and (d) $[250, 400]$ GeV p_{T}^{V}-region. In the bottom panel the ratio of the RNN over the BDT distributions is shown.

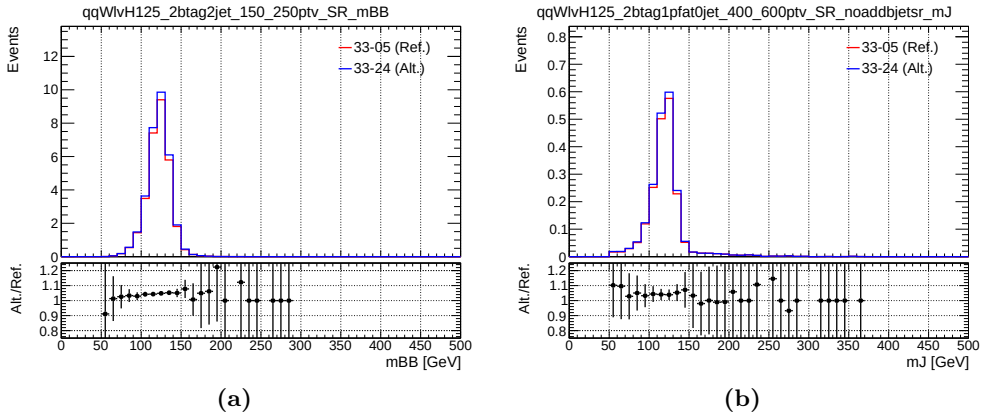


Figure 5.9: Comparison of the RNN *Loose* WP (blue ("33-24") histograms) versus the BDT *Medium* WP (red ("33-05") histograms) on the reconstructed invariant mass (m_{bb} or m_{J}) of the two small-R jets or the leading large-R jet Higgs candidate system for the $WH, H \to b\bar{b}$ events in the 0-lepton channel as predicted by the nominal simulation. (a) The m_{bb} distribution for the 2 b-tagged 2-jets SR in the $[150, 250]$ GeV p_{T}^{V}-region is shown while (b) presents the m_{J} distribution in the high purity SR in the $[400, 600]$ GeV p_{T}^{V}-region. In the bottom panel the ratio of the RNN over the BDT distributions is shown.

Small-R jets are divided in two categories: the *signal jets* which are central jets ($|\eta| < 2.5$), with $p_{\mathrm{T}} > 20\,\mathrm{GeV}$ (respectively $p_{\mathrm{T}} > 30\,\mathrm{GeV}$) for the resolved topology (for the boosted topology) and *forward jets* emitted in forward regions ($2.5 < |\eta| < 4.5$) of the ATLAS detector, outside of the

inner detector acceptance and are required to have $p_{\mathrm{T}} > 30$ GeV. Small-R jets should pass the jet cleaning procedure: the jet vertex tagger is used to remove small-R jets with $p_{\mathrm{T}} < 120$ GeV and $|\eta| < 2.5$ that are identified as not being associated with the primary vertex of the hard interaction.

Large-R jets are required to have $|\eta| < 2.0$ and $p_{\mathrm{T}} > 250$ GeV as it corresponds to the lowest possible p_{T}-threshold for being able to capture the two b-hadrons from $H \to b\bar{b}$ decays within a single large-R jet of radius $R = 0.1$ according to the Equation (5.1.1). Track-jets are selected requiring $p_{\mathrm{T}} > 10$ GeV and $|\eta| < 2.5$.

In addition to the standard jet calibration described in Section 3.6, small-R jets and large-R jet Higgs candidates receive residual corrections to improve their energy measurement (scale and resolution): if any muons are found within a p_{T}-dependent cone around the jet axis, the four-momentum of the closest muon is added to that of the jet, it is called the *OneMu* correction to account for heavy flavour semi-leptonic decays. Indeed, the leptonic and semi-leptonic b-quark branching ratio is of the order of 42% including electron and muon contributions: therefore, b-jets are expected to contain a muon approximately 20% of the times. Afterwards, a residual correction, referred to as *PtReco* correction, is applied to small-R jets to equalize the response to jets with leptonic or hadronic decays of heavy-flavour hadrons and to correct for resolution effects. Such corrections improve the resolution of the dijet mass by up to 17% for the resolved and the boosted topologies [65, 275, 310]. In the 2-lepton channel, instead of the *PtReco* correction, a kinematic likelihood fit, detailed in Section 5.3.7, is used both for small-R jets or the leading large-R jet (depending on the topology) exploiting the complete reconstruction of all final-state objects. Such correction is much more precise and largely improves the estimate of the energy of the Higgs jet candidates both for the resolved and boosted topologies as shown in Figure 5.12.

5.3.3 Specific selections for the resolved topology

For the resolved $VH, H \to b\bar{b}$ analysis, events are required to have exactly two b-tagged small-R jets which constitute the Higgs boson candidate with at least one of them having a p_{T} greater than 45 GeV. They are identified with the 70% b-tagging efficiency DL1r algorithm (efficiency measured in $t\bar{t}$ reference sample). For that same reference sample and b-tagging efficiency, the DL1r and MV2c10 taggers are found to have a c-jet rejection respectively of 10.6 and 8.6 while the light-jet rejection is measured to be equal to 441 and 258 [209]. As a consequence an increase of the rejection of c- and especially of light-jets is expected for Legacy Run 2 analysis in the resolved $VH, H \to b\bar{b}$ regime.

In the 0- and 1-lepton channels, only events with at most one additional un-tagged jet are allowed in order to suppress the $t\bar{t}$ background contribution and are classified in two separate event regions: the 2 b-tagged 2-jets and 2 b-tagged 3-jets categories. A study for a possible extension to a 4-jets category is currently ongoing especially for the 0-lepton channel where a 2σ significance is obtained from statistic only fit as the $t\bar{t}$ contamination for that channel and category is still reasonably small. In those two channels, the phase space is divided in two p_{T}^{V}-regions: the $[150, 250]$ and $[250, 400]$ GeV p_{T}^{V}-bins. Due to the limited amount of data available in the 2-lepton channel, events are categorized in the 2-jets and $\geq$ 3-jets categories. The extension from the 3-jets to the $\geq$ 3-jet category is possible because the 2-lepton channel is relatively pure in signal and the signal acceptance is increased in this category by 100% [65]. The phase space in the 2-lepton channel is divided in that case in 3 bins: the $[75, 150]$, the $[150, 250]$ and $[250, 400]$ GeV p_{T}^{V}-regions. The p_{T}^{V}-bound is lowered only for the 2-lepton channel because for $p_{\mathrm{T}}^{V} < 150$ GeV the multi-jet contamination is significantly increasing in the 0- and 1-lepton channels.

Since the angular distance ($\Delta R(b, b)$ or $\Delta R(c, c)$) between b-jets (or c-jets) is expected to decrease with the Higgs transverse momentum (Equation (5.1.1)) equivalently with p_{T}^{V}, lower and upper bounds $\Delta R(b, b)$ as a function of p_{T}^{V} are defined and delimit the central signal region (SR) from

the two control regions (CRs) called low and high ΔR CRs as shown in Figure 5.10 which are respectively enriched in V + jets and $t\bar{t}$ events for the $VH, H \to b\bar{b}$ analysis as top quarks are emitted "back to back" in $t\bar{t}$ events *i.e.* top-quarks are produced in opposite directions: hence the angular separation between the two b-jets is expected to be large. For the $VH, H \to c\bar{c}$ analysis, the high ΔR region is enriched in $V + cc$ and $V + cl$ events respectively for the TT and TL tagged regions. The lower and upper bounds are derived separately for each n-jet region from the

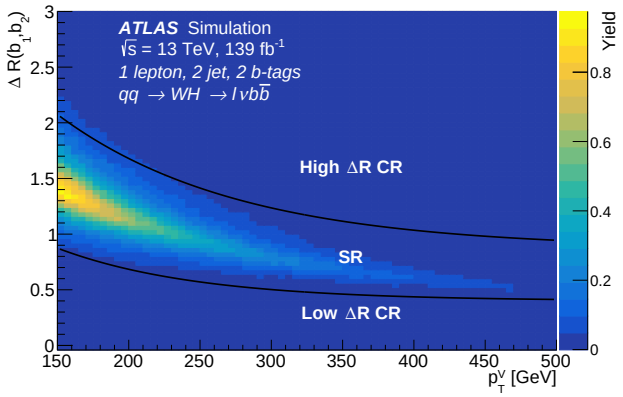


Figure 5.10: $WH, H \to b\bar{b}$ signal yield distribution of the angular distance ($\Delta R(b,b)$) as a function of p_{T}^{V} in the 1-lepton channel for the 2 b-tagged 2-jets events. The black lines delimit the continuous lower and upper selection on $\Delta R(b,b)$ used to categorize the events into the signal and control regions (taken from Ref. [65]).

$VH, H \to b\bar{b}$ analysis, and the side-band found are also adopted in the $VH, H \to c\bar{c}$ analysis:

- the high ΔR bound is defined thanks to the 2-lepton channel such that the high ΔR CR contains 5% (respectively 15%) of the signal for the 2-jets (respectively 3-jets or 4-jets or $\geq$ 5-jets) regions.
- The low ΔR bound is defined such that the low ΔR CR contains 10% of the diboson events in the 1-lepton channel, to ensure that a sufficient number of these events remain in the signal region when conducting the diboson validation analysis.

The ΔR selections derived in the 1-lepton and 2-lepton channels are applied in the other channels. The high ΔR cut is tighter in the 3-jets category than for the 2-jets in order to remove more $t\bar{t}$ events from the signal region especially in the 1-lepton channel. For the Run 2 Legacy publication it was decided that the low ΔR CR will be only kept in the 1-lepton channel for $VH, H \to b\bar{b}$ while it will be merged in all other channels with the signal region: it enables to control the normalization of the $V + bb$ (especially $W + bb$) contribution which is dominant in that region. The high ΔR regions are kept in all channels. For $VH, H \to c\bar{c}$ the low ΔR CR is always merged with the SR.

The fit-function used to define those two continuous boundaries is expressed as:

$$f(p_{\mathrm{T}}^{V}) = a \times \exp(b + c \times p_{\mathrm{T}}^{V}), \tag{5.3.3}$$

with the coefficients (a, b, c) for each jet-category associated to the low and high ΔR regions being reported in Table 5.4. Additional selections to enhance signal from background discrimination are described in Table 5.3. The invariant mass of the two b-tagged or c-tagged jets constituting the Higgs candidate is respectively denoted m_{bb} or m_{cc}. The invariant mass cut $m_{bb} > 50$ GeV was introduced to avoid a known mismodelling in the simulation below 50 GeV.

		a	b	c [GeV^{-1}]
Low ΔR	2-jets	0.410	0.818	-0.0106
	3-jets	0.430	0.399	-0.00931
	4-jets	0.411	0.347	-0.00774
	$\geq$ 5-jets (2L only)	0.501	-0.119	-0.00753
High ΔR	2-jets	0.787	1.387	-0.00700
	3-jets	0.684	1.204	-0.00600
	4-jets	0.863	0.984	-0.00408
	$\geq$ 5-jets (2L only)	1.667	0.519	-0.00504

Table 5.4: Coefficient a, b and c that determines the low and high ΔR regions as a function of p_{T}^{V} for the 3-channels in the resolved $VH, H \to b\bar{b}$ and $VH, H \to c\bar{c}$ analyses. The fit-function used to define those two boundaries is the following: $f(p_{\mathrm{T}}^{V}) = a \times \exp(b + c \times p_{\mathrm{T}}^{V})$. The $\geq$ 5-jets coefficients are only used for the 2-lepton channel, while the 4-jets ones may be used in other channels in case a 4-jets region is introduced for $VH, H \to b\bar{b}$ (on-going study).

5.3.4 Specific selections for the boosted topology

In the boosted analysis, events are required to have at least one large-radius jet with the leading large-R jet of the event (denoted J) being taken as the Higgs candidate. The leading large-R jet is required to have at least two ghost associated track-jets, and should pass the b-tagging *85% WP leading 3* requirement meaning that it should contain exactly 2 b-tagged track-jets among its 3 leading matched track-jets (or 2 leading ones if it contains only 2 track-jets) using the 85% WP DL1r tagging algorithm. In previous analysis rounds [5, 275], the *70% WP leading 2* strategy was used (with the MV2c10 tagger): the 2 leading track-jet were required to be b-tagged with the 70% MV2c10 WP. To avoid the ambiguous cases of concentric jets, events where the b-tagged VR track-jets overlap with other VR track-jets are removed. The overlapping VR track-jet being defined as $\Delta R/R_{\mathrm{s}} < 1$ (where ΔR corresponds to the distance among any pair of VR track-jets and R_{s} corresponds to the smaller radius of the considered pair).

A study with the multivariate discriminant used in the $VH, H \to b\bar{b}$ boosted analysis has been conducted to optimize the tagging strategy (either the leading 2 or leading 3 strategies) and the b-tagging working point (out of the 70%, the 77% or the 85% WPs). It led to the choice of the *85%WP leading 3* strategy as optimal setting because the significance is increased with respect to the 70% WP leading 2 one and the amount of data available in distributions both for signal and background is increased which is needed for fits in order to reduce statistical fluctuations and to properly normalize backgrounds by means of control regions. When comparing significances obtained with the leading 3 and leading 2 strategy for a same WP, they are similar in the 0- and 1-lepton channels while it is 5% lower in the 2-lepton channel for the leading 3 strategy compared to the leading 2 one. However, the leading 3 strategy was found to increase by around 15% the signal yields in the 0- and 1-lepton channels while it remains constant for the 2-lepton channel compared to the leading 2 strategy. As a result the leading 3 strategy was chosen. Concerning the b-tagging working point for the leading 3 strategy, the 85% WP is of course increasing statistics compared to the tighter 70% and 77% WPs, it was found that the significances are 3 to 5% higher than for the 70% and 77% WPs. Hence, the 85% WP was adopted.

For a $t\bar{t}$ reference sample, the DL1r and MV2c10 85% b-tagging WP are found to have a similar c-jet rejection of 3, while the light-jet rejections is respectively found to be of 56 against 31 [210]. As a result an increase of the rejection of light-jets is expected for Legacy Run 2 analysis in the boosted $VH, H \to b\bar{b}$ regime.

In the 0- and 1-lepton channels events are categorized according to the number of additional small-R jets with $p_{\mathrm{T}} > 30$ GeV in the event not matched to the large-R jet (matching performed with $\Delta R < 1$ requirement). If no additional small-R jet is present, the event falls into the so-called “high

purity signal region" (HP SR) otherwise it falls into the "low purity signal region" (LP SR). For the 2-lepton channel the HP SR, LP SR and the $t\bar{t}$ control regions are indeed merged together into a single signal region due to the limited number of data events available and the subdominant role of the $t\bar{t}$ background in this channel. In addition, in the 0- and 1-lepton channels a high purity $t\bar{t}$ control region is defined if there is at least one additional b-tagged track-jet, which is not associated to the Higgs candidate jet: this region targets events where the large-R jet is reconstructed from one of the top quark and the additional b-tagged track-jet comes from the other top quark since top quarks are emitted back to back in the transverse plane in $t\bar{t}$ events. In the three channels, events are categorized according to the reconstructed p_{T}^{V} in 2 bins: the $[400, 600]$ and $p_{\mathrm{T}}^{V} > 600$ GeV regions. Finally, the choice of $p_{\mathrm{T}}^{V} = 400$ GeV is motivated by the ability to capture the two b-jets from the Higgs boson decay inside the large-radius jet as their angular separation decreases with increasing p_{T}^{V}.

Additional selections to improve signal from background discrimination are described in Table 5.3. The Higgs candidate invariant mass is denoted m_J. As for the resolved topology, the invariant mass cut $m_J > 50$ GeV was introduced to avoid a known mismodelling in the simulation below 50 GeV.

5.3.5 Specific selections for the 0-lepton channel

For the 0-lepton channel, events with any *loose* lepton are rejected.

The $E_{\mathrm{T}}^{\mathrm{miss}}$ triggers [314] are used for the online selections with thresholds that varied from 70 GeV to 110 GeV between 2015 and 2018. Their efficiency is measured in W + jets, Z + jets and $t\bar{t}$ events using single-muon triggered data as at trigger level the muons are not included for the trigger $E_{\mathrm{T}}^{\mathrm{miss}}$ calculation. Such triggers allow to efficiently select events with large trigger-level $E_{\mathrm{T}}^{\mathrm{miss}}$ values. The resulting trigger correction factors applied to the simulated events are ranging from 0.95 for the offline $E_{\mathrm{T}}^{\mathrm{miss}}$ around 150 GeV and are close to unity for $E_{\mathrm{T}}^{\mathrm{miss}}$ values above 200 GeV.

High $E_{\mathrm{T}}^{\mathrm{miss}}$ multi-jet events which comes from mismeasured jets in the calorimeters are efficiently removed thanks to requirements on the angular separation of the $\boldsymbol{E}_{\mathbf{T}}^{\mathbf{miss}}$, jets, and $\boldsymbol{p}_{\mathbf{T}}^{\mathbf{miss}}$ as for example the cut $\min[\Delta\phi(\boldsymbol{E}_{\mathbf{T}}^{\mathbf{miss}}, \text{small-R jets})] > 20°$ (2-jets), $> 30°$ (3-jets) which removes a large fraction of the multi-jet background as shown in Figure 5.11. The reason for that angular selection is that, fake $E_{\mathrm{T}}^{\mathrm{miss}}$ can be due to a miscalibrated jet in a multijet event: in that case the $\boldsymbol{E}_{\mathbf{T}}^{\mathbf{miss}}$ will be aligned with the miscalibrated jet since it will correspond to the difference between the true and the miscalibrated reconstructed jet energy. The selection $\Delta\phi(\boldsymbol{E}_{\mathbf{T}}^{\mathbf{miss}}, \boldsymbol{p}_{\mathbf{T}}^{\mathbf{miss}}) < 90°$ is applied because the angular distance between $\boldsymbol{E}_{\mathbf{T}}^{\mathbf{miss}}$ and $\boldsymbol{p}_{\mathbf{T}}^{\mathbf{miss}}$, is expected to be randomly distributed for fake $E_{\mathrm{T}}^{\mathrm{miss}}$, while for true $E_{\mathrm{T}}^{\mathrm{miss}}$ from non-interacting particles $E_{\mathrm{T}}^{\mathrm{miss}}$ and $p_{\mathrm{T}}^{\mathrm{miss}}$ which are respectively reconstructed with calorimeter- and track-based information should be close to each other. Some of these variables are also used as input for the multivariate discriminant as described in Section 5.6.

Finally, for the resolved topology, a requirement on the scalar sum of the transverse momenta of the jets (H_{T}) removes a small part of events (less than 1%) to avoid a region of phase space where the trigger efficiency depends on the number of jets in the event.

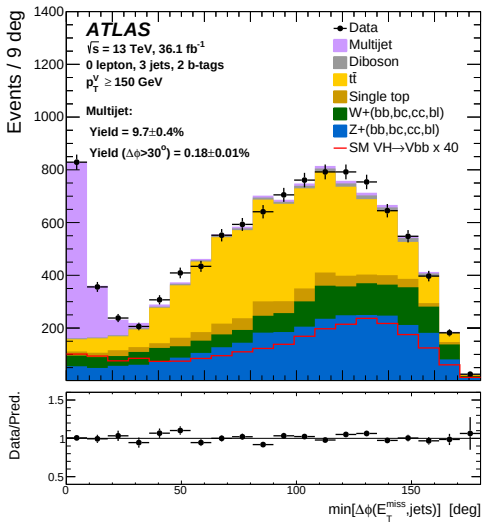


Figure 5.11: Distribution of the minimal azimuthal distance ($\min[\Delta\phi(\boldsymbol{E}_{\mathbf{T}}^{\mathbf{miss}}, \text{small-R jets})]$) between $E_{\mathrm{T}}^{\mathrm{miss}}$ and the small-R jets in the 0-lepton 2 b-tagged 3-jets category for 2015+2016 only ATLAS data-taking. The $\min[\Delta\phi(\boldsymbol{E}_{\mathbf{T}}^{\mathbf{miss}}, \text{small-R jets})] > 30°$ cut applied to the 3-jets category enables to remove most of the multi-jet events as after that cut the multi-jet background represents only $0.18\% \pm 0.01\%$ of the total yield compared to $9.7\% \pm 0.4\%$ before (taken from Ref. [310]).

5.3.6 Specific selections for the 1-lepton channel

For the 1-lepton channel, events are required to have exactly one *tight* lepton and no additional *loose* leptons.

In the electron sub-channel, the *tight* electron is required to have $p_{\mathrm{T}} > 27\,\mathrm{GeV}$. Events are required to satisfy a logical OR of single-electron triggers [236] with p_{T} thresholds that started at 24 GeV in 2015 and increased to 26 GeV in 2016-2018[2]. In addition, the $E_{\mathrm{T}}^{\mathrm{miss}} > 30\,\mathrm{GeV}$ cut is required in order to reduce the multi-jet background.

For the muon sub-channel, the *tight* muon is required to have $p_{\mathrm{T}} > 25$ GeV. The same $E_{\mathrm{T}}^{\mathrm{miss}}$ triggers [314] and correction factors as for the 0-lepton channel are used. As muons are not included in the $E_{\mathrm{T}}^{\mathrm{miss}}$ computation at trigger level, these triggers efficiently selects the $WH \to \mu\nu b\bar{b}$ signal events. Furthermore, they achieve higher performances than the single-muon triggers [237] which have a lower efficiency because of the more limited coverage of the muon trigger system in the central region.

5.3.7 Specific selections for the 2-lepton channel

For the 2-lepton channel, events are required to have exactly two *loose* leptons. The leptons should have the same flavour, the leading lepton should satisfy $p_{\mathrm{T}} > 27\,\mathrm{GeV}$ and the invariant mass of the pair must be close to the expected Z boson mass.

For the muon sub-channel, the muons are required to have opposite charges. An overlap removal of single-muon triggers [237] is used, with lowest p_{T} thresholds increasing from 2016–2018 and ranging from 20 GeV to 26 GeV.

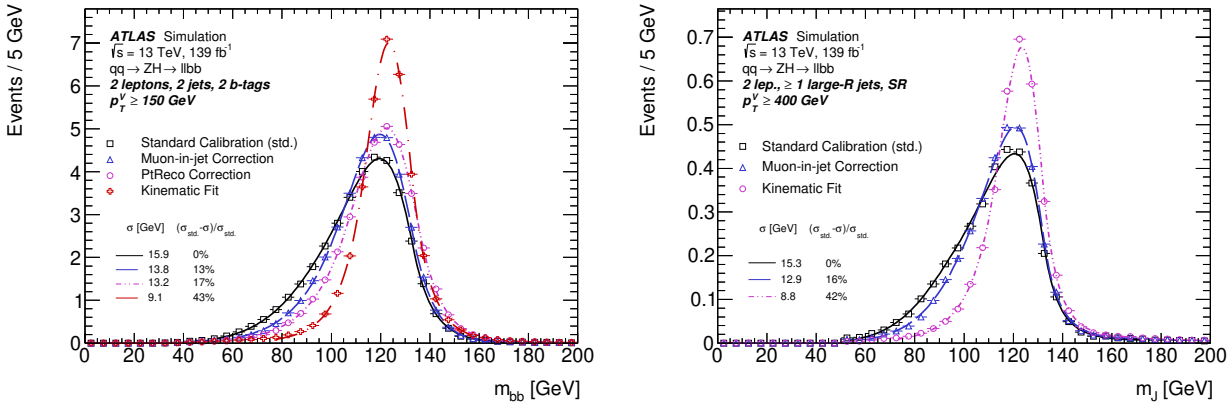
For the electron sub-channel, the opposite charges condition is not required because of the non-negligible electron charge misidentification rate. The trigger selection in that sub-channel is the

2. Additional identification and isolation requirements are applied to the trigger object to allow a low p_{T}-threshold to be maintained throughout Run 2.

same as in the 1-lepton channel.

For the resolved topology, a high purity *top*-$e\mu$ control region containing 99% of $t\bar{t}$ and single top Wt events is obtained by requiring exactly one electron and one muon instead of the same flavour condition. This $e\mu$-CR contains top events with two leptonically decaying W bosons: the CR is used to directly model the shape and normalization of the same flavour top background event using data which enables to eliminate all the experimental and theoretical uncertainties concerning top background thanks to this data-driven method.

For the resolved and boosted topologies, the Higgs candidate jets or large-R jet are corrected with a kinematic fit (KF) [315, 316] which exploits the conservation of the transverse momentum. Typically, the resolution of small-R jets or large-R jet is of the order of 10% while the energy resolution for the leptons is of the order of 1%. By means of a fit which constrains the $ZH \rightarrow l^+l^-b\bar{b}$ candidate system to be balanced (considering also the third jet for ≥ 3 jets events), taking into account constraints on the dilepton invariant mass, the transverse momentum, angular position (and their uncertainties) of the lepton and jets (for jets, transverse momenta are constrained by transfer functions predicting the true jet transverse momenta based on their reconstructed values), the four-momentum of each jet is corrected. The kinematic fit is applied after the *OneMu* correction which is accounting for muon in jets. Combining the muon in jet and kinematic fit corrections improves by approximately 40% the invariant mass resolution of the Higg candidate system both for the resolved and boosted topologies as illustrated in Figure 5.12.



(a) **(b)**

Figure 5.12: (a) m_{bb} and (b) m_J invariant masses distributions before and after additional corrections are applied in the resolved and boosted topologies for $qq \rightarrow ZH \rightarrow l^+l^-b\bar{b}$ events for the 2-lepton channel. The distributions are fit with a Bukin function [317] and the resolution values (σ) reported correspond to the width of the fitted function. Both for the resolved and boosted topologies, the kinematic fit (KF) correction is applied on b-tagged small-R and large-R jets after the *OneMu* correction. For the resolved topology, the *PtReco* correction is shown in Figure (a) simply for comparison with the KF corrections. Indeed, the *PtReco* correction (which is applied after the *OneMu* correction) is not used in the 2-lepton channel: this correction is only applied in the 0- and 1- lepton whereas in the 2-lepton channel it is replaced by the kinematic fit (taken from Refs. [65, 275]).

5.3.8 Specific $VH, H \rightarrow c\bar{c}$ selections

With the 2D b/c-tagger implementation presented in Figure 5.6, an inclusive c-tagging efficiency of 40% is achieved combining the loose and tight c-tagging WPs (estimated in $t\bar{t}$ reference sample), while the mistagging efficiencies are found to be of 10% for b-jets and 6.6% for light-jets. In the previous round of the $VH, H \rightarrow c\bar{c}$ analysis (the DL1r c-tagger was used against the MV2c10 b-tagger) the c-taggging efficiency, and c- and light-jet mistagging efficiencies were respectively equal to 27%, 8% and 1.6% [6]. As a higher c-tagging efficiency is expected for the Legacy Run 2

analysis and since the increase of the mistagging efficiency for b- and light-jets remains reasonable, it implies that an improvement of the sensitivity to $H \to c\bar{c}$ decays is expected.

The $VH, H \to c\bar{c}$ event selection is pretty similar to the $VH, H \to b\bar{b}$ resolved event selection replacing the b-tagging condition by the c-tagging condition, the signal and control regions of the $VH, H \to b\bar{b}$ resolved analysis are also used in the $VH, H \to c\bar{c}$ analysis but there is no low ΔR CR as it is always merged to the SR for the $VH, H \to c\bar{c}$ analysis. However, this analysis differs from $VH, H \to b\bar{b}$ for three points:

- the $VH, H \to c\bar{c}$ tagging requirement is not the exact equivalent as for the $VH, H \to b\bar{b}$ analysis. For the resolved $VH, H \to b\bar{b}$ analysis, exactly 2 b-tagged jets are required, while for $VH, H \to c\bar{c}$, at least 2 c-tagged jets are required to be found in the event with at least one of them being tightly c-tagged. The signal regions are further divided in the tight-tight (TT) and tight-loose (TL) categories.
- For the 3-jets category, jets are first ordered by their tagging priority (T > N > B) then ranked by p_{T}within each tagging priority categories. The two jets in the event with the highest overall priority are forming the Higgs-candidate system. The TT (respectively TL) regions gather events corresponding to the following flavour tagging triplets TTT, TTL and TTN (respectively TLL and TLN).
- The $VH, H \to c\bar{c}$ measurement is performed inclusively for $p_{\mathrm{T}}^{V} > 150\,\mathrm{GeV}$ in the 0-lepton and 1-lepton channels and in two p_{T}^{V}-regions for the 2-lepton channel: $[75, 150]$ and $p_{\mathrm{T}}^{V} > 150\,\mathrm{GeV}$ regions.

For the 0- and 1-lepton channels, in the 2-jets and 3-jets categories, respectively the CB and CNB regions (more precisely the TB and TNB regions) are used as a top control regions which are shared also with the $VH, H \to b\bar{b}$ analysis in order to constrain the *top(lq)* and the *top(bq)* components with *q* (respectively *l*) standing for c- or light-jet (respectively for light-jet) true flavour as shown in Figure 5.13. Indeed, the *top(bc)* component is one of the major background and is signal-like for both analyses: it is thus important to control it with data. Typically, the *top(lq)* events corresponds to c-tagging the two jets from the top hadronic W decay ($t \to Wb \to qqb$) while the *top(bq)* events corresponds to c-tagging the b-jet and a light- or c-jet coming from the top hadronic decay. Hence, the invariant mass for *top(lq)* events is expected to peak around the W mass while the *top(bq)* invariant mass is expected to be larger. Finally, in all channels, the NN-region for the 2-jets and 3-jets categories are enriched in $V + ll$ components and are used as control regions to normalize the $V + ll$ contribution.

An extension of the signal region to the 1 c-tagged tight-untagged (TN) region is currently under study.

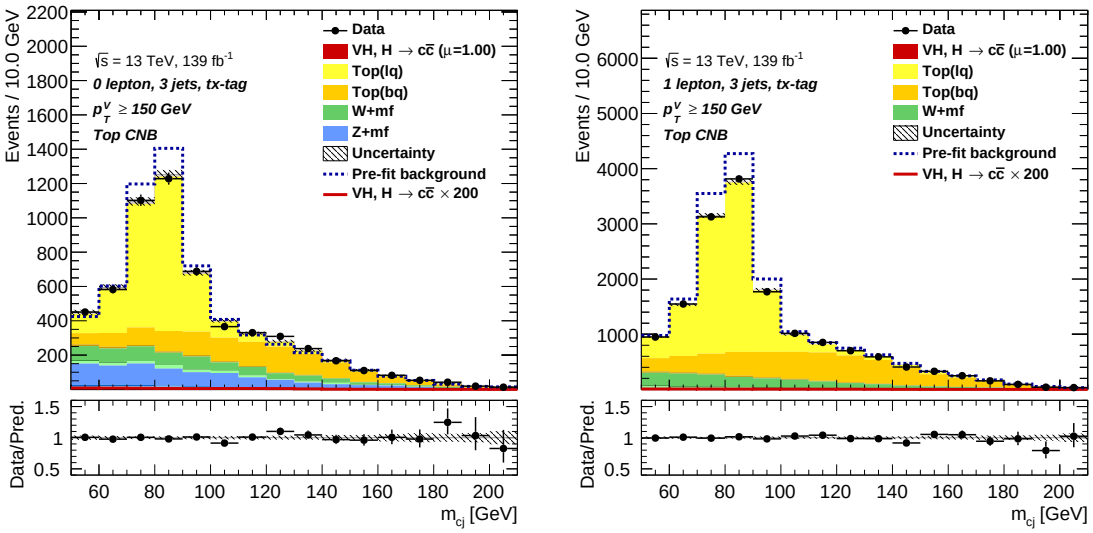


Figure 5.13: Top TNB control region (abusively called CNB CR) in the (a) 0- and (b) 1-lepton channels for $p_{\mathrm{T}}^{V} > 150$ GeV. The mixed flavour (mf) components refer to *bc*, *bl* and *cl* true flavour of the two jets constituting the Higgs candidate system. The ratios of the data to the sum of the fitted signal and background are shown in the lower panels.

5.4 Jet mass scale calibration for small-R jets

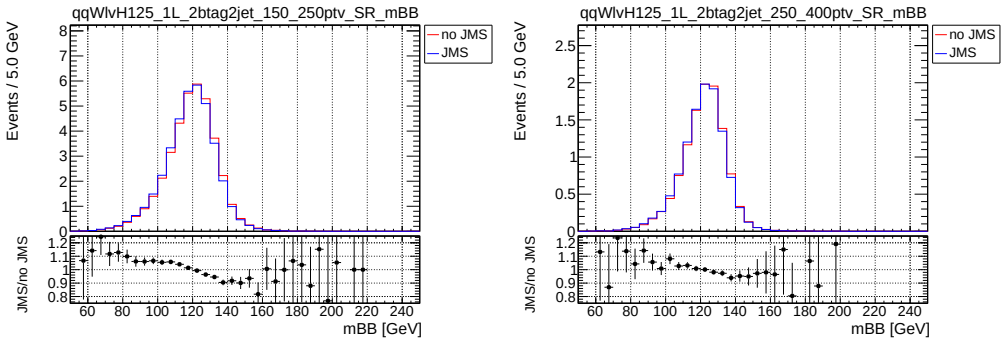
Recently, ATLAS developed a jet mass scale calibration for small-R jets similarly to the JMS calibration applied for large-R jets as described in Section 3.6.3.2. This calibration is advantageous as it corrects for discrepancies between the particle-level mass and the reconstructed jet mass hence reducing biases due to jet reconstruction algorithms and detector effects.

When applying the JMS calibration to small-R jets, a non-negligible shift of the m_{bb} invariant mass distribution toward lower values was observed for signal events in all leptonic channels in the $VH, H \to b\bar{b}$ resolved topology. The impact of the JMS calibration is shown for the 1-lepton channel in the Figure 5.14. Such differences have been further investigated as the variations induced by the JMS calibration were much larger than expected. A comparison of the leading and sub-leading b-jet characteristics (mass, energy and transverse momentum) at GSC calibration level has been performed in order to avoid the additional OneMu and PtReco corrections. The associated distributions are presented in Figure 5.15. Only the mass of jets is affected by the JMS calibration as expected, and applying the JMS reduces the invariant mass of the jets hence the lower m_{bb} values obtained in that case. The impact on the reconstructed mass m_{bb} of the JMS is though smaller than the discrepancy observed for the individual jet mass m_{b_1} and m_{b_2} because

$$m_{bb} = m_{b_1}^2 + m_{b_2}^2 + 2(E_{b_1}E_{b_2} - \boldsymbol{p_{b_1}} \cdot \boldsymbol{p_{b_2}}), \tag{5.4.1}$$

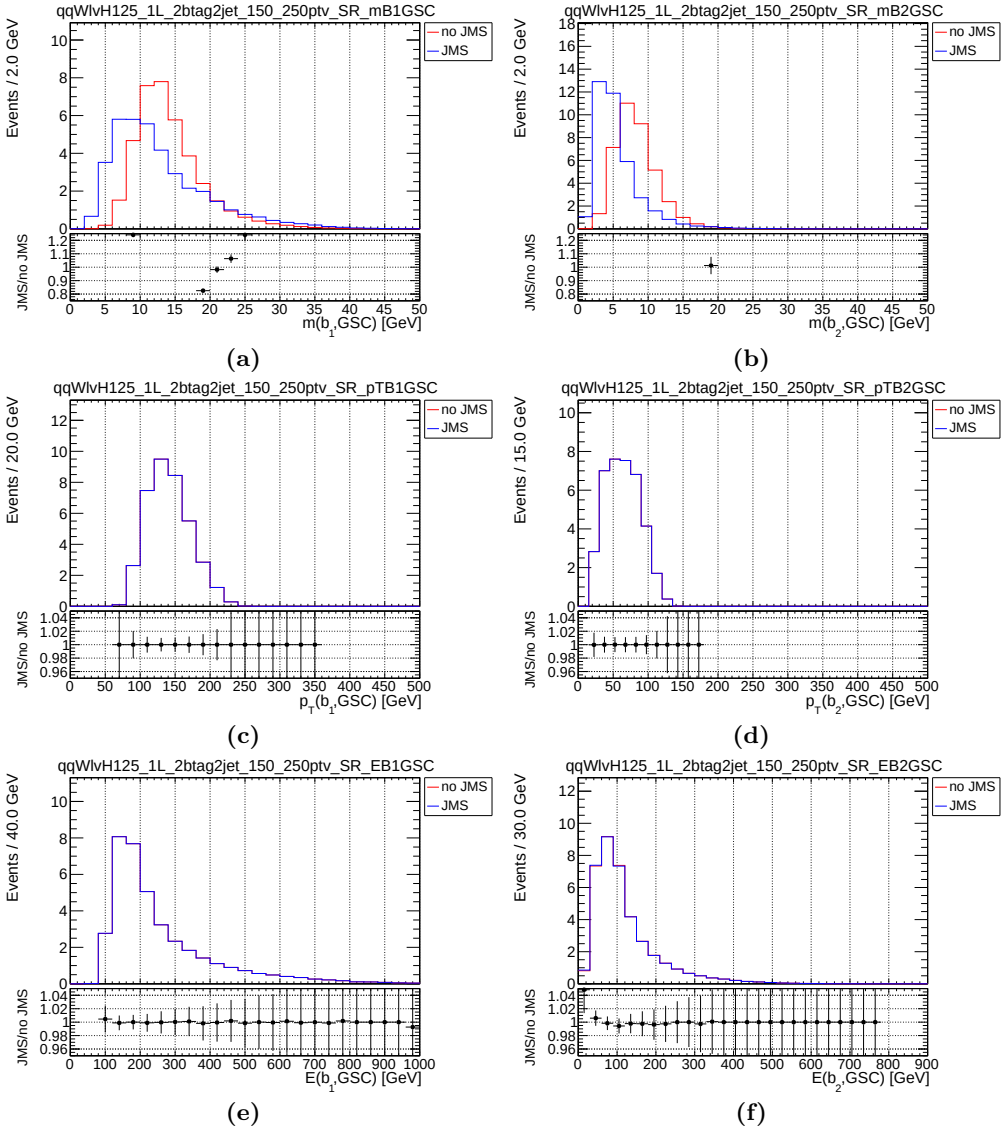
and $\left|p_{b_{1,2}}\right| \approx 100\text{-}150\ \text{GeV} > m_{b_{1,2}} \approx 10\text{-}15\ \text{GeV}$.

To determine if the JMS calibration should be used for small-R jets in the $VH, H \to b\bar{b}/c\bar{c}$ analyses, a comparison between truth and reconstructed small-R jet invariant mass has been studied in the 1-lepton channel. The truth jet invariant mass distributions of the leading and sub-leading b-jets are compared in Figure 5.16 to the reconstructed distributions with and without JMS applied for the different jet calibration level: GSC level, after muon in jet (OneMu) correction and after the full correction (OneMu+PtReco). As a result, a better agreement between truth and reconstructed invariant mass is obtained without applying the JMS with the truth WZ jets being the jets reconstructed from truth level particles with the anti-k_t $R = 0.4$ algorithm taking into account all final state particles including the products of the semi-leptonic decays in hadrons *i.e.* leptons and



(a) (b)

Figure 5.14: Comparison of the invariant mass (m_{bb}) distribution obtained with (blue histogram) and without (red histogram) the JMS calibration applied as predicted by the nominal simulation for the 2015+2016 recorded data for $WH, H \to b\bar{b}$ signal events for the 2 b-tagged 2-jets, (a) $150 < p_{\mathrm{T}}^{V} < 250$ GeV and (b) $250 < p_{\mathrm{T}}^{V} < 400$ GeV regions in the 1-lepton channel.



(a) (b) (c) (d) (e) (f)

Figure 5.15: Comparison of the mass, the transverse momentum and the energy of the (a), (c) and (e) leading (b_1), and (b), (d) and (f) sub-leading (b_2) calibrated at GSC level b-jets, with (blue histogram) and without (red histogram) JMS calibration applied as predicted by the nominal simulation for the 2015+2016 recorded data for $WH, H \to b\bar{b}$ signal events for the 2 b-tagged 2-jets, $150 < p_{\mathrm{T}}^{V} < 250$ GeV region in the 1-lepton channel. GSC level jet (a) and (b) invariant mass, (c) and (d) transverse momentum, and (e) and (f) energy.

neutrinos.

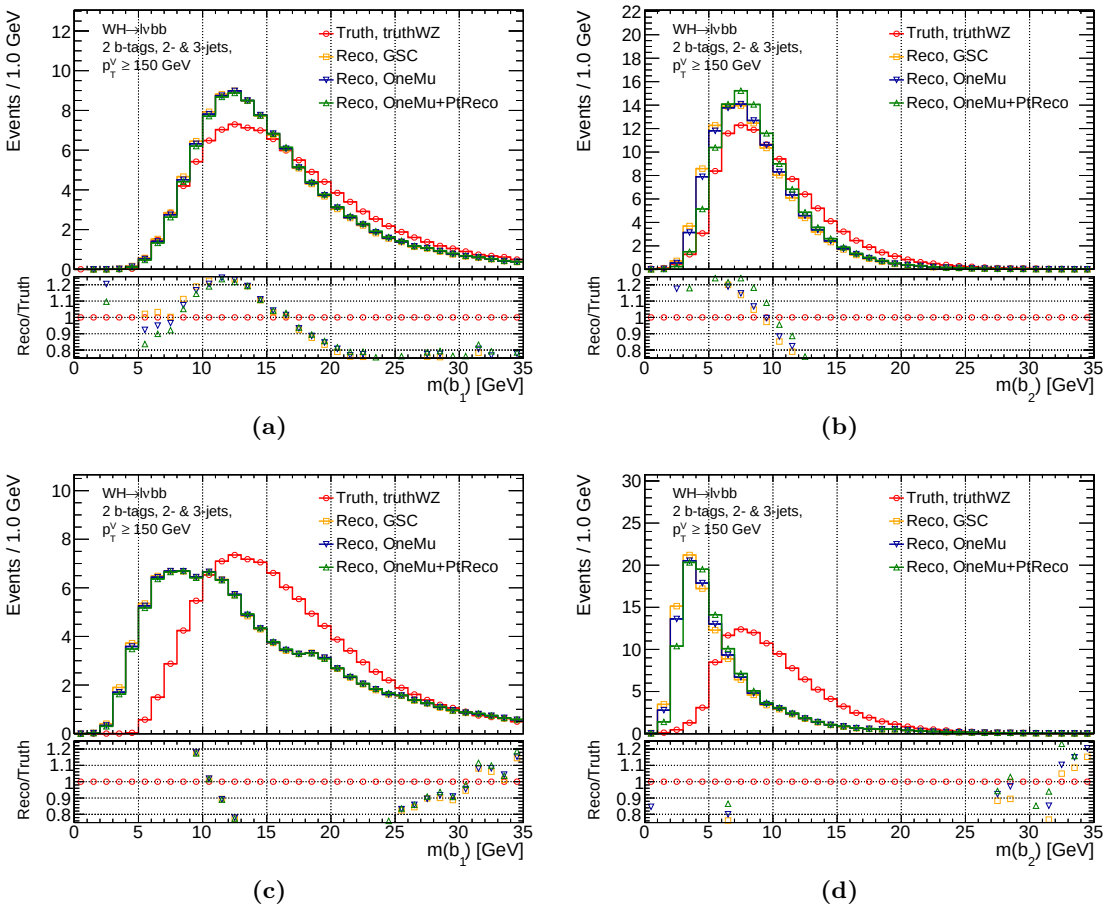


Figure 5.16: Comparison between the truth level (truth WZ) jet invariant mass and the reconstructed distributions corresponding to different level of jet calibration: GSC level, after muon-in-jet corrections (OneMu) and after applying the full correction (OneMu+PtReco). Invariant mass distributions (m_{b_1}) of the leading b-tagged jet (a) without and (c) with JMS applied, and invariant mass distributions (m_{b_2}) of the sub-leading b-tagged jet (b) without and (d) with JMS applied. The bottom panel shows the ratio between the reconstructed and truth invariant mass distributions. The distributions are derived for $WH, H \to b\bar{b}$ signal events in the 1-lepton channel for $p_{\mathrm{T}}^{V} > 150$ GeV region combining the 2 b-tagged 2- and 3-jets categories, and are predicted by the nominal simulation for the 2015+2016 data taking period.

Finally, a comparison of the data versus Monte-Carlo agreement has been performed in the three channels in order to test if a better data to simulation agreement could be achieved with the JMS calibration. The comparison is presented in Figure 5.17, the distribution are obtained without CR normalization scale factors applied. No improvement was obtained with the JMS both for m_{bb} or individual jet invariant masses m_{b_1} and m_{b_2}. This new JMS calibration has been developed and validated mainly in events using more massive jets ($m > 40\,\mathrm{GeV}$) than the targeted region of phase space of our analysis as $m_b \approx$ 10-15 GeV. Moreover, unlike large-R JMS, the small-R JMS does not include an in-situ JMS calibration hence no improvement of the data to simulation agreement is indeed expected. The impact of such calibration has been investigated, and it was decided to not apply the JMS calibration as it was found to be not appropriate for the $VH, H \to b\bar{b}/c\bar{c}$ analyses.

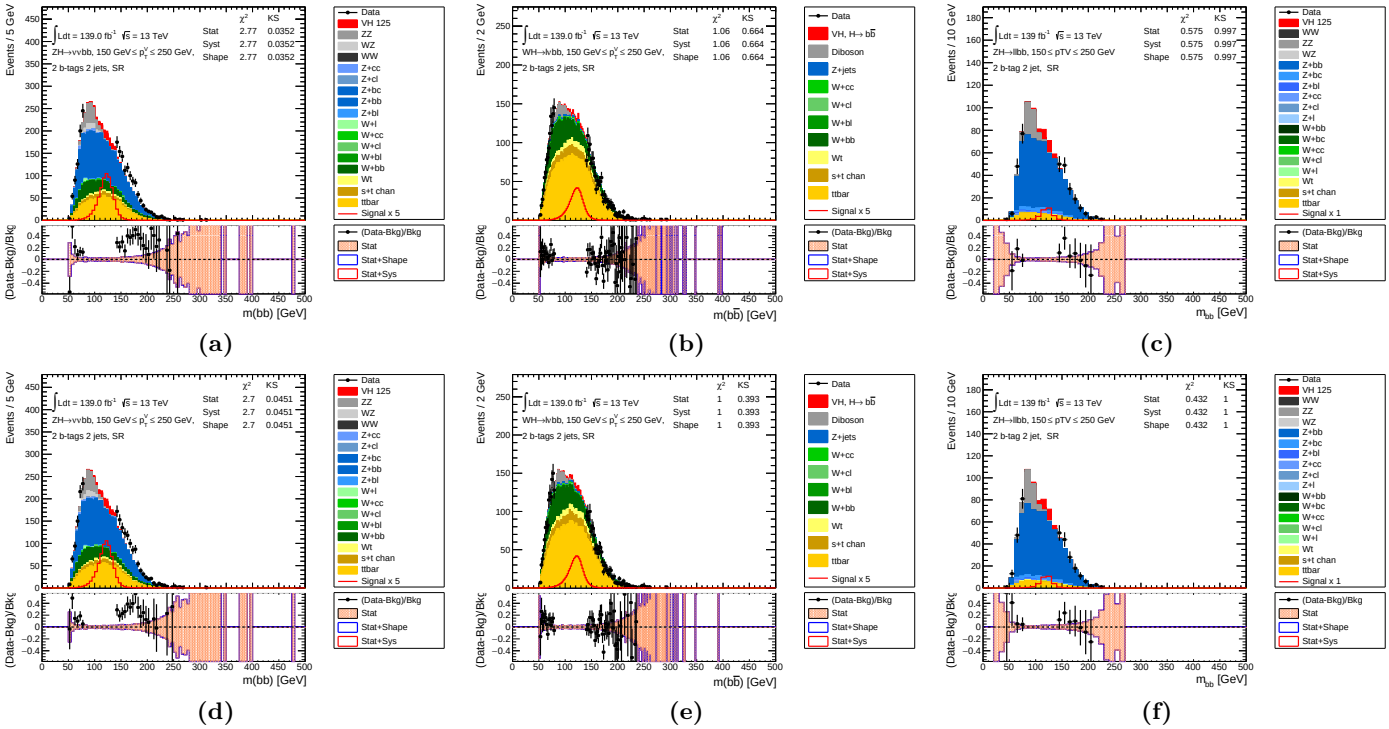


Figure 5.17: Comparison of the data-to-simulation agreement for the m_{bb} invariant mass distributions in the (a) and (d) 0-lepton, (b) and (e) 1-lepton, and (c) and (f) 2-lepton channels obtained (a), (b) and (c) without JMS and (d), (e) and (f) with JMS applied as predicted by the nominal simulation in the 2 b-tagged 2-jets signal regions. The normalization scale factors from control regions are not applied for those distributions.

5.5 Flavour tagging methods

For the $VH, H \to b\bar{b}$ analyses, the DL1r b-tagging algorithms for the resolved and boosted topologies have a high rejection of c- and light-jets resulting, after b-tagging requirements, in a lack of simulated background events whose final state is involving one or several c- and/or light-jets. The same reasoning applies for the $VH, H \to c\bar{c}$ analysis in terms of any background whose final states is not only involving c-jets. Such tagging method is referred to as *direct tagging* as only events satisfying flavour tagging requirements are taken into account.

For the $VH, H \to b\bar{b}$ analyses, to circumvent the lack of statistics for c- and light-component background events in simulations and avoid statistical fluctuations in the distributions of interest, several methods have been developed to accept simulated events which would have failed the b-tagging requirements otherwise. The same c-tagging techniques as listed below are used for the $VH, H \to c\bar{c}$ analysis.

➢ **The truth tagging method**: the idea behind this approach is to assign to each jet in simulated events a probability of passing the b-tagging requirements. All events are considered as passing b-tagging criteria and events are then simply reweighted based on their probability of satisfying such requirement. The reweighting is performed using the true flavour of each jet and ensures that the sum of all events weights remains consistent, and preserves the normalizations and shapes of the distributions of interest when comparing the ones obtained with direct tagging to the ones with truth tagging.

The truth tag weight of an event is computed as the product of the b-tagging efficiencies (ε_b) of the b-tagged jets, times the b-tagging inefficiencies ($1-\varepsilon_b$) of the non b-tagged jets. The tagging efficiencies are determined for each true flavour (b, c and light) component as a function of p_{T}

and $|\eta|$ thanks to a calibration performed in $t\bar{t}$ simulated events as shown in Figure 5.18. In case the number of jets in the event is greater than the number of required b-tagged jets, then all possible combinations of b-tagged and non b-tagged jets are considered, and the truth tag weight is the sum of those combinations. For example, for an event with three jets (with respective b-tagging efficiencies denoted ε_i) where 2 b-tagged jets are required then the truth tag weight (w) of the event is calculated as:

$$w = \varepsilon_1\varepsilon_2(1-\varepsilon_3) + \varepsilon_1(1-\varepsilon_2)\varepsilon_3 + (1-\varepsilon_1)\varepsilon_2\varepsilon_3. \tag{5.5.1}$$

Moreover, in that case the two jets considered as "b-tagged" among the three jets of the event are obtained by randomly selecting a combination of those jets among all possible pairs. The probability of a given pair to be chosen is based on the truth tag weight of the combination *e.g.* the probability ($p_{1,2}^{\text{tagged}}$) of the two leading jets to be chosen as the b-tagged pair is equal to:

$$p_{1,2}^{\text{tagged}} = \frac{\varepsilon_1\varepsilon_2(1-\varepsilon_3)}{w}. \tag{5.5.2}$$

More details about truth tagging and the generalization to higher jet multiplicities can be found in Reference [268].

The truth tagging allows for a large increase in the statistics of simulated events available in particular for c- and light-component background events. However, this approach relies on a precise knowledge of the b-tagging efficiencies for b-, c- and light-jets as a function of p_{T} and $|\eta|$ as otherwise an incorrect reweighting of events would be applied and would result in introducing mismodelling *i.e.* discrepancies between direct tagging and truth tagging distributions. Furthermore, the efficiency of a jet is only determined based on its p_{T} and $|\eta|$ which might not be a perfect assumption in some cases as it could depend on other criteria such as the jet being close-by to another jet and so on. For instance, for the previous $VH, H \to b\bar{b}$ boosted publications, such method was not used because the agreement between distributions obtained with the direct and the truth tagging methods was not enough satisfying.

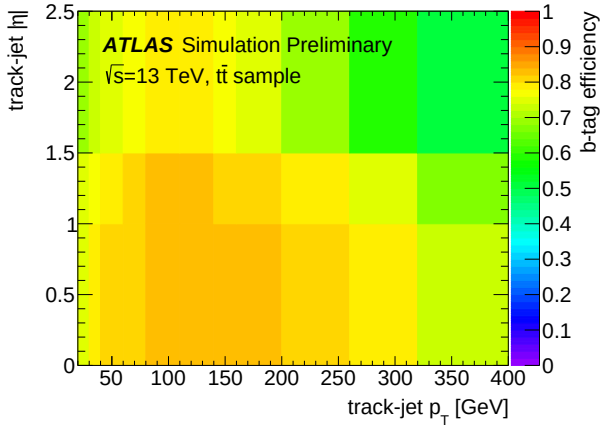


Figure 5.18: Efficiency map determined for true b-flavoured track-jets with simulated $t\bar{t}$ events using the 77% WP of the DL1r tagger (taken from Ref. [318]).

- **The hybrid tagging method**: is a mix between the direct tagging and the truth tagging approaches. For the true b-jets, direct tagging is used while the other jets (true c- and light-jets) are truth tagged. This technique avoid having to deal with true b-jets efficiency estimations hence should reduce mismodelling while still benefiting from the increase of statistics for the other components.
- **The GNN truth tagging method** [318]: this recently developed technique aims to tackle the drawbacks of the truth tagging. Indeed, for truth tagging technique, the weight are assigned to each jet solely based on the 2-dimensional information (p_{T},$|\eta|$) of the jet and without taking into

possible correlations between the jets of the event. The GNN truth tagging is taking into account more characteristics of the events that are listed in Table 5.5 for track-jets and predicts the flavour-tagging efficiency for each considered track-jet. Moreover, this prediction is performed simultaneously for all jets in the event hence correlations between neighbour jets are considered by the GNN whereas the truth tagging method assigns probabilities individually to each jet: b-tagging efficiencies could be affected by close to each other jets which is the case in particular for the track-jets used in the $VH, H \rightarrow b\bar{b}$ boosted analysis. A similar GNN truth tagging algorithm is trained for small-R jets for the resolved topology.

Track-jet variables
Track-jet p_{T}
Track-jet η
Track-jet ϕ
Track-jet flavour label
Mass of the p_{T}-leading b- or c-hadron in the track-jet
p_{T} of the p_{T}-leading b- or c-hadron in the track-jet
η of the p_{T}-leading b- or c-hadron in the track-jet
ϕ of the p_{T}-leading b- or c-hadron in the track-jet
Event variables
Average number of interactions per event, $\langle\mu\rangle$
Jet-pair variables
Angular separation between two track-jets, ΔR

Table 5.5: List of input variables used by the GNN truth tagging for track-jets (taken from Ref. [318]).

As shown in Figure 5.19, a better agreement between the GNN truth tagging and direct tagging distributions is achieved compared to truth tagging.

The GNN truth tagging is hence foreseen to be used both for the $VH, H \rightarrow b\bar{b}/c\bar{c}$ analyses as it allies the benefits from the truth tagging method while solving mismodelling issues.

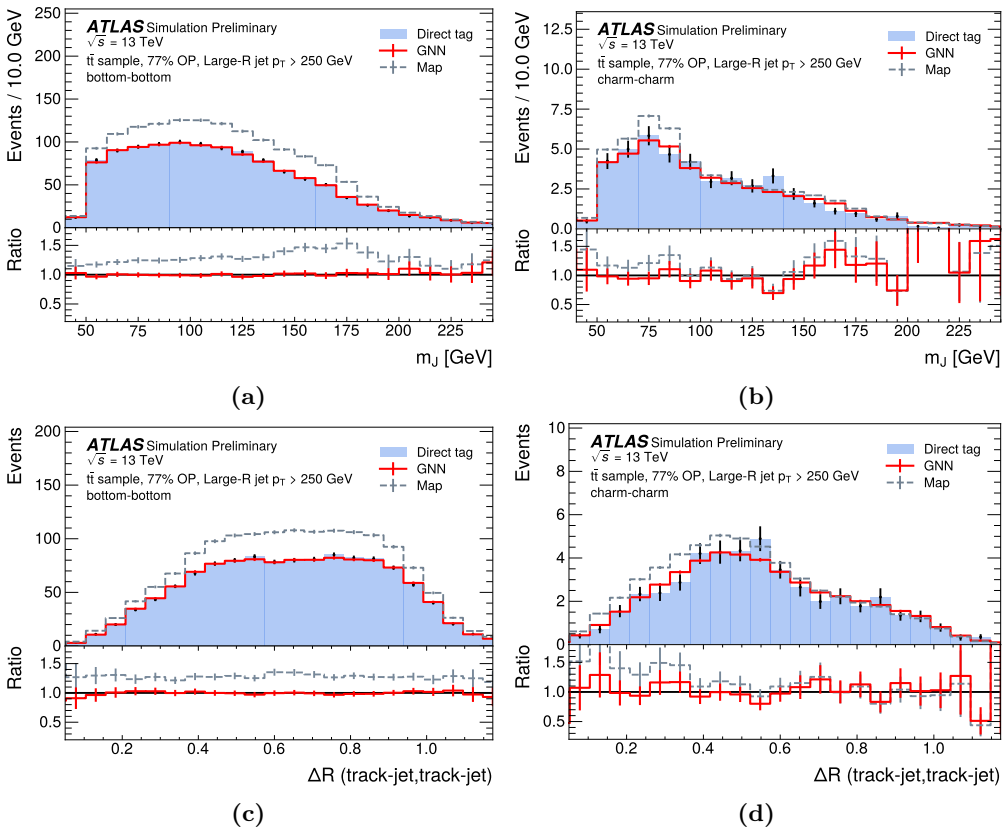


Figure 5.19: Distributions comparison of (a) and (b) the large-R jet mass, and (c) and (d) the angular distance between its two leading ghost-associated track-jets. The distributions are obtained with direct tagging (blue histogram), truth tagging (2-dimensional map, grey), and GNN truth tagging (red), in $t\bar{t}$ events selecting large-R jets with $p_{\mathrm{T}} > 250\,\mathrm{GeV}$ requiring the two leading track-jets to pass the b-tagged 77% WP requirement of the DL1r tagger. The Figures (a) and (c) correspond to a selection of a pair of true b-flavoured track-jets while (b) and (d) a pair of true c-flavoured track-jets. The black error bands represent the statistical error in direct tagging. The bottom panel shows the ratio with respect to the direct tagging distribution (taken from Ref. [318]).

5.6 Multivariate discriminants in the $VH, H \to b\bar{b}/c\bar{c}$ analyses

The $VH, H \to b\bar{b}/c\bar{c}$ analyses are using multivariate analysis (abbreviated MVA) algorithms to discriminate signal from background in order to increase the overall sensitivity to signal. The novelty for the Legacy Run 2 publication is that the MVA was previously only used for the $VH, H \to b\bar{b}$ resolved analysis while the $VH, H \to b\bar{b}$ boosted and $VH, H \to c\bar{c}$ were "cut-based" analyses *i.e.* only based on orthogonal selections on simple object-level or event-level quantities: the signal yield in that case was extracted from the invariant mass m_J or m_{cc} of the Higgs candidate system which is the most discriminating variable. The MVA approach has now been extended to the $VH, H \to b\bar{b}$ boosted and the $VH, H \to c\bar{c}$ analyses. Specific machine learning (ML) algorithms have been trained and optimized for each analysis and lepton channel for a total of 9 MVAs. Typically, it has been observed a 30% to 200% improvement in sensitivity per SR in each channel when using MVA algorithms compared to the mass-based analysis both for $VH, H \to b\bar{b}$ boosted and $VH, H \to c\bar{c}$. This increase can be explained by the fact that an MVA algorithm is provided several relevant input variables, it exploits correlations between them, and it is learning characteristics specific to different regions of phase space hence being more efficient than simple cut selections. Basically a machine learning algorithm decision is based on multidimensional information while individual cut selections are usually not exceeding 2 or 3 dimensions unless analytic formula exist which is pretty rare.

Machine learning for signal/background discrimination is divided in two steps:

- first the training of the ML algorithm: a *training sample* containing simulated events is provided to the MVA algorithm which is informed for each event if they belong to the signal or the background category. In this phase the ML is learning how to recognize them.
- Then there is the evaluation of the MVA performances: an *evaluation sample* (also called *test sample*), orthogonal to the *training sample*, is used to assess the performances of the algorithm for instance computing the obtained significance or other functions to be optimized (*e.g.* loss-functions). Another commonly looked at feature when assessing performances of an MVA algorithm is its *receiver operating characteristic* (ROC) curve in particular the *area under the curve* (AUC) of the ROC curve. The background inefficiency ($1 - \varepsilon_b$) is plotted as a function of the signal efficiency (ε_s), the curve obtained is referred to as the ROC curve, the area enclosed between that curve and the x-y axis is the AUC: it can vary between 0 and 1. The larger the AUC is, the better the algorithm is at discriminating signal from background as for large signal efficiencies the background efficiency (respectively background rejection) remains small (large).

The key points for machine learning algorithms to achieve better performances than cut based selections is:

- first to provide them a set of relevant variables that the algorithm can learn from in order to discriminate events. It is important to mention that some input variables provided may not appear at first glance useful when looking at their inclusive distributions. However, one has to keep in mind that the ML decisions are optimized in different regions of phase space and in multidimensional spaces: those variables may become relevant and could be discriminating variables in some specific regions.
- The second aspect is the optimization of the so-called *hyperparameters* of the algorithm. Those parameters which needs to be set before the training are influencing how it learns from the training sample, how efficient and robust its discrimination will be. The risk with ML is the phenomenon of *overtraining* which corresponds basically to the ML algorithm learning to discriminate signal from background based on the peculiarities of the training sample instead of the general features and characteristics of events. As a result performances of the algorithm

in that case in the evaluation phase worsen because the algorithm was not able to learn the general features and, in principle, the evaluation sample does not have the same peculiarities as the training sample. The overtraining is usually assessed by comparing the ROC curves and AUC obtained from the evaluation in the training and test samples. If large differences are observed for the ROC curves and their AUC it implies overtraining which should be avoided. On the other hand, a ML algorithm can *underfit* the training sample *i.e.* its learning and decisions are based on too conservative information: it is not managing to grasp the characteristics that distinguish signal from background resulting in poor performances in terms of discrimination reflected by obtaining a low AUC and a not optimal ROC curve. As a consequence, a careful tuning of the hyperparameters is crucial for optimizing the discriminating power of a machine learning algorithm.

The optimization of the hyperparameters and input variables set should be performed at the same time, because the optimal hyperparameters for a given set of variables may differ if the set is changed and vice-versa.

The MVA algorithm used by the $VH, H \to b\bar{b}/c\bar{c}$ analyses are boosted decisions trees (BDT) [205] (BDTs are introduced in Section 3.6.5 and schematized in Figure 3.27). A decision tree is a classifier performing repeated left/right (yes/no) decisions dividing phase space based on input variables until a stop criterion is fulfilled. Events are hence classified either as signal or background: a probability of being a signal event is assigned to them by the tree. Boosted decision trees is simply the extension from one to several decision trees which form a *forest*. The trees are derived from the same training ensemble by reweighting events (especially misclassified ones), and are finally combined into a single classifier which is given by a weighted average of the individual decision trees. Boosting enables to reduce overfitting peculiarities and prevents sensitivity to fluctuations in the training sample hence it enhances performances with respect to a single tree and avoids overtraining.

In terms of BDTs training in the 0- and 1-lepton channels:

- for the resolved topology, 2 BDTs are trained per lepton channel corresponding to the 2- and 3-jets categories for $p_{\mathrm{T}}^{V} > 150$ GeV both for $VH, H \to b\bar{b}$ and $VH, H \to c\bar{c}$.
- For the boosted topology, 1 BDT is trained per lepton channel corresponding to $p_{\mathrm{T}}^{V} > 400\,\mathrm{GeV}$.

For the 2-lepton channel:

- for the resolved topology, 4 BDTs are trained per lepton channel corresponding to the 2- and $\geq$ 3-jets categories for $p_{\mathrm{T}}^{V} \in [75, 150]$ GeV and $p_{\mathrm{T}}^{V} > 150$ GeV both for $VH, H \to b\bar{b}$ and $VH, H \to c\bar{c}$.
- For the boosted topology, 1 BDT is trained corresponding to $p_{\mathrm{T}}^{V} > 400$ GeV.

For the $VH, H \to b\bar{b}/c\bar{c}$ analyses, 2 different boosting algorithms described below are used:

- **the adaptive boost (AdaBoost) algorithm** [205]: the first tree (also called first classifier) starts with the original event weights and classify events into several regions containing either a majority of signal or background events. For the n^{th} classifier (with $n > 1$), the misclassified events from the $n-1^{\text{th}}$ tree are assigned higher weights by multiplying them by a common boost weight α_n obtained as:

$$\alpha_n = \frac{1 - \mathrm{err}_n}{\mathrm{err}_n}, \tag{5.6.1}$$

with err_n being the misclassification rate computed as

$$\mathrm{err}_n = \frac{\sum_i w_{n-1,i} M_{n-1}(i)}{\sum_i w_{n,i}}, \tag{5.6.2}$$

$w_{n-1,i}$ being the weight assigned to the i^{th} event in the training sample for the $n-1^{\text{th}}$ tree. And $M_{n-1}(i) = 1$ or 0 whether if the event was misclassified or not by the $n-1^{\text{th}}$ tree. As a result the event weights at the n^{th} step can always be expressed as:

$$w_{n,i} = \alpha_n^{M_{n-1}(i)} \cdot w_{n-1,i}. \tag{5.6.3}$$

Assuming that a tree selection is better than random guessing, then the misclassification fraction satisfies $\text{err}_n \leq 0.5$ implying that $\alpha_n \geq 1$. Once all weights have been recomputed for the n^{th} tree, they are all renormalized such that the sum of weights remains constant.

A generalization of the reweighting of events can be obtained by using a parameter β such that:

$$\alpha_n = \left(\frac{1-\text{err}_n}{\text{err}_n}\right)^{\beta}. \tag{5.6.4}$$

This parameter β is called the *learning rate* because it controls how large is the reweighting of misclassified events. High values of β will result in higher overtraining as a higher importance is given to misclassified events, in that case the algorithm could focus on a subset of hard to discriminate events or could learn specific features of the training sample. On the other hand low values of β could decrease the performances of the AdaBoost algorithm by not giving enough importance to the misclassified events.

The final output of the BDT classifier is a score between -1 and 1 obtained from each individual classifier:

$$y(\boldsymbol{x}) = \frac{1}{N}\sum_{n=1}^{N} \ln(\alpha_n) h_n(\boldsymbol{x}) \tag{5.6.5}$$

with $\boldsymbol{x}$ the set of input variables, $h_n(\boldsymbol{x}) = 1$ or -1 whether the event is classified as signal or background by the n^{th} classifier and N being the total number of classifiers (also referred to as the *number of trees*).

Another way to understand the AdaBoost algorithm, is to define the function F which corresponds to the classifier output score *i.e.* classifier prediction for an event:

$$F(\boldsymbol{x}) = \frac{1}{N}\sum_{n=1}^{N} \ln(\alpha_n) h_n(\boldsymbol{x}). \tag{5.6.6}$$

Then this algorithm is minimizing for every event the exponential loss function:

$$L(F, y, \boldsymbol{x}) = \exp\left(-yF(\boldsymbol{x})\right), \tag{5.6.7}$$

with y the expected value which is equal to 1 for a signal event and -1 for a background event. The total loss function to minimize is simply:

$$E(F; \{\boldsymbol{x_i}, y_i\}) = \sum_{i\in\{\text{events}\}} L(F, y_i, \boldsymbol{x_i}) = \sum_{i\in\{\text{events}\}} \exp\left(-y_i F(\boldsymbol{x_i})\right). \tag{5.6.8}$$

➢ **The gradient boost algorithm** [205]: the AdaBoost algorithm is lacking robustness with respect to outliers points in a training data set which can decrease performances. The gradient boost algorithm tries to circumvent this drawback by allowing for more robust loss functions while still keeping the good performances of AdaBoost: in principle it should be less susceptible to overtraining. The function used in that case is the binomial log-likelihood loss function:

$$L(F, y, \boldsymbol{x}) = \ln\left[1 + \exp\left(-2yF(\boldsymbol{x})\right)\right]. \tag{5.6.9}$$

The classifier algorithm and its output function F associated to that loss function cannot be obtained in a straightforward manner as for AdaBoost. The minimization is performed with a *steepest method* by computing iteratively the gradient of the loss function which provides information about how the algorithm should be modified to reduce the current loss function value.

The resolved $VH, H \to b\bar{b}$ analysis uses the gradient boost algorithm. Despite that the gradient boost algorithm is expected to achieve better performances than AdaBoost, the $VH, H \to b\bar{b}$ boosted and $VH, H \to c\bar{c}$ analyses are using the adaptive algorithm because a larger overtraining was observed for the gradient boost algorithm. This could perhaps be explained by the lower amount of signal events and larger background contamination for those two analyses compared to the $VH, H \to b\bar{b}$ resolved one, which could result in instabilities in the gradient descent procedure: hence the gradient boost method in that case would be more sensitive to fluctuations. The BDTs hyperparameters have been optimized in order to improve ROC curves and the statistical significance defined in Equation (5.7.27) while at the same time avoiding overtraining. For the $VH, H \to c\bar{c}$ analysis, some further studies are needed to optimize the set of input variables. The Table 5.6 summarizes all the input variables used by the BDTs. Some comparison of the signal versus background MVA input variables distributions can be found in Section 6.1.1 for the resolved topology MVA training of the 0-lepton channel. The variables not already introduced or that require some clarifications are described below.

➢ **Resolved topology specific variables ($\boldsymbol{VH, H \to b\bar{b}/c\bar{c}}$)**:

- The leading and sub-leading small-R jets chosen as the Higgs candidate system by means of b- or c-tagging requirements are respectively denoted j_1 and j_2.
- $m_{j_1 j_2}$ is the invariant mass of the Higgs candidate system.
- $\Delta R(j_1, j_2)$, $\Delta\phi(j_1, j_2)$ and $\Delta\eta(j_1, j_2)$ are respectively the angular, the azimuthal and the pseudorapidity distances between the two jets forming the Higgs candidate system.
- In the 3- and 4-jets categories, the third leading (untagged) jet of the event is denoted j_3. The invariant mass $m_{j_1 j_2 j_3}$ of the triplet is used as well as the transverse momentum of the third jet ($p_{\mathrm{T}}^{\mathrm{jet}_3}$). $m_{j_1 j_2 j_3}$ is for instance useful to reject events involving top quarks as if that reconstructed jet triplet comes from a top hadronic decay, its invariant mass should be close to the top quark mass $m_{\mathrm{top}} = 172.5\ \mathrm{GeV}$.
- The $\mathrm{bin}_{\mathrm{DL1r}}$ corresponds to the tagged-bin a jet belongs to (5 possible bins): the untagged, the loose (70% WP) and the tight (60% WP) b-tagged, and the loose and the tight c-tagged bins.

➢ **Boosted topology specific variables ($\boldsymbol{VH, H \to b\bar{b}}$ only)**:

- $j_{1,\mathrm{trk}}$ and $j_{2,\mathrm{trk}}$ are the leading and sub-leading b-tagged track-jets ghost associated to the leading large-R jet.
- $\Delta R(j_{1,\mathrm{trk}}, j_{2,\mathrm{trk}})$ is the angular distance between the two b-tagged track-jets.
- If the leading large-R jet contains at least 3 ghost-associated track-jets, the untagged track-jet among the three leading ones is denoted $j_{3,\mathrm{trk}}$ and $p_{\mathrm{T}}^{\mathrm{jet}_3,\mathrm{trk}}$ correspond to its transverse momentum.
- $N(\text{track-jets in } J)$ is the number of track-jets that are ghost-associated to the leading large-R jet.
- $N(\text{add. small-R jets})$ is the number of additional small-R jets that are not matched to the leading large-R jet.
- The $\mathrm{bin}_{\mathrm{DL1r}}$ corresponds to the tagged-bin the track-jet belongs to (4 possible bins): the 85%, the 77%, the 70% and 60% b-tagging efficiency bins.

➢ **0-lepton channel specific variables**:

- the m_{eff} is the scalar sum of $E_{\mathrm{T}}^{\mathrm{miss}}$ and the p_{T} of the small-R jets in the event.
- $p_{\mathrm{T}}^{\mathrm{miss,st}}$ is the track based soft $E_{\mathrm{T}}^{\mathrm{miss}}$ term, it corresponds to the vectorial sum of the p_{T} of all tracks in the event that are not associated to any reconstructed object in the event.
- $\Delta\phi(\boldsymbol{V}, \boldsymbol{H}_{\mathbf{cand}})$ is the azimuthal distance between the reconstructed $\boldsymbol{E}_{\mathbf{T}}^{\mathbf{miss}}$ and the Higgs boson candidate.

➢ **1-lepton channel specific variables**:

- m_{T}^{W} is the transverse mass of the W boson candidate defined as

$m_{\mathrm{T}}^{W} = \sqrt{2p_{\mathrm{T}}^{l} E_{\mathrm{T}}^{\mathrm{miss}}(1 - \cos(\Delta\phi(\boldsymbol{l}, \boldsymbol{E}_{\mathbf{T}}^{\mathbf{miss}})))}$.

- $\min[\Delta\phi(\boldsymbol{l}, \boldsymbol{b})]$ (or $\min[\Delta\phi(\boldsymbol{l}, \boldsymbol{c})]$ for the $VH, H \to c\bar{c}$ analysis) is the distance in ϕ between the lepton and the closest b-tagged (c-tagged) jet.
- m_{top} is the reconstructed mass of the leptonically decaying top quark. The reconstruction of m_{top} requires to determine the longitudinal momentum of the neutrino (p_z^{ν}) using the mass of the W boson:

$$p_z^{\nu} = \frac{1}{2(p_{\mathrm{T}}^{l})^2}\left[p_z^{l} X \pm E_l \sqrt{X^2 - 4(p_{\mathrm{T}}^{l})^2 (E_{\mathrm{T}}^{\mathrm{miss}})^2}\right], \tag{5.6.10}$$

  with X computed as

$$X = m_W^2 + 2(p_x^{l} E_x^{\mathrm{miss}} + p_y^{l} E_y^{\mathrm{miss}}). \tag{5.6.11}$$

  Then m_{top} is obtained by selecting the jet from the Higgs candidate 2-jets system and solution for p_z^{ν} that minimize m_{top}. In case $X^2 - 4(p_{\mathrm{T}}^{l})^2 (E_{\mathrm{T}}^{\mathrm{miss}})^2 < 0$ then $E_{\mathrm{T}}^{\mathrm{miss}}$ is corrected such that $X^2 - 4(p_{\mathrm{T}}^{l})^2 (E_{\mathrm{T}}^{\mathrm{miss}})^2 = 0$
- $\Delta y(\boldsymbol{V}, \boldsymbol{H}_{\mathbf{cand}})$ is the rapidity difference between the W boson and the Higgs boson candidates. It also requires to determine the longitudinal momentum of the neutrino (p_z^{ν}).
- $\Delta\phi(\boldsymbol{V}, \boldsymbol{H}_{\mathbf{cand}})$ is the azimuthal distance between the reconstructed W boson and the Higgs boson candidates.

➢ **2-lepton channel specific variables**:

- $(p_{\mathrm{T}}^{l_1} - p_{\mathrm{T}}^{l_2})/p_{\mathrm{T}}^{Z}$ is the lepton p_{T}-asymmetry with $p_{\mathrm{T}}^{l_1}$ and $p_{\mathrm{T}}^{l_2}$ being the transverse momenta of the leading and sub-leading leptons, and p_{T}^{Z} the transverse momentum of the Z boson candidate formed by the two leptons.
- $\cos\theta(\boldsymbol{\ell}^{-}, \boldsymbol{Z})$ is the Z boson polarization. It is the cosine of the angular distance between the direction of the negatively charged lepton in the Z boson rest frame and the flight direction of the Z boson in the laboratory frame. The polarization for the ZH process is expected to be different from the dominant $Z + \text{jets}$ background [319]. This variable was found to increase the sensitivity by 7% for the resolved topology [65].
- $E_{\mathrm{T}}^{\mathrm{miss}}/\sqrt{S_{\mathrm{T}}}$ is the quasi-significance of the $E_{\mathrm{T}}^{\mathrm{miss}}$ with S_{T} the scalar sum of the p_{T} of the leptons and jets in the event.
- $\Delta\eta(\boldsymbol{V}, \boldsymbol{H}_{\mathbf{cand}})$ is the pseudorapidity distance between the reconstructed Z-boson and Higgs boson candidates.

Some variables that were previously used in cut selections, as for instance the p_{T}-imbalance for the 2-lepton channel, have now been added to the list of input variables of the MVA algorithms and related cuts have been removed. Due to the switch to the MVA approach for the $VH, H \to b\bar{b}$ boosted and $VH, H \to c\bar{c}$ analyses, it is preferable to use them for the MVA which takes into account correlations between variables rather than using simple cuts.

		$VH, H \to b\bar{b}$			$VH, H \to c\bar{c}$ (Resolved only)		
Variable	Topology	0-lepton	1-lepton	2-lepton	0-lepton	1-lepton	2-lepton
$m_{j_1 j_2}$	R	×	×	×	×	×	×
m_J	B	×	×	×			
$m_{j_1 j_2 j_3}$	R (3-jets only)	×	×	×	×	×	×
$p_{\mathrm{T}}^{j_1}$	R	×	×	×	×	×	×
$p_{\mathrm{T}}^{j_1,\mathrm{trk}}$	B	×	×	×			
$p_{\mathrm{T}}^{j_2}$	R	×	×	×	×	×	×
$p_{\mathrm{T}}^{j_2,\mathrm{trk}}$	B	×	×	×			
$p_{\mathrm{T}}^{j_3}$	R (3-jets only)	×	×	×	×	×	×
$p_{\mathrm{T}}^{j_3,\mathrm{trk}}$	B	×	×	×			
$\Delta R(\boldsymbol{j_1}, \boldsymbol{j_2})$	R	×	×	×	×	×	×
$\Delta R(\boldsymbol{j_{1,trk}}, \boldsymbol{j_{2,trk}})$	B	×	×	×			
$\lvert\Delta\eta(\boldsymbol{j_1}, \boldsymbol{j_2})\rvert$	R	×			×		
$\lvert\Delta\phi(j_1, j_2)\rvert$	R				×		
$\mathrm{bin}_{\mathrm{DL1r}}(j_1)$	R	×	×		×	×	
$\mathrm{bin}_{\mathrm{DL1r}}(j_{1,\mathrm{trk}})$	B	×	×	×			
$\mathrm{bin}_{\mathrm{DL1r}}(j_2)$	R	×	×		×	×	
$\mathrm{bin}_{\mathrm{DL1r}}(j_{2,\mathrm{trk}})$	B	×	×	×			
p_{T}^{V}	R & B	$\equiv E_{\mathrm{T}}^{\mathrm{miss}}$	×	×	$\equiv E_{\mathrm{T}}^{\mathrm{miss}}$	×	×
$E_{\mathrm{T}}^{\mathrm{miss}}$	R	×	×		×	×	
$E_{\mathrm{T}}^{\mathrm{miss}}$	B	×					
$E_{\mathrm{T}}^{\mathrm{miss}}/\sqrt{S_{\mathrm{T}}}$	R			×			×
$\lvert\Delta y(\boldsymbol{V}, \boldsymbol{H_{\mathrm{cand}}})\rvert$	R & B		×			×	
$\lvert\Delta\phi(\boldsymbol{V}, \boldsymbol{H_{\mathrm{cand}}})\rvert$	R & B	×	×		×	×	
$\lvert\Delta\eta(\boldsymbol{V}, \boldsymbol{H_{\mathrm{cand}}})\rvert$	R			×			×
$\min[\Delta\phi(\boldsymbol{\ell}, \boldsymbol{b} \text{ or } \boldsymbol{c})]$	R		×			×	
m_{eff}	R	×			×		
$p_{\mathrm{T}}^{\mathrm{miss,st}}$	R	×			×		
m_{T}^{W}	R		×			×	
m_{top}	R		×			×	
$m_{\ell\ell}$	R			×			×
$\cos\theta(\boldsymbol{\ell^-}, \boldsymbol{Z})$	R			×			×
$(p_{\mathrm{T}}^{l_1} - p_{\mathrm{T}}^{l_2})/p_{\mathrm{T}}^{Z}$	R & B			×			×
N(track-jets in J)	B	×	×	×			
N(add. small-R jets)	B	×	×	×			

Table 5.6: MVA variables used for the 0-, 1- and 2-lepton channels in the resolved (denoted R) and boosted (denoted B) topologies for the $VH, H \to b\bar{b}$ and $VH, H \to c\bar{c}$ analyses. The $VH, H \to c\bar{c}$ analysis only probes the resolved topology. The Higgs candidate system, composed of the 2 small-R jets ($b\bar{b}$ or $c\bar{c}$) for the resolved regime or of the leading large-R jet (J) of the event for the boosted regime, is sometimes denoted H_{cand}.

5.7 Fit formalism

The signal and background yields are determined through maximum likelihood fits to the distributions of the boosted decision tree output in data for each topology, channel and signal region that were presented in the previous Section 5.6. Background contributions are normalized thanks to the control regions defined in Section 5.3, the procedure will be emphasized in the next Section 5.8.

This section describes the fit formalism in detail. First, Section 5.7.1 explains the hypothesis testing procedure, defines the p-value and its relation with the so-called significance. Then in Section 5.7.2, the likelihood function and the different types of nuisance parameters used in the analysis are presented. The determination, thanks to the minimization of the likelihood function, of the observed significance and of the parameters of interest such as the signal strength are also explained. On the other hand, Section 5.7.3 focuses on the computation of the expected significance by means of an Asimov dataset. The Sections 5.7.4 and 5.7.5 are highlighting the different and important cross-checks that allow to understand how the nuisance parameters affect the result of a fit in particular by looking at the pulls and the impact of the nuisance parameters, the correlation matrix between them, and the breakdown of uncertainties. The Section 5.7.6 is explaining how the Asimov significance can be computed from a distribution. Finally, the transformation D which is a rebinning applied to the BDT output distributions in order to enhance the signal from background discrimination is presented in Section 5.7.7.

5.7.1 Hypothesis testing, p-value and significance

In the context of searches for a signal process, the null hypothesis (H_0) corresponds to the assumption that the observed experimental data can be fully described with background only processes while the alternative hypothesis (H_1) includes both background and the sought signal. On the other hand, for setting limits, the signal-plus-background model is taken as the H_0 hypothesis and is tested against the background-only hypothesis H_1.

The p-value of a hypothesis (H) corresponds to the probability of finding data sets with equal or greater incompatibility than the observed data assuming the hypothesis H. Hence, the p-value quantifies the level of agreement of the measured data for a given hypothesis. In other word, for signal searches, when testing the background only hypothesis H_0, the associated probability (p) found can be interpreted as the chances that the discrepancies observed in data are only due to statistical fluctuations in the measurements: the lower is p the less likely such discrepancies can be explained only by fluctuations.

In particle physics, the significance[3] (S, expressed in standard deviation denoted σ) is usually preferred and is computed with the p-value as [320]

$$S = \Phi^{-1}(1-p) \Leftrightarrow p = \int_S^{+\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx = 1 - \Phi(S), \tag{5.7.1}$$

with Φ^{-1} the *quantile of the standard normal (Gaussian) distribution i.e.* the inverse of the cumulative distribution function (CDF) of the standard Gaussian. For a significance exceeding $3\,\sigma$ it is commonly referred to as the evidence of the corresponding signal process while above $5\,\sigma$ it is the observation *i.e.* the official discovery of a new signal process as the p-value in that case is $p < 3 \times 10^{-7}$. Hence, the background only hypothesis can be rejected in favor of the signal-plus-background assumption. For a signal hypothesis rejection, a threshold p-value greater than 0.05 corresponding to a 95% confidence level is often used. The Table 5.7 summarizes some characteristic values for the significances, the p-values and the confidence levels.

3. The significance is usually denoted Z instead of S. This choice of different notation was made to avoid confusion with another unrelated notion called the transformation D which is detailed in Section 5.7.7.

S	p-value	CL
$1\,\sigma$	15.9%	84.1%
$1.645\,\sigma$	5%	95%
$2\,\sigma$	2.28%	97.7%
$2.326\,\sigma$	1%	99%
$3\,\sigma$	0.135%	99.9%
$4\,\sigma$	3.16×10^{-5}	99.997%
$5\,\sigma$	2.87×10^{-7}	99.99997%

Table 5.7: Significances (S), p-values and confidence levels (CL $= 1 - p$).

5.7.2 Likelihood function, signal strength and observed significance

For a given distribution of interest, with N bins containing respectively $\boldsymbol{n} = (n_1, \ldots, n_N)$ events, of a variable x (*e.g.* the BDT output distribution or the invariant mass etc) which has been experimentally measured in data, then the expectation value ($E[n_i]$) for the content (n_i) of the i^{th} bin in the context of the signal-plus-background hypothesis can be expressed as

$$E[n_i] = \mu s_i + b_i, \tag{5.7.2}$$

with s_i and b_i the number of signal and background events expected for that bin which can be calculated from the respective probability density function (PDF) f_s and f_b for signal and background as

$$s_i = s_{\text{tot}} \int_{\text{bin } i} f_s(x; \boldsymbol{\theta_s}) dx \qquad b_i = b_{\text{tot}} \int_{\text{bin } i} f_b(x; \boldsymbol{\theta_b}) dx, \tag{5.7.3}$$

s_{tot} and b_{tot} being the total number of signal and background events as predicted by the simulation and $\boldsymbol{\theta_s}$ and $\boldsymbol{\theta_b}$ are the set of parameters controlling the shapes of the PDFs. The parameter μ is the so-called *signal strength* as $\mu = 0$ corresponds to the background only hypothesis, while $\mu = 1$ is for the nominal signal-plus-background hypothesis. The signal strength can be defined as the ratio

$$\mu := \frac{\sigma \cdot \text{BR}}{\sigma_{\text{SM}} \cdot \text{BR}_{\text{SM}}} \tag{5.7.4}$$

of the product of the cross-section and the branching ratio of the searched signal process measured in data over the same product but predicted by the Standard Model.

Let $\boldsymbol{\theta} = (\boldsymbol{\theta_s}, \boldsymbol{\theta_b}, b_{\text{tot}})$ be the set of all *nuisance parameters* (NPs) *i.e.* the set of parameters which must be accounted for in the hypothesis testing but which are not of interest. The parameter μ is the only *parameter of interest* (POI) in this context while s_{tot} is kept constant and computed thanks to the simulation.

For the $VH, H \to b\bar{b}/c\bar{c}$ analyses, a binned likelihood function is used to measure the signal strength thanks to the BDT output distributions. It can be decomposed in the product of 3 different likelihoods

$$\mathcal{L}(\mu, \boldsymbol{\theta}) = \mathcal{L}_{\text{EML}}(\mu, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\tau}) \cdot \mathcal{L}_{\text{syst}}(\boldsymbol{\alpha}) \cdot \mathcal{L}_{\text{stat}}(\boldsymbol{\gamma}), \tag{5.7.5}$$

with:

➢ μ the parameter of interest,

➢ $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\tau})$ the nuisance parameters which are respectively:

- $\boldsymbol{\alpha}$ the NPs for the experimental and modelling uncertainties. By convention $\alpha_j = 0$ is for the nominal prediction while $\alpha_j = \pm 1$ are associated to the $\pm 1\sigma$ variations related to a given uncertainty, uncertainty which can affect one or several processes.

- $\boldsymbol{\gamma}$ are the NPs for the MC statistical uncertainty on the sum of background processes in each bin (one NP per bin) [321]: the normalization of the sum of backgrounds (b_i) in each bin is modified with a multiplicative factor γ_i such that $\gamma_i b_i$ is allowed to vary within the background MC statistical uncertainty of the bin i considered. Hence, if $\gamma_i = 1$ the normalization of the background is unchanged.
- $\boldsymbol{\tau}$ are the floating normalizations for the different background processes which are free parameters determined during the fit without any prior.

➢ $\mathcal{L}_{\text{EML}}$ is the *extended maximum likelihood*. The signal-plus-background events are assumed to follow a Poisson distribution as events are occurring independently and the average rate at which they occur is independent of any occurrences. Moreover, the content of each bin of the distribution of interest is independent from the other bins, hence the likelihood is simply the product of the Poisson distributions in each bin (considering all SRs and CRs)

$$\mathcal{L}_{\text{EML}}(\mu, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\tau}) = \prod_{i=1}^{N_{\text{bins}}} \frac{(\mu s_i(\boldsymbol{\alpha}) + b_i(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\tau}))^{n_i}}{n_i!} \cdot e^{-(\mu s_i(\boldsymbol{\alpha}) + b_i(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\tau}))}. \tag{5.7.6}$$

➢ $\mathcal{L}_{\text{syst}}$ is the likelihood associated to the experimental and modelling uncertainties. Under the assumption that the systematic uncertainties are not correlated, it is computed as the product of Gaussian distributions

$$\mathcal{L}_{\text{syst}}(\boldsymbol{\alpha}) = \prod_{j=1}^{N_{\text{syst}}} \frac{e^{-\frac{\alpha_j^2}{2}}}{\sqrt{2\pi}}. \tag{5.7.7}$$

➢ $\mathcal{L}_{\text{stat}}$ is the likelihood associated to the MC statistical uncertainty on the sum of all backgrounds and is obtained from the products of the (3-dimensional) Γ functions over all bins

$$\mathcal{L}_{\text{stat}}(\boldsymbol{\gamma}) = \prod_{i=1}^{N_{\text{bins}}} \Gamma(\gamma_i b_i, b_i + 1, 1) = \prod_{i=1}^{N_{\text{bins}}} \frac{(\gamma_i b_i)^{b_i}}{\Gamma(b_i + 1)} \cdot e^{-\gamma_i b_i}, \tag{5.7.8}$$

with

$$\Gamma(\mu, \alpha, \beta) = \frac{\beta^\alpha \mu^{\alpha-1}}{\Gamma(\alpha)} \cdot e^{-\beta\mu}, \tag{5.7.9}$$

and the so called (1-dimensional) Γ function

$$\Gamma(\alpha) = \int_0^{+\infty} t^{\alpha-1} e^{-t} dt. \tag{5.7.10}$$

The terms in the product of the Equation (5.7.8) are the probabilities for each bin i to have b_i background events predicted when the true number of events is $\gamma_i b_i$. Moreover, those terms are indeed "Poisson terms" as $\Gamma(n+1) = n!$ for any positive integer n. Since b_i is not necessarily an integer, the Γ function is needed in the denominator instead of the factorial term.

The best agreement between data and simulation is found when maximizing the likelihood function defined in Equation (5.7.5) or equivalently minimizing the *negative log-likelihood* ($-2\ln(\mathcal{L})$). Two different types of fits can be distinguished:

➢ **the conditional fits**: the likelihood is maximized for a particular (fixed) value of the signal strength and only the NPs are allowed to vary.

➢ **The unconditional fit**: the likelihood is maximized without fixing any POIs and NPs, allowing to measure the signal strength in data.

The profile likelihood ratio to test a hypothesized value of μ can be defined as:

$$\lambda(\mu) = \frac{\mathcal{L}(\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu))}{\mathcal{L}(\hat{\mu}, \hat{\boldsymbol{\theta}}(\hat{\mu}))} \tag{5.7.11}$$

where $\hat{\hat{\boldsymbol{\theta}}}(\mu)$, in the numerator, is referred to as the *conditional maximum likelihood estimator* and corresponds to the value of $\boldsymbol{\theta}$ obtained from the conditional fit with the specific signal strength value μ, while $\hat{\mu}$ and $\hat{\boldsymbol{\theta}}(\hat{\mu})$ correspond to the *unconditional maximum likelihood estimators i.e.* the parameters for which the likelihood is maximal and are derived with the unconditional fit. $\lambda(\mu)$ can be equivalently replaced by:

$$t_\mu = -2\ln(\lambda(\mu)). \tag{5.7.12}$$

Based on its definition, $0 \leq \lambda(\mu) \leq 1$ and λ close to 1 (respectively t_μ close to 0) indicates a good agreement between data and the hypothesized value of μ while if λ is near 0 (respectively t_μ tends to $+\infty$) it implies an incompatibility. As a result t_μ quantifies the level of disagreement with data, level which can be assessed for the hypothesized value μ computing the p-value hence the significance as

$$p_\mu = \int_{t_{\mu_{\text{obs}}}}^{+\infty} f(t_\mu|\mu)dt_\mu, \tag{5.7.13}$$

with μ_{obs} the signal strength value observed in data and $f(t_\mu|\mu)$ the PDF for t_μ under the assumption of the signal strength μ. In particular, for a positive signal test discovery, one can define the test statistic for the null hypothesis *i.e.* for $\mu = 0$ as

$$q_0 = \begin{cases} -2\ln(\lambda(0)) & \text{if } \hat{\mu} \geq 0 \\ 0 & \text{if } \hat{\mu} < 0 \end{cases}, \tag{5.7.14}$$

with $\lambda(0)$ the profile likelihood ratio for $\mu = 0$ as defined in Equation (5.7.11). Then the p-value quantifying the disagreement between the data and the background only hypothesis is

$$p_0 = \int_{q_{0_{\text{obs}}}}^{+\infty} f(q_0|0)dq_0, \tag{5.7.15}$$

where $f(q_0|0)$ is the PDF of the statistic q_0 assuming the background only hypothesis.

Most of the time determining analytically the PDF $f(q_0|0)$ is difficult if not impossible leading to the use of estimation techniques to deduce that PDF as for instance the Monte Carlo method [322]. Indeed, a simpler and less computational expensive solution exists, as if the background only hypothesis is true, then the Wilks' theorem [323, 324], ensures that t_0 is asymptotically following a χ^2 distribution with a number of degree of freedom being equal to the difference in dimensions between the numerator and denominator in the Equation (5.7.11). Since in that equation, μ is a fixed parameter only in the numerator, q_0 is distributed as a χ^2 with one degree of freedom and so the significance (S_0) for the background hypothesis can simply be computed as [320]:

$$S_0 = \Phi^{-1}(1 - p_0) = \sqrt{q_0}. \tag{5.7.16}$$

Indeed, the Wilks' theorem is valid for any test value μ: t_μ follows a χ^2 distribution of degree 1 (as here only one POI is considered). As for the $\Delta\chi^2 = 1$ criterion to find the statistical uncertainty for the parameter of interest in a χ^2 context as presented in Section 4.4.2, the statistical uncertainties $\sigma_\mu^\pm$ for the signal strength $\hat{\mu}^{+\sigma_\mu^+}_{-\sigma_\mu^-}$ that maximizes the likelihood, can be obtained finding the two value $\mu_\pm$ such that [323, 324]

$$t_{\mu_\pm} - t_{\hat{\mu}} = t_{\mu_\pm} = -2\ln\left(\frac{\mathcal{L}(\mu_\pm, \hat{\hat{\boldsymbol{\theta}}}(\mu_\pm))}{\mathcal{L}(\hat{\mu}, \hat{\boldsymbol{\theta}}(\hat{\mu}))}\right) = 1, \tag{5.7.17}$$

as shown in Figure 5.20, and then $\sigma_\mu^\pm = |\mu_\pm - \hat{\mu}|$.

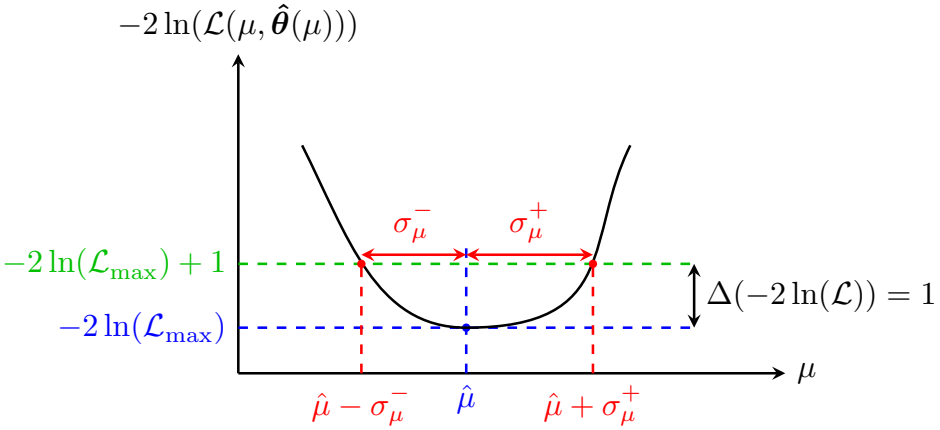


Figure 5.20: Scheme of the determination of the statistical uncertainties $\sigma_\mu^\pm$ on the signal strength $\hat{\mu}$ that minimizes the negative log-likelihood and $\mathcal{L}_{\max} = \mathcal{L}(\hat{\mu}, \hat{\boldsymbol{\theta}}(\hat{\mu}))$ being the maximum likelihood value.

5.7.3 Asimov dataset and expected significance

For a test of the signal strength parameter μ, either $\mu = 0$ for a discovery or $\mu \neq 0$ for an upper limit, and for a dataset distributed according to a signal strength parameter μ' then Wald [320, 325] proved that for a single POI:

$$-2\ln(\lambda(\mu)) = \frac{(\mu - \hat{\mu})^2}{\sigma^2} + \mathcal{O}(1/\sqrt{N}), \tag{5.7.18}$$

with $\hat{\mu}$ following a Gaussian distribution with a mean μ' and standard deviation σ, N being the data sample size. The standard deviation (σ) of $\hat{\mu}$ can be computed thanks to the covariance matrix of the estimators of all parameters (POI and NPs) $V_{ij} = \mathrm{cov}(\hat{\vartheta}_i, \hat{\vartheta}_j)$ with $\boldsymbol{\vartheta} = (\mu, \boldsymbol{\theta})$ *i.e.* ϑ_i can be the POI or an NP. In the large sample limit, the bias of the maximum likelihood estimators (MLEs) tends to zero and $\mathcal{O}(1/\sqrt{N})$ can be neglected leading to the formula

$$V_{ij}^{-1} = \mathrm{cov}(\hat{\vartheta}_i, \hat{\vartheta}_j)^{-1} = -E\left[\frac{\partial^2 \ln(\mathcal{L}(\boldsymbol{\vartheta}))}{\partial\vartheta_i \partial\vartheta_j}\right], \tag{5.7.19}$$

that allows for an estimation of σ since $\vartheta_0 = \mu$ implying that $\sigma^2 = V_{00}$.

An Asimov dataset is an artificial representative (simulated) distribution (or sample) used to calculate the expected sensitivities and which is generated for a given signal strength μ' *e.g.* $\mu' = 1$. An Asimov dataset is defined by the fact that the true parameter values used to generate that dataset should be obtained when estimating all the ML parameters. Such dataset is found by imposing the derivatives of the log-likelihood, with respect to each parameter, to be equal to 0 [320]:

$$\left.\frac{\partial \ln(\mathcal{L}(\boldsymbol{\vartheta}))}{\partial\vartheta_i}\right|_{\mu=\mu'} = 0, \tag{5.7.20}$$

hence all MLEs ϑ_i are determined which allows to generate the Asimov dataset.

To characterize the sensitivity of an experiment to a potential signal discovery, the expected (median) significance should be determined as indeed one is not interested in the significance for a single dataset but rather in the significance for which one would be able to reject several μ values. In particular, in the context of a discovery one would like to estimate the median significance under the nominal signal hypothesis ($\mu = 1$) and reject the background only hypothesis ($\mu = 0$). Two possible methods exist: either generating a large quantity of MC datasets which is computationally

expensive and time-consuming or use approximation formula (approach described below) which is much simpler and faster.

The expected probability ($p_{0_{\text{exp}}}$) value for a given signal strength μ' can be expressed as

$$p_{0_{\text{exp}}} = \int_{t_{0_{\text{exp}}}}^{+\infty} f(t_0|\mu')dt_0, \tag{5.7.21}$$

with $t_{0_{\text{exp}}} = \text{med}[t_0|\mu']$ the median value of t_0 for an assumed signal strength μ' which could be determined with $f(t_0|\mu')$ but indeed would require estimating the PDF $f(t_0|\mu')$. For the Asimov dataset, from Equation (5.7.18), follows that:

$$\sigma_A^2 \approx \frac{(\mu-\mu')^2}{t_{\mu,A}} \underset{\substack{\text{for a discovery,}\\ \text{test } \mu=0}}{=} \frac{\mu'^2}{t_{0,A}} \Leftrightarrow t_{0,A} = \frac{\mu'^2}{\sigma_A^2} \tag{5.7.22}$$

with $\sigma_A^2 = V_{00}$ being the standard deviation *i.e.* the uncertainty on the estimation of the signal strength μ' of the Asimov dataset providing the expected sensitivity to μ' and $t_{\mu,A} = -2\ln(\lambda_A(\mu))$. Therefore, $t_{0,A}$ can be deduced from Equation (5.7.22) and the expected significance follows as one can show that it is equal to:

$$\text{med}[S_0|\mu'] = \sqrt{t_{0,A}}. \tag{5.7.23}$$

5.7.4 Pulls, impacts of nuisance parameters and correlations in fit

The results of a fit can be visualized and checked with some useful quantities and plots to understand the impact of the nuisance parameters on the fit.

For each NP θ_i, its so-called *pull* can be defined as [326]

$$\text{pull}(\theta_i) = \frac{\hat{\theta}_i - \theta_i^{\text{exp}}}{\sigma_{\theta_i^{\text{exp}}}}, \tag{5.7.24}$$

with $\hat{\theta}_i$ the NP value obtained from the maximum likelihood fit, θ_i^{exp} the value expected for that NP and $\sigma_{\theta_i^{\text{exp}}}$ the uncertainty expected for that parameter. $\hat{\theta}_i$ and θ_i^{exp} are also referred to as post-fit and pre-fit NP values and $\sigma_{\theta_i^{\text{exp}}}$ as the pre-fit uncertainty. For instance in the case of the $VH, H \to b\bar{b}/c\bar{c}$ fits, the NPs defined in Equation (5.7.5) are expected to be equal to $\alpha_i^{\text{exp}} = 0$ for the experimental and modelling uncertainties, and for the normalization related NPs $\gamma_i^{\text{exp}} = 1$ and $\tau_i^{\text{exp}} = 1$. As a result the pull of a NP quantifies how far is the NP from the expected value and in case the pull is not compatible with zero the reason should be investigated. Moreover, if the size of the error on the pull ($\delta_{\text{pull}(\theta_i)}$) is smaller (respectively larger) than unity it means that the NP is more (respectively less) constrained by the data than expected which should be understood.

The *impact* of a NP for a given POI, such as the signal strength μ, is defined as [326]

$$\text{impact}(\theta_i) = \Delta\mu^{\pm}(\theta_i) = \hat{\hat{\mu}}_{\theta_i \pm \sigma_{\theta_i}} - \hat{\mu}, \tag{5.7.25}$$

with $\hat{\hat{\mu}}_{\theta_i \pm \sigma_{\theta_i}}$ the maximum likelihood estimator for the signal strength where all parameters are allowed to vary except θ_i which is fixed at its most extreme values $\theta_i \pm \sigma_{\theta_i}$. The impact of a NP can be assessed using the pre-fit or post-fit values of $\theta_i \pm \sigma_{\theta_i}$, they are hence respectively called pre-fit and post-fit impact of the NP. The impact of a NP allows assessing by how much the NP can affect the determination of the signal strength as not all the NPs are equally important.

The correlations between the fit parameters (POIs and NPs), which can be derived thanks to the Equation (5.7.19) in the large sample limit, are by definition contained in the interval $-1 \leq$

$\text{cov}(\hat{\vartheta}_i, \hat{\vartheta}_j) \leq 1$. Two parameters with a correlation close to 1 are referred to as fully correlated implying that if the value of a parameter is increased (respectively decreased) then the other parameter value will increase (respectively decrease) as well. If the correlation is close to -1, the parameters are anti-correlated their value changes in opposite directions an increase of one of them will decrease the value of the other. Finally, for a correlation close to 0 the parameters are said to be uncorrelated meaning that a variation of one of them will not affect the other parameter. Usually some NPs are expected to be correlated but most of them should be uncorrelated, if non-expected correlations are observed this could indicate a problem in the fit procedure or fit model. In addition, the covariance matrix obtained helps to understand why some NPs are constrained or pulled.

5.7.5 Breakdown of uncertainties

Some uncertainties are usually grouped together as they are associated to a same type of systematic as for instance jet energy scale (JES) uncertainties or the b-tagging efficiency uncertainties and so on. In order to study the effect of a group of uncertainties, their impact is computed as:

$$\text{impact(group)} = \sqrt{\sigma^2_{\hat{\mu}} - \sigma^2_{\hat{\mu}_{\text{group}}}} \tag{5.7.26}$$

with $\sigma_{\hat{\mu}}$ being the uncertainty on the signal strength obtained for the maximum likelihood fit when all POIs and NPs are allowed to float, while $\sigma_{\hat{\mu}_{\text{group}}}$ corresponds to the uncertainty when the fit is performed fixing all the NPs of the group to their best fit values.

5.7.6 Asimov significance

An asymptotic formula for the significance from Equation (5.7.1) is provided by the Asimov significance (here assuming $\mu = 1$) which is computed from the BDT output distribution (or other distributions such as the mass) as follows [320, 326]:

$$S = \sqrt{\sum_i 2\left[(s_i + b_i)\ln\left(1 + \frac{s_i}{b_i}\right) - s_i\right]}, \tag{5.7.27}$$

with s_i and b_i being the total number of signal and background events contained in the i^{th} bin of the distribution of interest. The Asimov significance is in particular useful for studies which are carried out without relying on a fitting procedure. The associated statistical uncertainty (δS) on the Asimov significance by propagation of errors is equal to:

$$\delta S = \sqrt{\sum_i \left(\frac{\partial S}{\partial s_i}\delta s_i\right)^2 + \left(\frac{\partial S}{\partial b_i}\delta b_i\right)^2}, \tag{5.7.28}$$

where

$$\frac{\partial S}{\partial s_i} = \frac{1}{S}\ln\left(1 + \frac{s_i}{b_i}\right) \qquad \frac{\partial S}{\partial b_i} = \frac{1}{S}\left[\ln\left(1 + \frac{s_i}{b_i}\right) - \frac{s_i}{b_i}\right], \tag{5.7.29}$$

and δs_i and δb_i are the statistical uncertainties for signal and background events associated to the i^{th} bin which are calculated as

$$\delta s_i = \sqrt{\sum_{j \in \mathcal{S}_i} w_j^2}, \qquad \delta b_i = \sqrt{\sum_{j \in \mathcal{B}_i} w_j^2}, \tag{5.7.30}$$

with $\mathcal{S}_i$ and $\mathcal{B}_i$ being respectively the set of all signal and background events contributing to the i^{th} bin and w_j is the weight of such individual events.

5.7.7 Transformation D

For the BDT output distributions, prior to the computation of the significance, a transformation, called *transformation D*, is applied to the bins of that distribution in order to maximize the separation of signal from background events while requiring a minimum number of data and simulation events in each bin. The bins of the original distribution are merged together as the low BDT scores bins are almost only populated by background events while for the high BDT score bins the signal to background ratio is the highest. In this merging procedure, the statistical uncertainties are taken into account. The aim of the transformation D is to obtain a smoother distribution both for the background and the signal while having a fine enough binning for the bins with the largest signal contribution and keeping the statistical uncertainty under a certain threshold for each bin.

A general formula describing such merging, also referred to as remapping, is given by

$$Z_i(j,k) = Z(z_s, z_b, N_s, N_b, n_s(j,k), n_b(j,k)), \tag{5.7.31}$$

where:

- $Z_i(j,k)$ is the i^{th} bin of the transformed distribution and is obtained by merging all bins from the original distribution between the j^{th} and the k^{th} bins included.
- z_s and z_b are parameters used to tune the merging.
- N_s and N_b are respectively the total number of signal and background events.
- $n_s(j,k)$ and $n_b(j,k)$ are respectively the total number of signal and background events contained between the j^{th} and k^{th} bins of the original distribution.

The Equation (5.7.31) is a general formula, several possible ways to remap a distribution exist, the transformation D is one of them and is defined by

$$Z_i(j, k_{\text{last}}) = z_s \frac{n_s(j, k_{\text{last}})}{N_s} + z_b \frac{n_b(j, k_{\text{last}})}{N_b}. \tag{5.7.32}$$

The original BDT distributions have 500 equidistant bins. For the transformation D, the bins are merged with the following steps below:

- starting from the highest bin (k_{last}), *i.e.* the last bin on the right of the original distribution. The range of the interval is increased by adding, one after the other, the bins on the left of the k_{last} bin.
- The value $Z(j, k_{\text{last}})$ is computed at each step until both $Z(j, k_{\text{last}}) > 1$ and the statistical uncertainty obtained merging all bins between j and k_{last} is smaller than 20%. Once those criterions are reached for a given bin j, all the bins between j and k_{last} included are merged into a single one.
- The process is repeated starting this time from the new last bin $k_{\text{last}} = j - 1$. It stops when reaching the first bin *i.e.* the bin on the left of the original distribution.

The bins of the distribution after transformation D have variable widths. For aesthetic reasons, the width of the bins is usually simply modified so that the transformed distribution is plotted with the same width for each bin.

Due to its definition in Equation (5.7.32) and its implementation, when using the transformation D, the sum $z_s + z_b$ of the two hyperparameters of the transformation D is equal to the number of bins of the transformed distribution. In extreme cases if $z_s = 0$ and $z_b > 0$ then $N_{\text{bin}} = z_b$ the background would be equally distributed among bins (same for signal if $z_s > 0$ and $z_b = 0$). Hence, z_s and z_b control the shape of the signal and background transformed BDT distributions, shapes which are not of interest for the physical interpretations but which are important for signal-background discrimination.

The hyperparameters are tuned such that the number of bins remains reasonable to have enough statistics while at the same time ensuring sensitivity to signal. Their optimization is still under study but typically for the $VH, H \to b\bar{b}$ analysis, for the resolved topology $(z_s, z_b) = (10, 5)$ while for the boosted topology $(z_s, z_b) = (3, 2)$ as statistics decreases in the high-p_{T}^{V} regions. For the $VH, H \to c\bar{c}$ analysis they are set to $(z_s, z_b) = (5, 5)$.

5.8 Background normalizations and systematic uncertainties

The current section describes the background normalization procedure and the uncertainties to be taken into account in the likelihood fit model. Such uncertainties can indeed be split in two categories: the modelling and the experimental uncertainties detailed hereafter.

The ultimate goal of the $VH, H \to b\bar{b}/c\bar{c}$ analysis is to perform a simultaneous $VH, H \to b\bar{b}$ and $VH, H \to c\bar{c}$ fit in order to jointly determine the corresponding signal strengths $(\mu_{VH}^{H\to b\bar{b}}, \mu_{VH}^{H\to c\bar{c}})$. It requires to build a single likelihood model that takes into account all the analysis regions. However, as I am writing this section, the $VH, H \to b\bar{b}$ and $VH, H \to c\bar{c}$ combination in a unique likelihood fit has not yet been implemented, and the $VH(b\bar{b}/c\bar{c})$ analysis team has just started the modelling studies. As a result, the current description of the fit and modelling uncertainties is heavily based on the previous publications [5, 6, 65, 275] and may not perfectly reflect the final choices concerning the incoming Legacy Run 2 publication.

Moreover, this section mainly focuses on the $VH, H \to b\bar{b}$ analysis. The reader is referred to the former ATLAS $VH, H \to c\bar{c}$ publication [6] for more details about the fit model in particular the background normalization which is different for some aspects from the $VH, H \to b\bar{b}$ analysis. A brief description of the control regions that differ from the $VH, H \to b\bar{b}$ analysis has been provided in Section 5.3.

5.8.1 Background normalizations and modelling uncertainties

The fits follow the division of phase space of the $VH, H \to b\bar{b}$ and $VH, H \to c\bar{c}$ analyses, in the 0- and 1-lepton channels, the event reconstruction starts for $p_{\mathrm{T}}^{V} > 150$ GeV, while the 2-lepton channel also exploits the $[75, 150]$ GeV p_{T}^{V}-region. In total for the resolved topology 3 p_{T}^{V}-region are defined: the $[75, 150]$ GeV (2-lepton channel only), the $[150, 250]$ GeV and the $[250, 400]$ GeV p_{T}^{V}-bins splitting events in the 2- and 3-jets ($\geq$ 3-jets in the 2-lepton channel).

For the $VH, H \to b\bar{b}$ boosted topology, the phase space is split in 2 regions: the $[400, 600]$ GeV and the $p_{\mathrm{T}}^{V} > 600$ GeV p_{T}^{V}-bins. The $p_{\mathrm{T}}^{V} > 600$ GeV is a new region introduced for the Legacy Run 2 analysis as previously the measurement was performed inclusively for $p_{\mathrm{T}}^{V} > 400$ GeV. The boosted signal region is divided in two categories the high purity (no additional small-R jet) and the low purity (at least one additional small-R jet) for the 0- and 1-lepton channels. In the $p_{\mathrm{T}}^{V} > 600$ GeV-region, to address the decrease in the number of events the HP and LP are merged into a single signal region. For the same reasons, the 2-lepton channel only consist in a 1 signal region without even a top control region.

For the $VH, H \to b\bar{b}$ analysis, the V + jets simulated events are categorized based on the true flavour of the two b-tagged small-R jets (or track-jets) as bb, bc, bl, cc, cl and l components. The heavy flavour (HF) component is referring to the sum of $bb + bc + bl + cc$ contributions. The leading backgrounds normalizations in each channel, *i.e.* $t\bar{t}$, W + HF and Z + HF, are floating parameters of the fit and are determined from data by means of dedicated control regions: normalizations are obtained from the regions with the highest background purity and propagated to the other channels. The normalizations are furthermore determined independently per p_{T}^{V}-regions and categories

(2-jets, 3-jets for the resolved regime and the HP and LP for the boosted regime) as shown in Figure 5.21.

For the resolved topology, for the $VH, H \to b\bar{b}$ analysis all the normalizations are obtained simply from the yields in CRs. The $t\bar{t}$ and the $W + \text{HF}$ normalizations are respectively obtained from the yields of the high ΔR and the low ΔR CRs in the 1-lepton channel and are extrapolated to the 0-lepton channel (the $t\bar{t}$ contribution being determined with a data-driven technique in the 2-lepton channel). Moreover, for the $VH, H \to b\bar{b}/c\bar{c}$ simultaneous fit, the addition of the common top CNB control region (discussed in Section 5.3.8) will allow constraining in particular the *top(lq)* ($q = c, l$) component both for the $VH, H \to b\bar{b}$ and $VH, H \to c\bar{c}$ analyses. The $Z + \text{HF}$ normalization is computed from the yields of the high ΔR CR of the 2-lepton channel and is extrapolated to the 0-lepton channel. The $VH, H \to c\bar{c}$ analysis also relies on the shape of the m_{cc} distribution in CRs.

The boosted topology only has a top control region defined in the 0- and 1-lepton channels. The $t\bar{t}$ normalization is obtained from the invariant mass distribution of the leading large-R jet (m_J) in that top CR. On the other hand, the normalizations of the $W + \text{HF}$ and $Z + \text{HF}$ background are extrapolated from the resolved $[250, 400]$ GeV p_{T}^V-region and uncertainties related to that extrapolation (both for the extrapolation of the normalization and the modelling on the relative acceptance between the boosted and resolved regimes) are taken into account in the likelihood fit model.

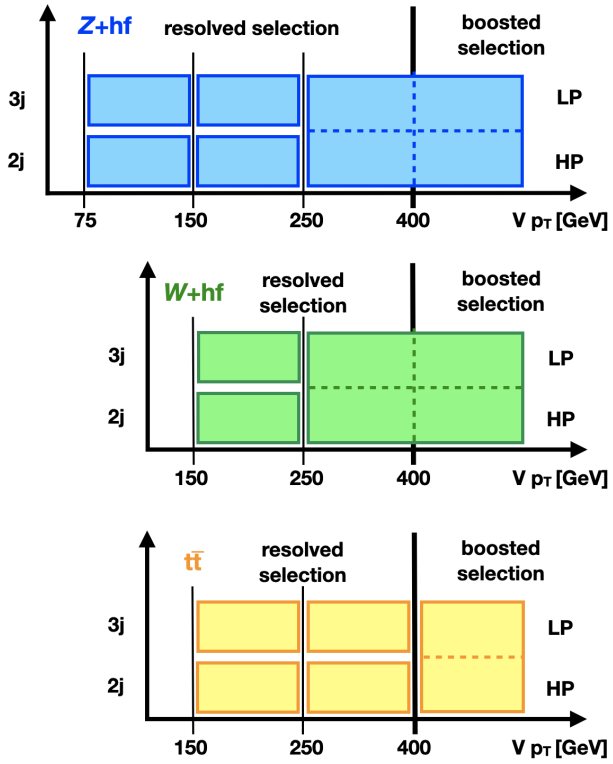


Figure 5.21: Division of phase space for the determination of the floating normalization factors for the $Z + \text{HF}$ background in the 0- and 2-lepton channels (top), the $W + \text{HF}$ background in the 0- and 1-lepton channels (middle) and the $t\bar{t}$ background in the 0- and 1-lepton channels (bottom). The main categorization of events in multiple p_{T}^V regions and different jet multiplicity requirements is shown. Regions isolated by solid lines receive independent normalization factors. Regions separated by dashed lines share a common normalization factor and dedicated extrapolation uncertainties are implemented among them (taken from Ref. [5]).

Only the uncertainties related to the flavour composition of the $V + \text{jets}$ background are treated as correlated between the resolved and boosted regimes. The remaining $t\bar{t}$ and $V + \text{HF}$ nuisance parameters dedicated to the description of those backgrounds are considered uncorrelated between

the resolved and the boosted analysis regions due to the difference in kinematic regimes covered by the two selections. It concerns in particular all the background modelling uncertainties affecting the shape of the fitted distributions.

The modelling uncertainties cover three different effects: uncertainties for the normalization of the samples, and/or the impact of the relative acceptance between the categories defined for the fit model and/or uncertainties on the shapes of the main discriminant variables (in the $VH, H \to b\bar{b}/c\bar{c}$ analysis BDT output distribution, the mass of the Higgs candidate and the transverse momentum of the vector boson). All those uncertainties are estimated by comparing the nominal MC prediction to an alternative set of predictions that are obtained under different hypotheses. For instance the uncertainties can originate from different choices for the renormalization and factorization scales (μ_R, μ_F), for the strong coupling constant α_s or for the parton distribution function used. It also concerns the effects of matrix element and parton showering models chosen.

- **The normalization uncertainties** are calculated as an overall uncertainty on the yield of the process considered for one or several regions. These uncertainties are derived in the region with the highest purity, usually a control region, and are extrapolated to the other regions: additional uncertainties are added to account for these extrapolations.
- **The acceptance uncertainties** account for possible changes in the relative fractions of events of events within the phase space considered by the analysis. The repartition of events could be affected for instance by the choice of the event generator. Those differences induce migration of events between regions and more generally results in a change of the ratio between the number of events across the regions as predicted by the nominal simulation and an alternative one. To assess and take into account such acceptance differences the following double-ratios between event categories are computed:

$$\delta_{A,B} = \left(\frac{N_A}{N_B}\right)_{\text{Alternative}} \Bigg/ \left(\frac{N_A}{N_B}\right)_{\text{Nominal}}, \tag{5.8.1}$$

  where A is the region with the highest purity (*e.g.* a control region) and B is the other region considered (*e.g.* a signal region), N_A and N_B being the number of events found respectively in those regions. For such acceptance uncertainties, the samples are normalized to the same cross-section because the yields differences are already accounted for by dedicated normalization uncertainties.
- **The shape uncertainties** are evaluating the possible effects on the shape of the most important variables of the analysis when comparing the nominal prediction to alternative simulations. They are computed using truth level quantities for the most discriminating variables of the BDT input variables ($m_{H_{\text{cand}}}$ and p_{T}^{V}) for different event generators including the nominal generator. Such uncertainties have been tested to be sufficiently large to cover the shape uncertainties of the other BDT input variables. To reduce statistical fluctuations the maximal difference between the nominal and alternative generators is fitted with a function (that can be different across different regions of phase space, channels…) and nominal events are reweighted with that fitting function to predict the shape of the alternative distribution. Another solution is to predict the shape of the alternative distribution with a machine learning reweighting technique [327] such as the BDT reweighting (denoted BDTr) or the newly developed neural network CARL which comes with improved performances and which is for technical reasons easier to use. The reweighting algorithms are trained to predict the shape of an alternative event generator distribution using the nominal prediction based on the characteristics of the nominal event ($p_{\text{T}}^{b_1}, p_{\text{T}}^{b_2}, m_{bb}, p_{\text{T}}^{V}, \ldots$). Basically they are able to transform the probability density function from one event generator to another by means of a reweighting of events. In both cases, either using a simple fit or a multidimensional algorithm, the usually low amount of simulated event in alternative samples is avoided, and thus it reduces statistical uncertainties as all distributions (the nominal and alternatives ones) are obtained only based on the nominal prediction (once the

reweighting of events has been determined). In any case, after reweighting each event an alternative BDT output distribution close to the prediction of the original alternative generators is obtained. Those alternative distributions are thus accounting for a shape uncertainty systematic and are less suffering from a lack of the amount of data compared to the original alternative samples.

The modelling uncertainties considered for the signal and background processes are described below.

- **Signal**: the systematic uncertainties affecting the calculations of the VH production cross-sections and the $H \to b\bar{b}$ branching fraction are assigned based on the recommendations of the LHC Higgs Cross-Section Working Group [57]. The acceptance and shape uncertainties are derived to account for missing higher-order QCD and EW corrections, for PDF+α_{S} uncertainties, and for variations of the ME, PS and UE models. The ME uncertainty is evaluated comparing the nominal (POWHEG) and alternative matrix-element generators (MADGRAPH5_AMC@NLO), both interfaced to the same program (PYTHIA 8) for parton showering. The PS and UE uncertainty is evaluated by comparing the nominal POWHEG BOX signal samples showered by PYTHIA 8 with alternative samples showered by HERWIG 7. Factorization and renormalization scales are varied by factors of 0.5 and 2. PDF-related uncertainties are derived using alternative PDF sets. The effects of the uncertainties from missing higher-order EW corrections, PDF+α_{S} and QCD scale variations on the jet mass shape are negligible.
- **$t\bar{t}$ background**: the ME and PS uncertainty in the 0- and 1-lepton channels (for the 2-lepton the $t\bar{t}$ contribution is data-driven) are estimated comparing POWHEG+PYTHIA 8 respectively to MADGRAPH5_AMC@NLO+PYTHIA 8 and POWHEG+HERWIG 7. The dominant flavour component in $t\bar{t}$ events is $b\bar{b}$. However, there is a non-negligible $b\bar{c}$ component which has a more signal-like topology. The relative composition uncertainties of the three flavour components ($b\bar{b}$, $b\bar{c}$, and other) are estimated from the difference in the ratio of the $b\bar{c}$ or other components with respect to the $b\bar{b}$ yield between the nominal sample and the alternative ME and PS uncertainty samples. Shape uncertainties are obtained by comparison of the nominal sample with the ME and PS samples in the 0+1 channels. Factorization and renormalization scales are varied by factors of 0.5 and 2. ISR and FSR uncertainties are also taken into account through special tune of PYTHIA 8.
- **Single top quark background**: for the Wt- and t-channels, uncertainties are derived for the normalization, relative acceptance and shapes of the m_{bb} (or m_J) and p_{T}^V distributions. Normalization uncertainties are obtained summing in quadrature the scale variations due to renormalization and factorization, the uncertainties on α_S and the PDF uncertainties. For the dominant Wt-channel, ISR/FSR uncertainties as well as a comparison to MADGRAPH5_AMC@NLO and HERWIG7 are taken into account. Differences between the diagram subtraction and the diagram removal schemes is considered to account for possible ambiguities and interferences between $t\bar{t}$ and Wt processes. The estimated modelling uncertainties are applied independently according to the flavour of the candidate system, due to the different regions of phase space being probed for the $b\bar{b}$, $b\bar{c}$ and other flavour components. Only a normalization uncertainty is derived for the s-channel, since its contribution is very low.
- **W + jets background**: the W + HF is the dominant background component in the 1-lepton channel. An extrapolation uncertainty is considered in the 0-lepton channel for the extrapolation from the 1-lepton to the 0-lepton channel. The remaining flavour components ($W + c\bar{l}$ and $W + l\bar{l}$) are strongly suppressed: only normalization uncertainties are included and are correlated across all regions. Flavour composition relative uncertainties for the $b\bar{c}$, $b\bar{l}$ and $c\bar{c}$ components with respect to the dominant $b\bar{b}$ component are considered for the W + HF normalization and are estimated separately in each lepton channel. Relative acceptance uncertainties are estimated for the ratio of the event yield in the SR to that in the CRs. Shape uncertainties are derived for p_{T}^V and using the reweighting method from comparisons of the nominal sample (SHERPA) with

an alternative sample (MADGRAPH5_AMC@NLO+PYTHIA 8).

- ➢ **Z + jets background**: the Z + HF process is the dominant background component in the 0- and 2-lepton channels. The remaining flavour components ($Z + cl$ and $Z + ll$) constitute a very small fraction of the total background yield thus only normalization uncertainties are included. Flavour composition relative uncertainties for the bc, bl and cc components with respect to the dominant bb component are considered for the Z + HF normalization and are estimated separately in each lepton channel. For the Z + HF background a relative acceptance uncertainty in the ratio of the event yield in the 0-lepton channel to that in the 2-lepton channel is considered. Relative acceptance uncertainties are estimated for the ratio of the event yield in the SR to that in the CRs. Shape uncertainties are derived from comparisons of the nominal sample (SHERPA) with an alternative sample (MADGRAPH5_AMC@NLO+PYTHIA 8). For the resolved topology, shape uncertainties are derived for m_{bb} and p_{T}^{V}, and are evaluated from comparisons with data in the m_{bb} distribution side-bands ($m_{bb} < 80$ GeV or $m_{bb} > 140$ GeV), after subtraction of the backgrounds other than Z + jets. For the boosted topology, shape uncertainties are derived for the m_J distribution.
- ➢ **Diboson background**: is composed of three production modes (WZ, WW and ZZ). The WW process has a very low contribution to the total background yields, thus only a normalization uncertainty is considered. The WZ and ZZ backgrounds, that have a more important contribution, are assigned the following uncertainties: normalization, relative acceptance between regions and the m_{bb} (or m_J) and p_{T}^{V} shapes. Those uncertainties are derived from comparison of the nominal simulation (SHERPA) with alternative samples (POWHEG+PYTHIA 8 and POWHEG+HERWIG++).
- ➢ **Multijet background**: its contribution is negligible and thus ignored in the 0- and 2-lepton channels. The MJ background is estimated in the 1-lepton channel with a data-driven procedure. Normalization and shape uncertainties are taken into account for the MJ background. The normalization uncertainty is estimated summing in quadrature several contributions. The lowest p_{T} lepton trigger instead of the usual OR combination is used. It enables to quantify the impact of the trigger choice. The impact of the extrapolation from the inverted isolation region is evaluated by tightening the isolation requirement. This results in the selection of events with a signature closer to that of the signal thus with a lower extrapolation uncertainty. The shape difference between the two selections is considered as uncertainty. The impact of the EW background normalization is obtained by extracting the MJ template with and without the normalization corrections applied to the EW components. The difference between the two templates is taken as an uncertainty. The shape uncertainties are obtained normalizing the template distributions to the nominal MJ template and are considered as uncorrelated between the electron and muon sub-channels.

5.8.2 Experimental uncertainties

The experimental uncertainties can be divided in the categories listed below. It concerns uncertainties affecting objects used by the analysis such as their calibrations, reconstruction or identification efficiencies and the experimental setup, i.e. luminosity or pile-up uncertainties. Those uncertainties are usually derived, by dedicated ATLAS working groups, in reference control samples with external measurements and are propagated to physics analyses.

- ➢ **Small-R jets.** All the systematic uncertainties concerning the small-R jets (b-tagging efficiency uncertainties, jet energy scale and energy resolution, η-intercalibration, flavor uncertainties, pile-up related uncertainties, punch through and single particle response uncertainties), have already been presented for the b-JES calibration in Chapter 4 in Section 4.4.3.
- ➢ **VR track-jets.** In this analysis VR track-jets are only used for testing the b-tagging requirement of large-R jets for the boosted $VH, H \to b\bar{b}$ analysis. The only uncertainty associated to them

is thus the b-tagging efficiency uncertainties.

- **Large-R jets.** The large-R jet uncertainties include the jet energy scale, jet energy resolution and jet mass scale uncertainties. Those uncertainties arise from the large-R jet calibration chain described in Section 3.6.3. The JES and JMS uncertainties are treated as correlated in the $VH, H \to b\bar{b}$ boosted analysis.
- **Triggers.** Efficiency (statistical and systematic) uncertainties for single lepton and $E_{\mathrm{T}}^{\mathrm{miss}}$ triggers need to be accounted for. Scale factors are applied to simulation to correct for data-simulation triggering efficiency discrepancies. Sources of uncertainties corresponding to those scale factors are thus considered. For the $E_{\mathrm{T}}^{\mathrm{miss}}$ trigger uncertainties, variations are directly derived by the $VH(b\bar{b}/c\bar{c})$ group with $W(\mu,\nu)$+jets events. The statistical error to derive those SFs is taken into account, and in addition uncertainties accounting for differences in SFs observed for $t\bar{t}$, $Z+\mathrm{jets}$ processes are also included. Finally, the kinematic dependencies of the sum of transverse momenta used to compute the $E_{\mathrm{T}}^{\mathrm{miss}}$ are considered for the $E_{\mathrm{T}}^{\mathrm{miss}}$ trigger uncertainty.
- **Leptons ($l = e, \mu$).** Electron and muon calibration (energy scale and energy resolution) as well as their identification, reconstruction and isolation efficiency uncertainties are taken into account. For muons the track-to-vertex association, the muon momentum (inner detector and muon spectrometer components) resolution and the identification of muons with low transverse momenta ($p_{\mathrm{T}} < 15\,\mathrm{GeV}$) are also included as sources of uncertainty.
- **Taus**. Systematic uncertainties associated to the reconstruction, the tau-identification efficiency, the tau identification against electron and the tau energy scale are taken into account.
- $\boldsymbol{E_{\mathrm{T}}^{\mathrm{miss}}}$. Uncertainties on the energy calibration and resolution of the soft transverse missing momentum ($p_{\mathrm{T}}^{\mathrm{miss,st}}$) are considered. Moreover, the uncertainties on the lepton and jet energy scales are propagated for the $E_{\mathrm{T}}^{\mathrm{miss}}$ uncertainty.
- **Luminosity**. The luminosity of LHC is determined thanks to van der Meer scans[4]. The uncertainty on the integrated luminosity associated to that measurement for the full ATLAS Run 2 is 1.7% [278].
- **Pile-up reweighting**. In simulations, only the best-guess of the data pile-up conditions is considered. Once the data has been collected, the pile-up conditions are corrected in the simulation to match the conditions observed in data: this procedure is referred to as *pile-up reweighting*. As a result, the distribution of the average number of interactions per bunch crossing in the simulated samples is corrected. A systematic uncertainty to account for that rescaling procedure is considered.

The experimental uncertainties related to the online event selection (*i.e.* affecting the recording of events), the luminosity, the leptons reconstruction and identification, and missing transverse momentum are considered as correlated between the resolved and boosted analyses in the likelihood fit. The uncertainties related to the small-R jets energy calibration, are also treated as correlated between the two regimes since small-R jets are used as Higgs boson candidate in the resolved regime, and are defining the HP SR and LP SR in the boosted regime. On the other hand, the uncertainties of small-R and large-R jets are treated as uncorrelated due to the differences in reconstruction and calibration procedures. In the same way, the uncertainties related to the b-tagging efficiency correction factors for small-R and large-R jets are treated as uncorrelated because the identification of b-jets relies on inputs from jet reconstruction which differ between the two regimes as the tagging relies on calorimeter-based information for the resolved topology and track-based information for the boosted topology.

4. The van der Meer scan allows to determine the luminosity of the LHC. Beams of protons are transversely separated with an increasing distance. A scan of the number of interactions as function of the transverse distance between the beams allows for a determination of the shape of proton bunches at the interaction point. This information once fitted with dedicated models allows deriving the luminosity of ATLAS for nominal conditions when beams are perfectly aligned.

5.9 Results from the previous publications

In this section the latest ATLAS $VH(b\bar{b}/c\bar{c})$ public results are presented. When relevant a comparison with the latest results of CMS is also provided and the main differences between the two analyses are briefly emphasized. The theoretical notions and frameworks used *i.e.* STXS cross-section measurement, coupling strength modifiers definition and EFT interpretations have been introduced in Chapter 1 respectively in Sections 1.4.4, 1.4.6 and 1.4.5.

Below is a quick summary of the Run 2 related ATLAS results concerning the $VH, H \to b\bar{b}/c\bar{c}$ analyses.

- In 2018 was announced the observation of the $H \to b\bar{b}$ decay mode and of the VH production mode [3] which yielded observed (expected) significances respectively of 5.4 (5.5) and 5.3 (4.8) standard deviations for those two processes.
  - The $H \to b\bar{b}$ decay observation was made possible combining a partial dataset from Run 2 ($80\ \text{fb}^{-1}$ of data) with the full dataset collected for the Run 1 ($25\ \text{fb}^{-1}$) of the LHC. With this amount of data, the $H \to b\bar{b}$ observation has required a combination of the VH, $t\bar{t}H$, VBF and ggF Higgs production modes selecting events with a signature compatible with a Higgs decay to a pair of b-quarks. The $VH, H \to b\bar{b}$ alone measurement was found to reach already a 4.9 (5.1) observed (expected) significance (combining Run 1 and partial Run 2 results) but was not enough to claim the observation of the $H \to b\bar{b}$ decay mode.
  - On the other hand the observation of the VH production mode was obtained only using the partial Run 2 dataset ($80\ \text{fb}^{-1}$ of data) selecting events compatible with a VH signature and combining $H \to b\bar{b}$, $H \to ZZ^* \to 4l$ and $H \to \gamma\gamma$ decay mode measurements. The $VH, H \to b\bar{b}$ alone measurement was found to reach already a 4.9 (4.3) observed (expected) significance (for the partial Run 2 dataset) but was not enough to claim the observation of the VH production mode.
- In 2019, the first resolved $VH, H \to b\bar{b}$ STXS measurement [328] was performed in 5 bins with the partial Run 2 dataset ($80\ \text{fb}^{-1}$).
- In 2020 the resolved 5-STXS bin measurement was updated [65] exploiting the full Run 2 dataset ($139\ \text{fb}^{-1}$ of data). It led to the observation of the VH production mode based only on $VH, H \to b\bar{b}$ measurement. The observed (expected) significances of 4.0 (4.1) and 5.3 (5.1) were respectively found for the WH and ZH production modes.
- The same year, the first boosted $VH, H \to b\bar{b}$ STXS measurement [275] with the full Run 2 was performed with 4 bins and an observed (expected) significance of 2.1 (2.7) was found.
- In 2021, a preliminary resolved and boosted combination $VH, H \to b\bar{b}$ STXS measurement (7 bins) with the full Run 2 [5] was carried out: an observed (expected) significances of 4.1 (3.9) and 4.6 (5.0) standard deviations for the WH and ZH production modes were obtained respectively.
- At the end of 2021, a combination of the $VH, H \to c\bar{c}$ and $VH, H \to b\bar{b}$ resolved analyses has been performed [6]. The $VH, H \to c\bar{c}$ search yields an observed (expected) upper limit of 26 (31) times the predicted SM $\sigma(VH) \times \mathcal{B}(H \to c\bar{c})$ product. For the first time, the Higgs-charm coupling was measured to be weaker than the Higgs-bottom coupling with a 95% CL.
- In 2022, the first fiducial and differential cross-section measurement [276] was performed in the 0-lepton channel using the full Run 2 data-taking period.

Only the most general results that provide the largest overview of the analysis are detailed in this section. Thus, despite being an interesting result, the new fiducial measurement will not be covered in this thesis. In Section 5.9.1, the resolved and boosted combination $VH, H \to b\bar{b}$ measurement is detailed, the Section 5.9.2 presents the $VH, H \to c\bar{c}$ measurement and the combination of the resolved $VH, H \to b\bar{b}$ and $VH, H \to c\bar{c}$ analyses. Finally, the Section 5.9.3 explains two possible

cross-checks that can be performed to validate the event selection of the VH analysis, and check that no potential undetected biases have been introduced in the VH measurement.

5.9.1 $VH, H \to b\bar{b}$ resolved and boosted combination

The general concept of an STXS measurement is introduced in Section 1.4.4. Here, the STXS measurement is performed combining the $VH, H \to b\bar{b}$ resolved and boosted topologies [5] and follows closely the event selection and the signal and control regions division of phase space described in Section 5.3. The background modelling and sources of uncertainties taken into account are pretty similar to what is presented in Section 5.8. The fits are performed following the fit formalism described in Section 5.7.

The main differences for that publication compared to the current analysis are listed below.

- The small-R jets are reconstructed with the EMTopo algorithm whereas currently the PFlow algorithm is used.
- For the resolved topology, the MV2c10 b-tagger was used with the 70% WP requirement. For such regime dedicated BDT algorithms have been trained per leptonic channels.
- For the boosted topology, the MV2c10 b-tagger was also used with the 70% WP leading 2 b-tagging strategy. The significances are obtained with a fit of the invariant mass of the leading large-R jet of the event as the BDT approach was not yet used at that time for that regime.

5.9.1.1 Combination strategy

An orthogonal division of phase space is required to perform the STXS measurement combining the resolved and boosted topologies. The Figure 5.22 shows the fraction of events that can be reconstructed either with only the resolved or with only the boosted or with both topologies (referred to as overlap events in that case) as a function of the reconstructed p_{T}^{V}. The fraction of signal events that are only captured by the boosted analysis increases with p_{T}^{V}. As for the current analysis, it was decided to reconstruct events only with the resolved topology if $p_{\mathrm{T}}^{V} < 400$ GeV and only with the boosted topology if $p_{\mathrm{T}}^{V} > 400$ GeV which corresponds to the kinematic properties expected for the Higgs decay to b-quark in VH events and the ability to capture both of them in a single large-R jet of radius $R = 1.0$ instead or two small-R jets of radius $R = 0.4$ as for $p_{\mathrm{T}}^{V} = 400$ GeV the angular distance is expected to be $\Delta R(b, \bar{b}) \approx 0.6$. Moreover, that bound for the reconstruction of those two regimes is further motivated by the boundaries chosen for the STXS categories in ATLAS for which a bound at 400 GeV is imposed in order to be able to combine STXS measurements of several Higgs analyses.

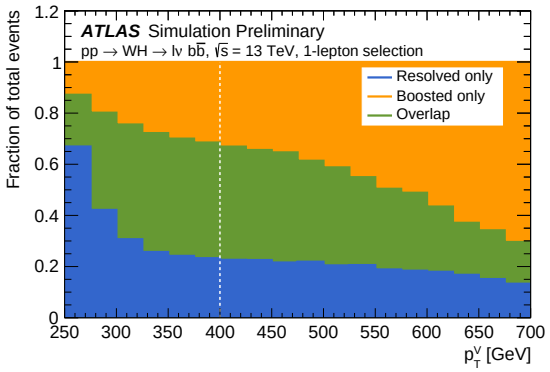


Figure 5.22: Fraction of selected $VH, H \to b\bar{b}$ signal events that can be reconstructed only with the resolved topology (blue), only with the boosted topology (orange) or with the two topologies (referred to as overlap events, green) as predicted by the nominal simulation in the 1-lepton channel as a function of the reconstructed p_{T}^{V}. Only events with $p_{\mathrm{T}}^{V} > 250$ GeV are shown since the boosted selection is not considered for lower p_{T}^{V} (taken from Ref. [5]).

5.9.1.2 STXS division of phase space

The STXS measurement is using a total of 7 p_{T}^{V}-bins listed in Figure 5.23 and follows a reduced stage-1.2 STXS scheme as here the measurement is performed inclusively in terms of jet multiplicity *i.e.* all jet categories are merged both at reconstruction and the truth levels compared to the original VH stage-1.2 STXS scheme (presented in Figure 1.9a). The measurement is performed separating the WH (1-lepton channel) and ZH (0- and 2-lepton channels) processes at particle level.

- For the resolved topology two p_{T}^{V}-bins are defined: the $[150, 250]$ GeV and $[250, 400]$ GeV p_{T}^{V}-regions. For the 2-lepton channel the event reconstruction starts at $p_{\mathrm{T}}^{V} = 75$ GeV hence there is the additional $[75, 150]$ GeV p_{T}^{V}-region since that channel is not polluted by the multijet background compared to the 0- and 1-lepton channels.
- For the boosted topology a single p_{T}^{V}-region is defined: $p_{\mathrm{T}}^{V} > 400$ GeV

The truth level categories are the same as the reconstructed ones but as a function of the truth p_{T}^{V} which is denoted $p_{\mathrm{T}}^{V,t}$.

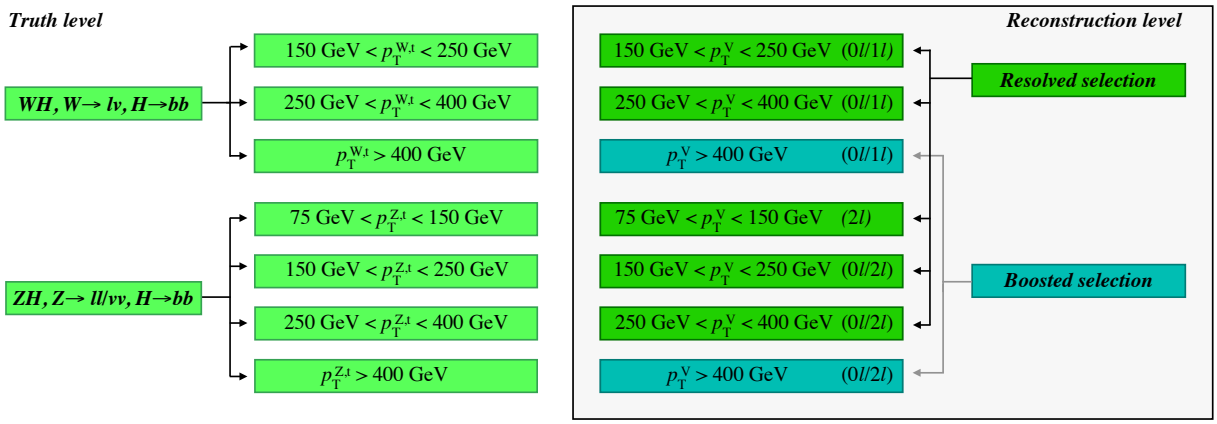
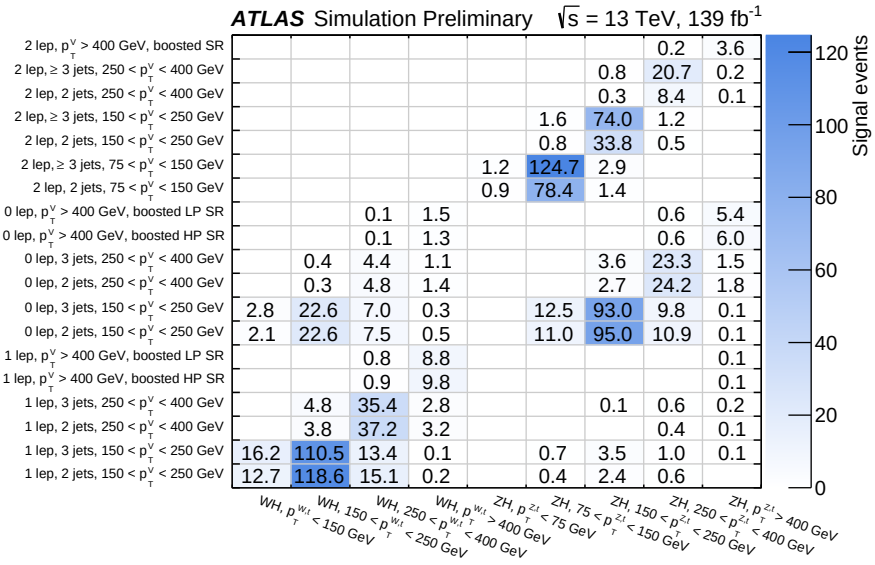
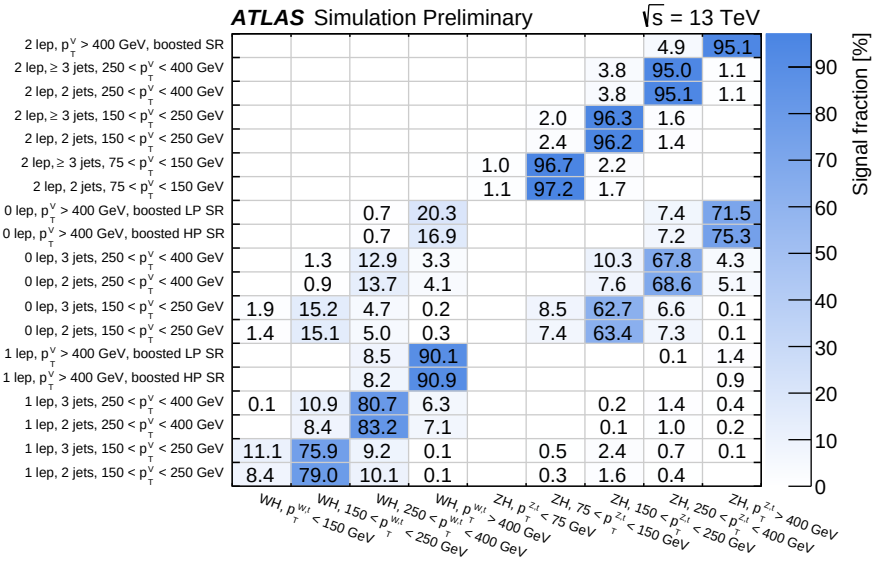


Figure 5.23: Reduced stage-1.2 STXS scheme: the particle level ZH and WH processes are divided into truth level transverse momentum ($p_{\mathrm{T}}^{V,t}$) categories for which cross-sections are extracted in the combination. The corresponding p_{T}^{V} categories at reconstruction level follow the same division of phase space as the truth level (taken from Ref. [5]).

The expected signal yields in each reconstructed categories as well as the corresponding fraction of signal events are reported in Figure 5.24. The block-diagonal matrices obtained prove there is a good correspondence between the reconstructed p_{T}^{V} and truth $p_{\mathrm{T}}^{V,t}$ categories. For the 0-lepton channel, the WH signal event as explained in Section 5.3.2 are indeed representing 13%-20% of reconstructed events in particular because of $WH \to \tau\nu b\bar{b}$ events.



(a)



(b)

Figure 5.24: Predicted (a) signal event yields and (b) fraction (expressed in percent) of signal events passing all selection criteria in every reconstructed event category (y-axis) for each each STXS signal region (x-axis) and every $VH, H \to b\bar{b}$ processes. Entries with event yield below 0.1 or signal fractions below 0.1% are not shown (taken from Ref. [5]).

5.9.1.3 Results of the STXS measurement

The results of the combined STXS measurement can be found in Figure 5.25 and are as well reported in Table 5.8. In that figure, a comparison with the previous resolved only [65] and boosted only [275] STXS measurements is provided. The combined STXS results are compatible with the Standard Model prediction within uncertainties and are compatible with the previous measure-

ments. The details of the different sources of uncertainties is presented in Table 5.9. The relative precision on the signal strengths measured ranges from 30% to 40% in the $150 < p_{\mathrm{T}}^{V,t} < 400$ GeV regions, to 75% in the lowest and up to 300% in the highest p_{T}^{V}-regions. The largest uncertainty comes from the ZH $p_{\mathrm{T}}^{V,t} > 400$ GeV region where an under-fluctuation in data is observed compared to the signal-plus-background SM prediction. The statistical uncertainty includes the floating normalization factors on the leading background contributions is the single largest uncertainty in all regions though for the $p_{\mathrm{T}}^{V,t} < 250\,\mathrm{GeV}$ STXS bins systematic uncertainties are also significantly contributing to the total uncertainty. The largest systematic uncertainties are related to the modelling of background, jet calibration and b-tagging with as well a significant contribution of the signal uncertainties for the ZH processes because of the limited precision on the gluon initiated processes theoretical predictions. The observed correlations between the measured cross-sections for the 7 STXS bins are shown in Figure 5.26 and are of the order of few percent for most of the phase space up to 10-15% for $p_{\mathrm{T}}^{V,t} > 400$ GeV.

An inclusive signal strength for the $VH, H \to b\bar{b}$ processes of

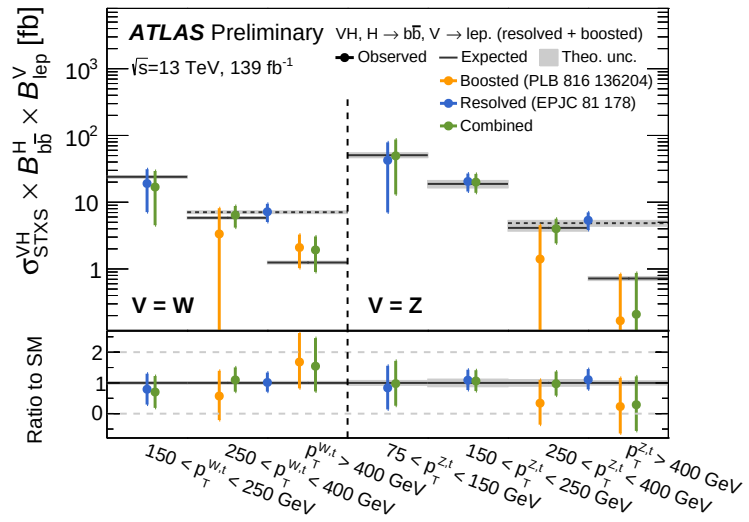
$$\mu_{VH}^{H\to b\bar{b}} = 1.00^{+0.18}_{-0.17} = 1.00^{+0.12}_{-0.11}\,(\text{stat.})^{+0.14}_{-0.13}\,(\text{syst.})$$

is found. Moreover, when measured separately, the $WH$and ZH signal strengths are found to be equal to:

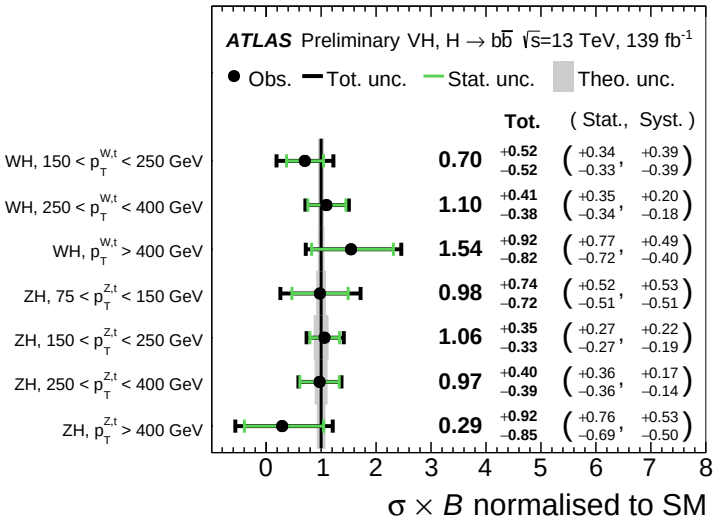
$$\mu_{WH}^{H\to b\bar{b}} = 1.03^{+0.28}_{-0.27} = 1.03^{+0.19}_{-0.19}\,(\text{stat.})^{+0.21}_{-0.19}\,(\text{syst.}),$$
$$\mu_{ZH}^{H\to b\bar{b}} = 0.97^{+0.25}_{-0.23} = 0.97^{+0.17}_{-0.17}\,(\text{stat.})^{+0.18}_{-0.15}\,(\text{syst.}),$$

and the background-only hypothesis is rejected with observed (expected) significances of 4.1 (3.9) and 4.6 (5.0) standard deviations respectively for the WH and ZH production modes. A few post-fit distributions used for the fit of the BDT outputs for the resolved and boosted topology are shown in Figure 5.27 and 5.28. The 0-lepton and 2-lepton channels are dominated by the $Z + \text{jets}$ background while the 1-lepton is mainly contaminated by the $t\bar{t}$(semi-leptonic) and $W + \text{jets}$ backgrounds. In the 2-lepton channel despite the lack of statistic a high purity is achieved especially in the resolved regime thanks to the BDT as shown in Figure 5.27c.

This $VH, H \to b\bar{b}$ STXS measurement is cross-validated thanks to the VZ (WZ and ZZ) analysis, whose principle is detailed in Section 5.9.3, in the resolved regions where an alternative BDT discriminant (BDT_{VZ}) is trained and simultaneously in the boosted region fitting the mass of the leading large-R jet of the event. The VH events are considered as background in that analysis with a NP on its normalization constrained to the SM prediction. The signal strength measured for those processes is $\mu_{VZ}^{Z\to b\bar{b}} = 0.96^{+0.15}_{-0.13}$ which is in good agreement with the SM predictions and compatible with the result already obtained for the published resolved only analysis.



(a)



(b)

Figure 5.25: Expected (lines) versus observed (markers) VH ($V \to$ lepton) cross-sections times the $H \to b\bar{b}$ branching fraction in the reduced 1.2 STXS scheme (a) for the combination (green) compared to the resolved only [65] (blue) and boosted only [275] (orange) analyses results. The horizontal dashed lines indicate the expected values in the $p_{\mathrm{T}}^{V,t} > 250$ GeV category since the published resolved analysis did not include a split at a $p_{\mathrm{T}}^{V,t}$ of 400 GeV. (b) Signal strength measurement for the 7 STXS bins. The grey shaded area shows the theoretical uncertainty on the SM prediction (taken from Ref. [5]).

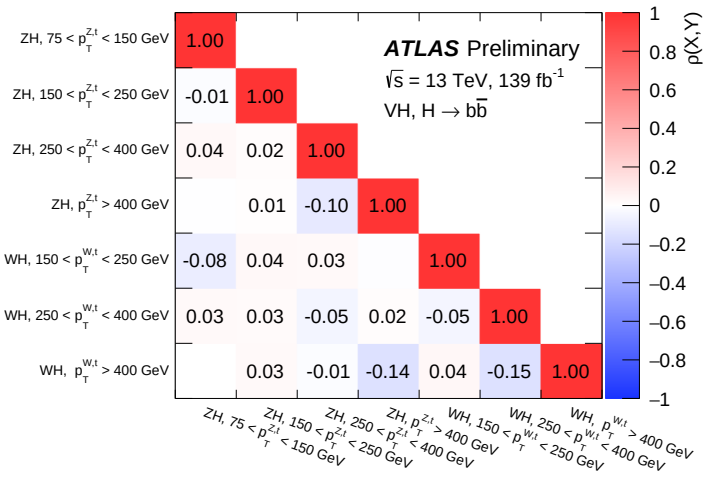


Figure 5.26: Observed correlations between the measured reduced stage-1.2 STXS for $VH(V \to \text{leptons}, H \to b\bar{b}$ processes, including both the statistical and systematic uncertainties (taken from Ref. [5]).

STXS region		SM prediction			Measurement			Stat. unc.	Syst. unc. [fb]		
Process	$p_{\mathrm{T}}^{V,\,\mathrm{t}}$ interval	[fb]			[fb]			[fb]	Th. sig.	Th. bkg.	Exp.
$W(\ell\nu)H$	150–250 GeV	24.0	±	1.1	16.9	±	12.4	8.1	0.8	7.3	5.7
$W(\ell\nu)H$	250–400 GeV	5.8	±	0.3	6.4	±	2.3	2.0	0.3	0.9	0.6
$W(\ell\nu)H$	$>$ 400 GeV	1.3	±	0.1	1.9	±	1.1	0.9	0.2	0.4	0.3
$Z(\ell\ell/\nu\nu)H$	75–150 GeV	50.6	±	4.1	49.5	±	36.9	25.9	6.3	18.5	20.6
$Z(\ell\ell/\nu\nu)H$	150–250 GeV	18.8	±	2.4	20.0	±	6.4	5.1	1.8	2.5	2.2
$Z(\ell\ell/\nu\nu)H$	250–400 GeV	4.1	±	0.5	4.0	±	1.6	1.5	0.4	0.4	0.3
$Z(\ell\ell/\nu\nu)H$	$>$ 400 GeV	0.7	±	0.1	0.2	±	0.6	0.5	0.1	0.3	0.2

Table 5.8: Best-fit values and uncertainties for the VH ($V \to$ leptons) cross-section times the $H \to b\bar{b}$ branching fraction, in the reduced 1.2 STXS scheme. The SM predictions for each region, computed using the inclusive cross-section calculations and the event simulations are also reported. The symmetrized contributions to the total measurement uncertainty from statistical (Stat. unc.) and systematic uncertainties (Syst. unc.) related to the signal (Th. sig.) and background prediction (Th. bkg.), and the experimental performance (Exp.) are provided separately. The total systematic uncertainty (difference in quadrature between the total uncertainty and the statistical uncertainty) differs from the sum in quadrature of the Th. sig., Th. bkg., and Exp. systematic uncertainties due to correlations. All leptonic decays of the vector bosons (including those to τ-leptons, $l = e, \mu, \tau$) are considered (taken from Ref. [5]).

Category	WH, 150–250 GeV	WH, 250–400 GeV	WH, >400 GeV	ZH, 75–150 GeV	ZH, 150–250 GeV	ZH, 250–400 GeV	ZH, >400 GeV
Total	73%	36%	56%	75%	32%	40%	305%
Statistical	48%	32%	48%	52%	26%	37%	248%
Systematic	56%	17%	29%	53%	19%	16%	177%
Statistical uncertainties							
Data stat only	44%	29%	45%	44%	23%	34%	242%
Floating normalisations	26%	9%	14%	21%	11%	10%	57%
$t\bar{t}$ $e\mu$ control region	4%	≤1%	≤1%	23%	4%	8%	3%
Experimental uncertainties							
Total experimental	34%	9%	14%	42%	11%	7%	95%
↪ lepton	4%	2%	3%	3%	≤1%	≤1%	4%
↪ $E_{\mathrm{T}}^{\mathrm{miss}}$	24%	≤1%	≤1%	25%	3%	≤1%	9%
↪ Small-R jets	17%	6%	6%	28%	7%	4%	19%
↪ Large-R jets	≤1%	≤1%	9%	≤1%	≤1%	≤1%	89%
↪ Calo b-tagging (b-jets)	8%	2%	≤1%	18%	8%	5%	6%
↪ Calo b-tagging (c-jets)	14%	6%	≤1%	2%	2%	≤1%	3%
↪ Calo b-tagging (light-flavour jets)	≤1%	≤1%	≤1%	≤1%	≤1%	2%	2%
↪ Calo b-tagging (extrap. from charm)	≤1%	≤1%	≤1%	≤1%	≤1%	≤1%	≤1%
↪ VR b-tagging (b-jets)	≤1%	≤1%	7%	≤1%	≤1%	≤1%	15%
↪ VR b-tagging (c-jets)	≤1%	≤1%	2%	≤1%	≤1%	≤1%	8%
↪ VR b-tagging (light-flavour jets)	≤1%	≤1%	≤1%	≤1%	≤1%	≤1%	21%
↪ VR b-tagging (extrap. from charm)	≤1%	≤1%	≤1%	≤1%	≤1%	≤1%	3%
↪ Pile-up	2%	≤1%	≤1%	5%	≤1%	≤1%	≤1%
↪ Luminosity	2%	2%	2%	≤1%	≤1%	≤1%	5%
Theoretical and modelling uncertainties							
Signal	5%	5%	9%	13%	9%	9%	23%
Backgrounds	43%	13%	23%	37%	12%	11%	145%
↪ single top	17%	8%	8%	5%	≤1%	2%	4%
↪ $t\bar{t}$	22%	3%	9%	10%	4%	3%	13%
↪ W+jets	25%	6%	8%	2%	≤1%	≤1%	14%
↪ Z+jets	7%	2%	5%	29%	9%	6%	115%
↪ Diboson	6%	3%	3%	9%	4%	4%	27%
↪ Multi-jet	≤1%	≤1%	3%	≤1%	≤1%	≤1%	3%
↪ MC statistical	17%	8%	13%	17%	6%	7%	77%

Table 5.9: Breakdown of the relative uncertainty in the cross-section for the WH and ZH STXS measurements in the reduced STXS scheme (taken from Ref. [5]).

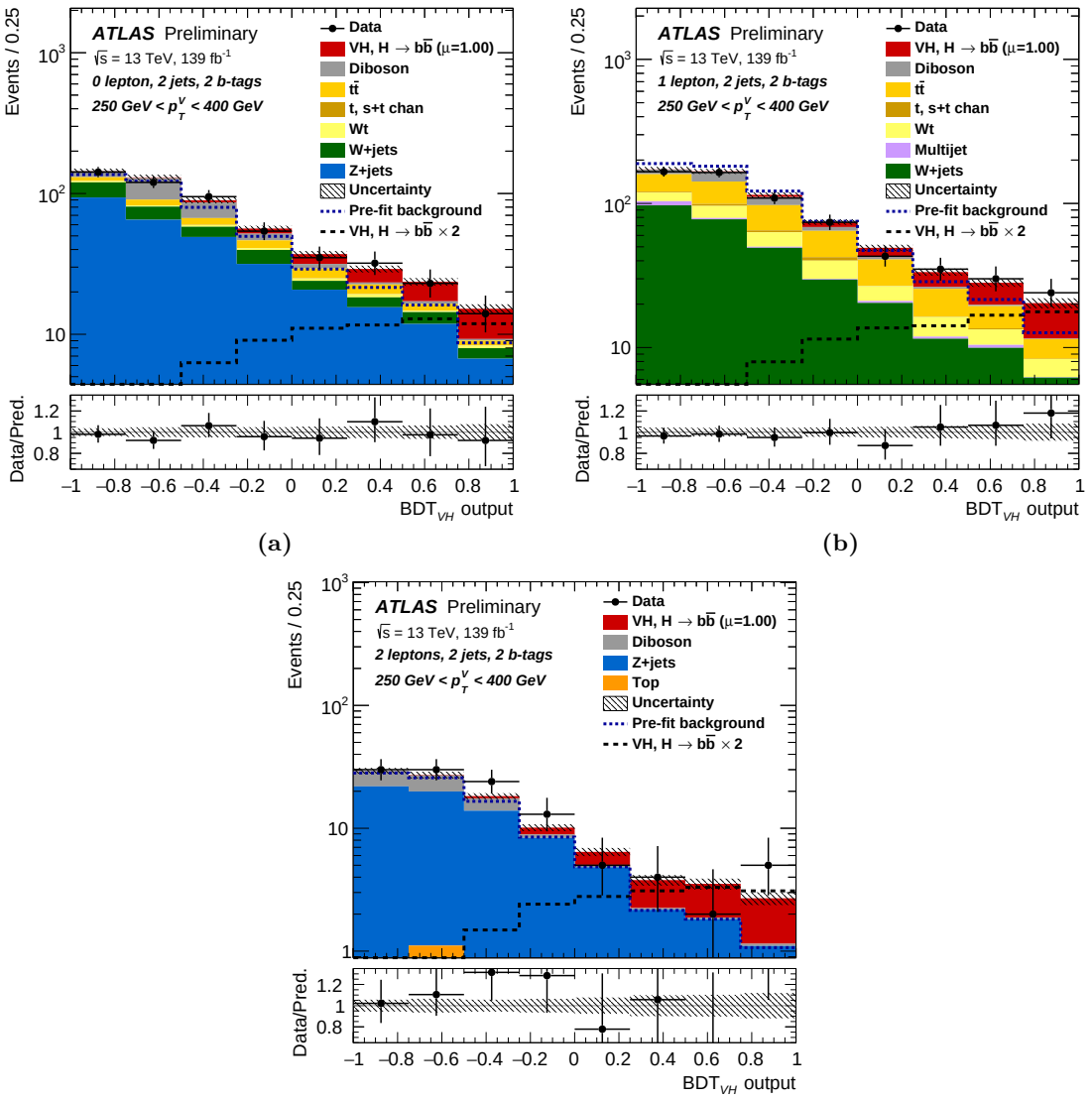


Figure 5.27: (a) 0-lepton, (b) 1-lepton and (c) 2-lepton post-fit distributions of the BDT discriminant in the resolved regime 2 b-tagged 2-jets signal region for $250 < p_{\mathrm{T}}^{V} < 400\,\mathrm{GeV}$. The background processes are normalized thanks to the global likelihood fit, the total pre-fit background distribution is shown as a dashed blue line. The Higgs boson signal is shown as a filled histogram on top of the fitted backgrounds normalized to the signal yield extracted from data ($\mu = 1.00$), and unstacked as an unfilled histogram (dashed black line), scaled by the multiplicative factor indicated in the legend. The size of the combined statistical and systematic uncertainties for the sum of the fitted signal and background are indicated by the hatched band. The ratios of the data to the sum of the fitted signal and background are shown in the lower panels (taken from Ref. [5]).

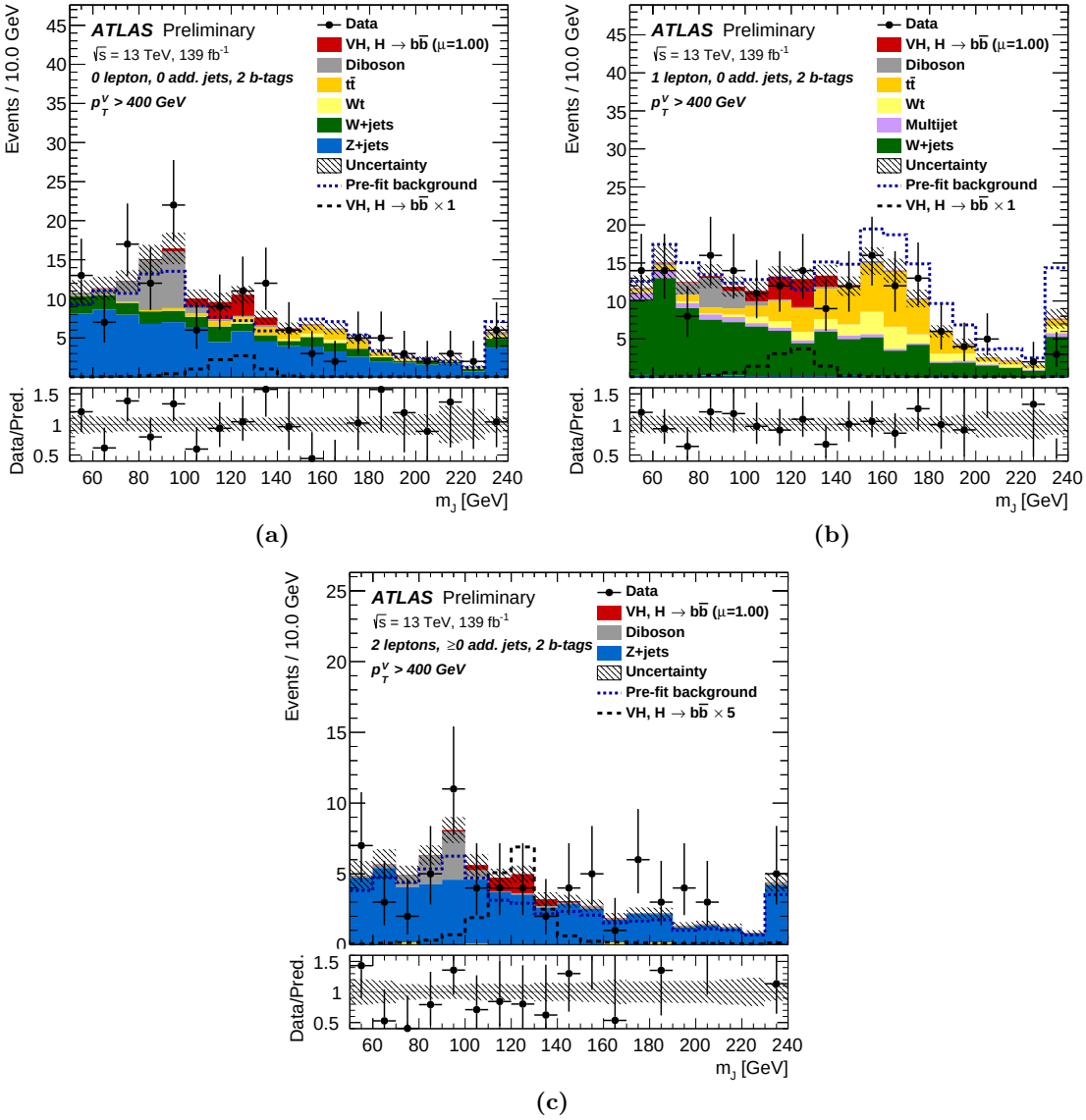


Figure 5.28: (a) 0-lepton, (b) 1-lepton and (c) 2-lepton post-fit distributions of the invariant mass of the leading large-R jet in the boosted regime high for $p_{\mathrm{T}}^{V} > 400\,\mathrm{GeV}$ in the (a) and (b) high purity signal regions, and (c) in the signal region as there is indeed no HP SR, LP SR and CR in the 2-lepton channel: all regions are merged together to circumvent the low statistic available. The contributions of each process are plotted after the global likelihood fit. The Higgs boson signal is shown as a filled histogram on top of the fitted backgrounds normalized to the signal yield extracted from data ($\mu = 1.00$), and unstacked as an unfilled histogram (dashed black line), scaled by the multiplicative factor indicated in the legend. The dashed histograms show the total pre-fit background. The size of the combined statistical and systematic uncertainties for the sum of the fitted signal and background are indicated by the hatched band. The ratios of the data to the sum of the fitted signal and background are shown in the lower panels (taken from Ref. [5]).

5.9.1.4 EFT interpretation

The $VH, H \to b\bar{b}$ STXS measurement can be interpreted with an effective field theory approach whose general principle is detailed in Section 1.4.6. It allows constraining potential modifications of the signal strength and tensor structure of such $VH, H \to b\bar{b}$ processes. The effective Lagrangian ($\mathcal{L}_{\text{SMEFT}}$) is obtained from the SM Lagrangian ($\mathcal{L}_{\text{SM}}$) by adding extra terms following [60–62] to the Lagrangian only considering dimension-6 operators in the combination paper [5] (and thus in this section) as odd operators induce a violation of lepton and/or baryon number conservation and even operators with dimension $D \geq 8$ are suppressed by the energy scale. The assumed value for new physics energy scale is taken to be equal to $\Lambda = 1$ TeV.

Under those hypotheses, the Figure 5.29 shows the expected and observed constraints on the Wilson coefficient $c_{Hq}^{(3)}$ in the context of a one dimensional fit assuming that other coefficients are vanishing, and the Figure 5.30 presents the impact of three operators on the predicted $VH, H \to b\bar{b}$ signal strength as a function of $p_{\text{T}}^{V,t}$. Hence, non-negligible deviations with respect to the STXS SM predictions could be expected especially a high $p_{\text{T}}^{V,t}$ in case of EFT related phenomenon. The Figure 5.31 summarizes the confidence level obtained for one-dimensional fits for 3 different Wilson coefficients either using the linear or the linear-plus-quadratic parametrizations. Finally, in the Figure 5.32 the results of a 2-dimensional fit for the pair of Wilson coefficient $c_{Hq}^{(3)}$ and c_{HW} are provided.

In view of the EFT interpretations, the combination of the resolved and boosted topology appears to be of high interests because it allows lifting some degeneracy on some Wilson coefficients, and it reduces the ranges of their possible value. The results found are all compatible with the SM prediction showing no sign of deviation from the SM as the observed value of the Wilson coefficient are consistent with zero within uncertainties.

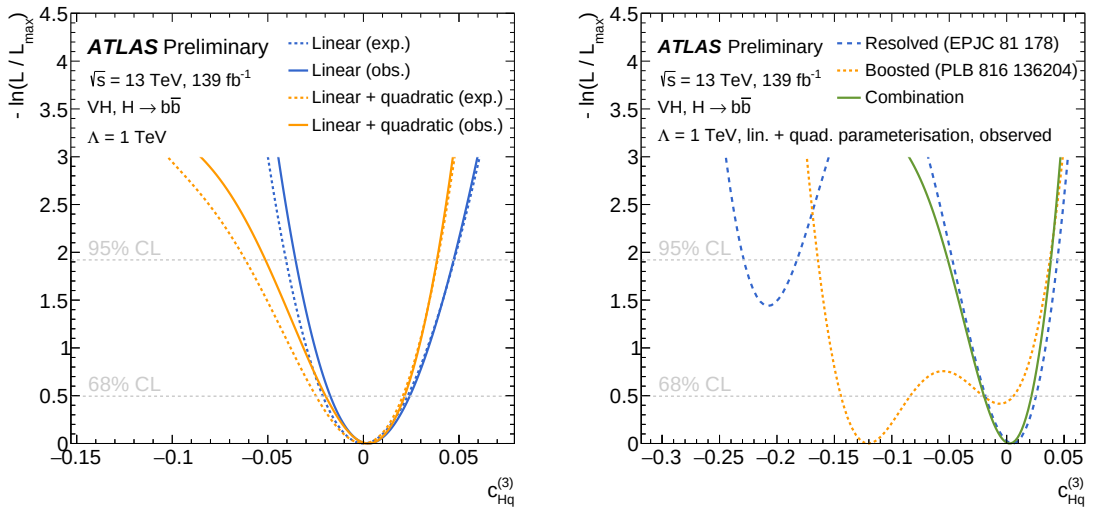
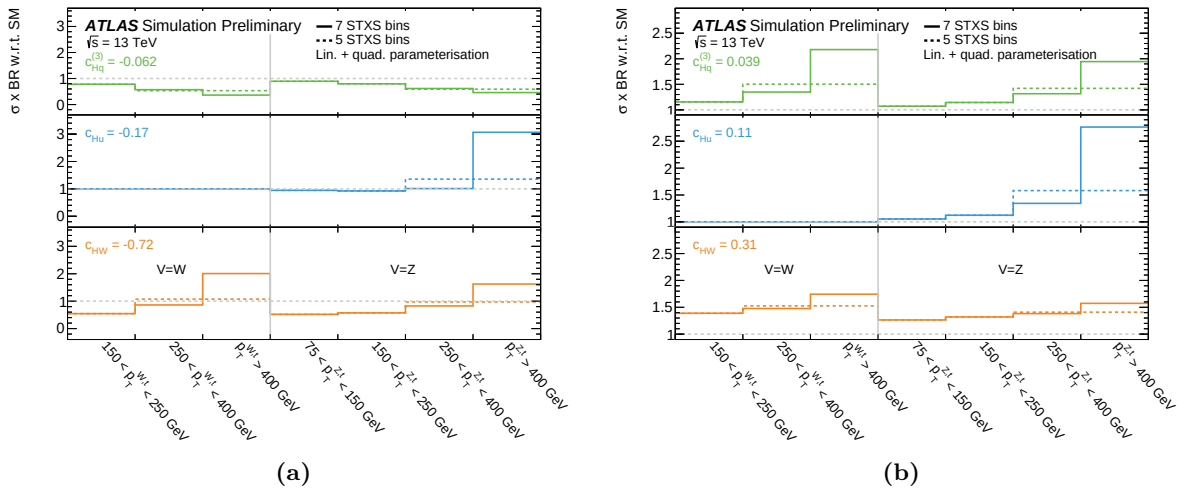
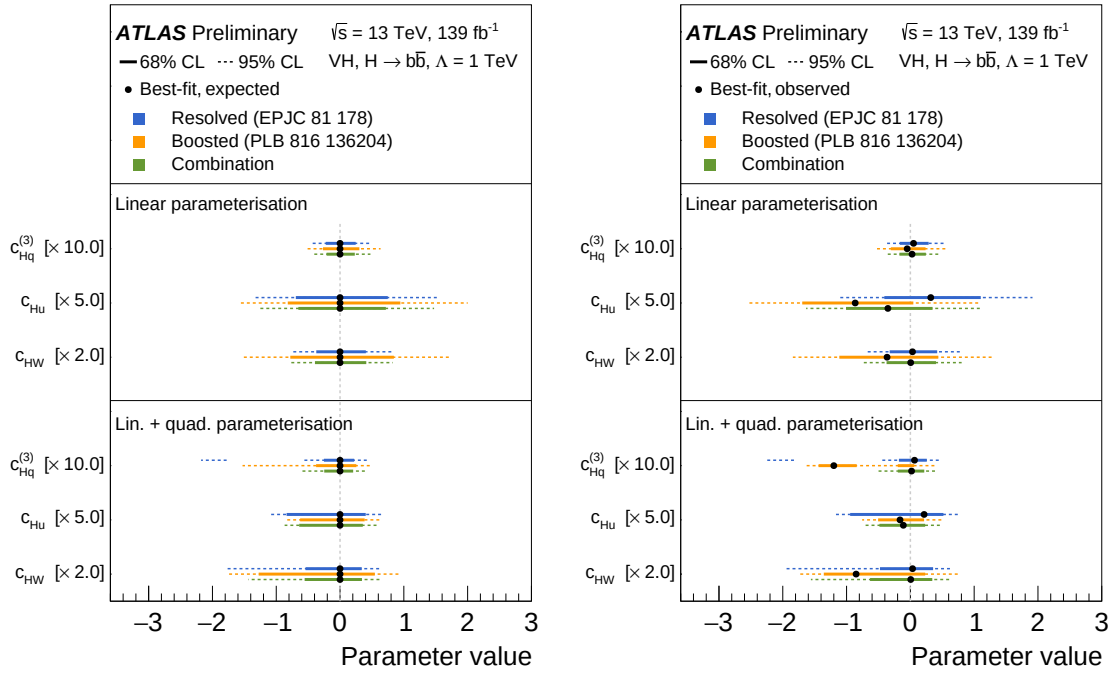


Figure 5.29: One-dimensional fits to constrain the Wilson coefficient $c_{Hq}^{(3)}$ of an effective Lagrangian when the other coefficients are assumed to vanish. (a) The observed (solid) and expected (dotted) negative-log-likelihood functions are shown for the case where only the linear (blue) or the linear-plus-quadratic (orange) terms are considered. (b) The observed negative-log-likelihood functions in the individual published (resolved only [65] and boosted only [275]) analyses and the combination for the one-dimensional fits taking into account linear-plus-quadratic terms (taken from Ref. [5]).



(a) **(b)**

Figure 5.30: Predicted modification of the $VH, H \to b\bar{b}$ signal strength as a function of $p_{\mathrm{T}}^{V,t}$ in the STXS regions considered for three different operators of the effective Lagrangian: $\mathcal{O}_{Hq}^{(3)}$, $\mathcal{O}_{Hu}$ and $\mathcal{O}_{HW}$. The values chosen for the Wilson coefficient correspond to the (a) lower and (b) upper end of the expected 95% confidence interval obtained by the combination analysis [5]. The cross-sections are calculated using the SMEFTSim package [62] and only considering linear plus quadratic terms for the operators. The solid line illustrates the case where each STXS bin receives an independent parameterization (7 STXS bins) while for the dotted line, a common parameterization is computed for the $p_{\mathrm{T}}^{V,t} > 250$ GeV regions for WH and ZH, respectively (5 STXS bins, taken from Ref. [5]).



(a) **(b)**

Figure 5.31: (a) Expected and (b) observed 68% and 95% CL intervals for three selected Wilson coefficients associated to dimension-6 operators of the effective Lagrangian when considering the linear and the linear-plus-quadratic parametrizations for the individual (resolved only [65] and boosted only [275]) published analyses and the combination in a one dimensional fit context *i.e.* assuming that only the coefficient considered can be non-zero (taken from Ref. [5]).

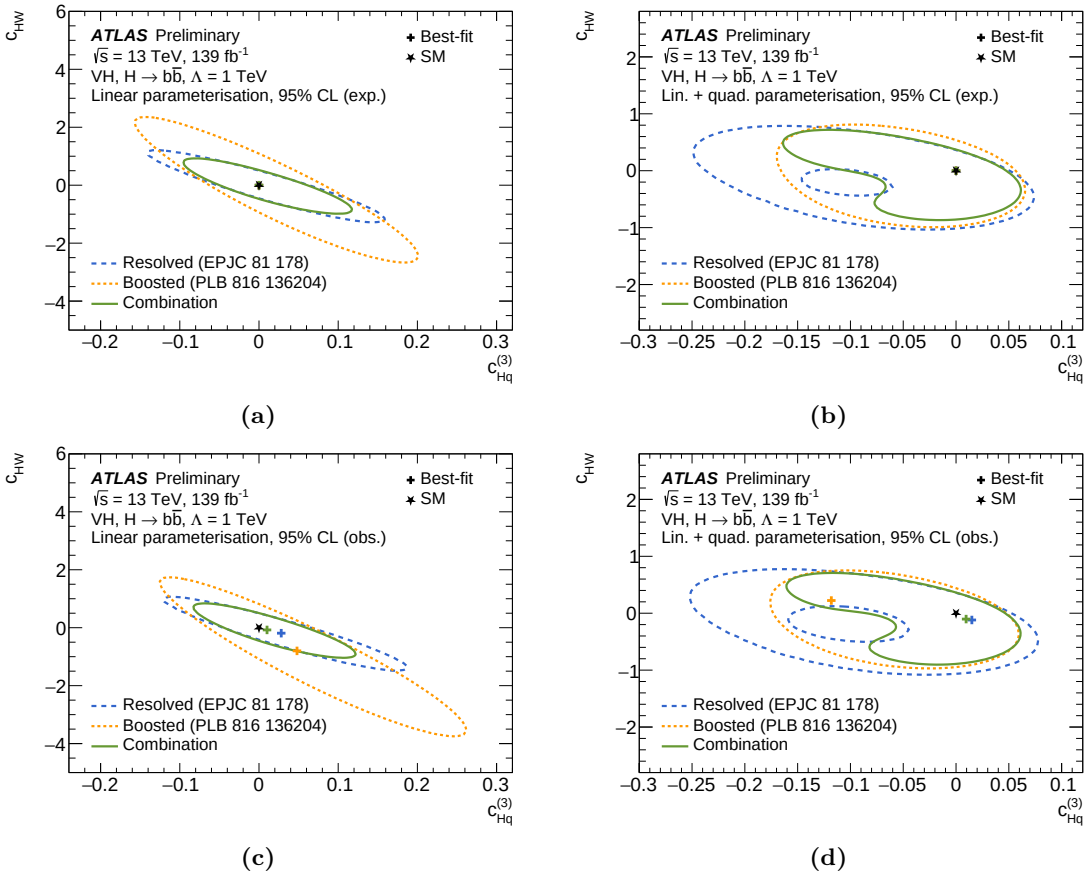


Figure 5.32: (a) and (b) Expected, and (c) and (d) observed 95% confidence intervals for a simultaneous likelihood fits for the pair $c_{Hq}^{(3)}$ and c_{HW} of Wilson coefficients of an effective Lagrangian in (a) and (c) the linear, and (b) and (d) the linear-plus-quadratic parametrizations for the individual (resolved only [65] and boosted only [275]) published analyses and the combination. The best-fit point is marked by a cross and the expected SM value by a star (taken from Ref. [5]).

5.9.1.5 Comparison with the CMS latest results

In this section the latest CMS results concerning the $VH, H \to b\bar{b}$ STXS measurement are discussed, only the main differences are highlighted but the details are left apart. The reader can refer to the references provided for more information.

In 2022, CMS released its first full Run 2 $VH, H \to b\bar{b}$ STXS measurement [329] measuring an inclusive signal strength of

$$\mu_{VH}^{H\to b\bar{b}} = 0.58^{+0.19}_{-0.18} = 0.58 \pm 0.15\,(\text{stat.}) \pm 0.12\,(\text{syst.}),$$

with an observed (expected) significance of $3.3\,\sigma$ ($5.2\,\sigma$). The separately measured STXS signal strengths for the WH and ZH processes are equal to

$$\mu_{WH}^{H\to b\bar{b}} = 0.97^{+0.32}_{-0.32} = 0.97 \pm 0.23\,(\text{stat.}) \pm 0.22\,(\text{syst.}),$$
$$\mu_{ZH}^{H\to b\bar{b}} = 0.37^{+0.21}_{-0.21} = 0.37 \pm 0.16\,(\text{stat.}) \pm 0.14\,(\text{syst.}).$$

The signal strength is smaller than that found by a previous CMS analysis [4] of a partial Run 2 dataset, corresponding to $77.2\,\text{fb}^{-1}$ of data taken between 2016 and 2017 which led to a measured signal strength of

$$\mu_{VH}^{H\to b\bar{b}} = 1.06 \pm 0.26 = 1.06 \pm 0.20\,(\text{stat.}) \pm 0.17\,(\text{syst.}),$$

which is in that case compatible with the SM predictions.

The STXS measurements obtained by CMS are presented in Figure 5.33. A reduced STXS stage 1.2 scheme is also used but slightly differs from the reduced STXS scheme of ATLAS as for CMS the p_{T}^{Z} category is split into the 0 and ≥ 1 additional jets in the $150 < p_{\text{T}}^{Z} < 250$ GeV region resulting in 8 STXS bins (compared to 7 bins for ATLAS). CMS uses as well a combination of the resolved and boosted topologies however the reconstruction strategy is different as events are only reconstructed as resolved for $p_{\text{T}}^{V} < 250$ GeV while for $p_{\text{T}}^{V} > 250$ GeV events can be reconstructed either as resolved or boosted prioritizing the resolved regime. In the Figure 5.33b deviation from SM are observed both for the WH and ZH processes for $p_{\text{T}}^{V,t} < 250$ GeV while deviation from the SM prediction for $p_{\text{T}}^{V,t} > 250$ GeV are only observed for the WH process. When performing the STXS measurement decoupling lepton channels, the main disagreement is coming from the 2-lepton channel as shown in Figure 5.33c which has the lowest uncertainty hence which must pull the global fit for the 8 STXS bin towards a lower signal strength value. Finally, CMS performed the cross-check analyses of the dijet invariant mass resulting in a measured signal strength $\mu_{VH}^{H\to b\bar{b}} = 0.34 \pm 0.34$ compatible with the global likelihood fit result and the VZ cross-check analysis measured $\mu_{VZ}^{Z\to b\bar{b}} = 1.16 \pm 0.13$ compatible with the SM prediction.

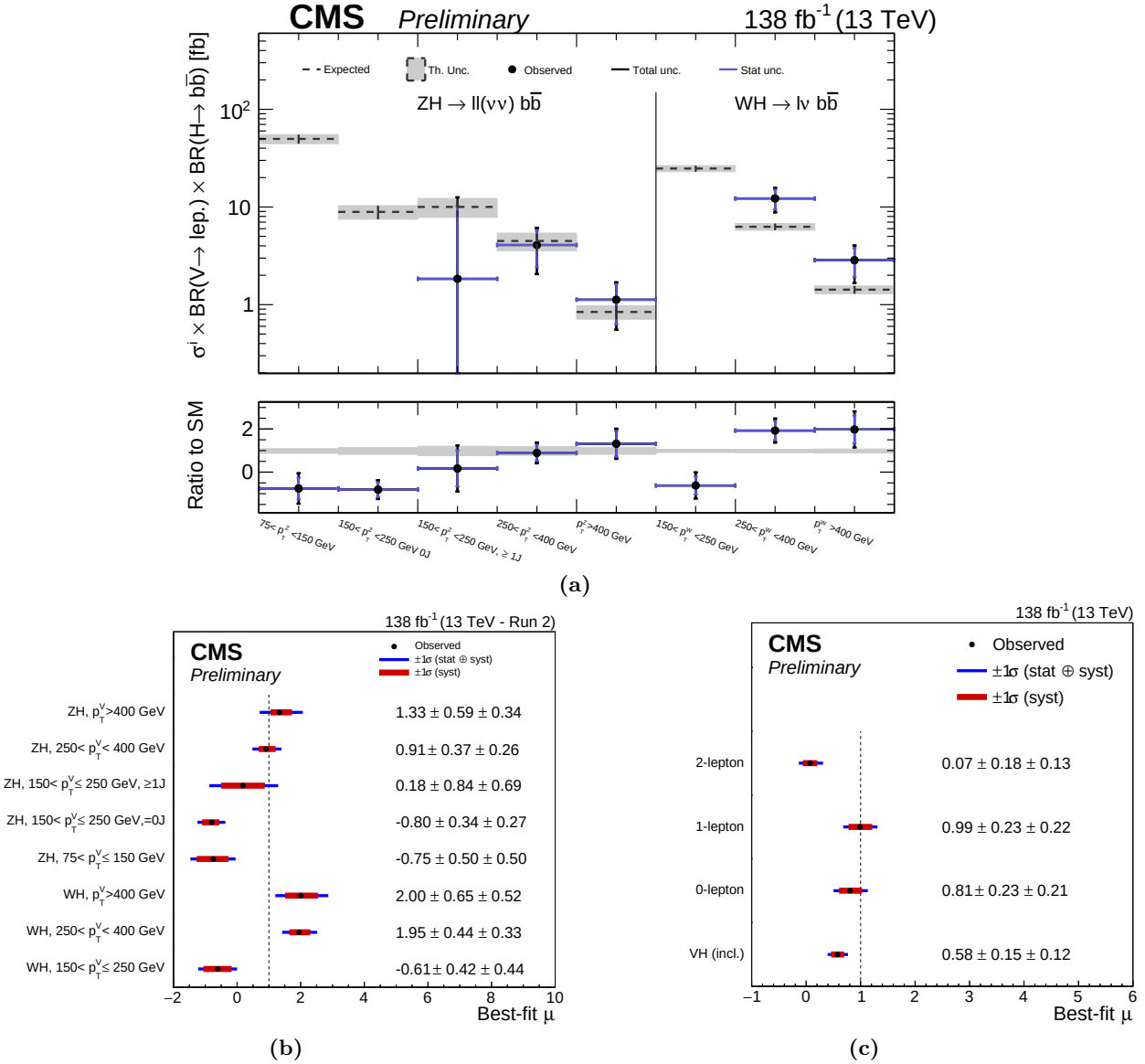


Figure 5.33: CMS full Run 2 (a) best-fit values and uncertainties for the VH ($V \to$ leptons) cross-section times the $H \to b\bar{b}$ branching fraction, in the reduced 1.2 STXS scheme. In the bottom panel, the ratio of the observed results with associated uncertainties to the SM expectations is shown. If the observed signal strength for a given STXS bin is negative, no value is quoted for $\sigma\mathcal{B}$.STXS signal strengths for (b) the 8 STXS bins of the analysis and (c) for the 0-, 1- and 2-lepton channels, as well as the combined signal strength (taken from Ref. [329]).

5.9.2 $VH, H \to b\bar{b}/c\bar{c}$ combination

The ATLAS collaboration published a combination of the resolved $VH, H \to b\bar{b}$ and (resolved) $VH, H \to c\bar{c}$ processes in 2022 [6]. To perform such combination an orthogonal b/c-tagging method was used with the MV2c10 tagger and DL1r tagger respectively for b- and c-tagging jets following the same principle as presented in Section 5.3.1. The 70% b-tagging efficiency WP is used for the MV2c10 tagger while the DL1r tagger has been optimized for the $VH, H \to c\bar{c}$ analysis leading to an average efficiency measured in simulated $t\bar{t}$ samples of 27% for c-jets, and misidentification rates of 8% and 1.6% respectively for b- and light-jets. The event categorization is significantly different from the current division of phase space of the $VH, H \to c\bar{c}$ analysis as the 1-c tag regions were used as signal region and fixed upper ΔR cuts were defined as a function of p_{T}^{V} instead of the now continuous ΔR-p_{T}^{V} side-bands. Moreover, the global likelihood fits were performed over the invariant m_{cc} mass distributions whereas for the $VH, H \to b\bar{b}/c\bar{c}$ Legacy publication the BDT distributions will be used instead. The reader is referred to the references provided in this section for further information about the details of the event selections and combination of events.

5.9.2.1 $VH, H \to c\bar{c}$ only likelihood fit

The global likelihood fit is performed with 3 signal strength corresponding to the $VH, H \to c\bar{c}$, $VW, W \to cq$ and $VZ, Z \to c\bar{c}$ processes:

$$\begin{aligned}
\mu_{VH}^{H\to c\bar{c}} &= -9 \pm 16 = -9 \pm 10\,(\text{stat.}) \pm 12\,(\text{syst.}) \\
\mu_{VW}^{W\to cq} &= 0.83 \pm 0.24 = 0.83 \pm 0.11\,(\text{stat.}) \pm 0.21\,(\text{syst.}) \\
\mu_{VZ}^{Z\to c\bar{c}} &= 1.16 \pm 0.48 = 1.16 \pm 0.32\,(\text{stat.}) \pm 0.36\,(\text{syst.}),
\end{aligned}$$

with the observed (expected) significance of the $VW, W \to cq$ process being equal 3.8 (4.6), and 2.6 (2.2) for the $VZ, Z \to c\bar{c}$ signal. For the $\mu_{VH}^{H\to c\bar{c}}$ signal strength an observed (expected) upper limit of 26 (31^{+12}_{-8}) at 95% confidence level is obtained with using a modified frequentist $\mathrm{CL_s}$ method [330] with the profile-likelihood ratio as the test statistic. The Figure 5.34 summarizes the results obtained per leptonic channels with as well as the ones derived combining leptonic channels. The best fit value $\mu_{VH}^{H\to c\bar{c}}$ is then interpreted with the kappa framework [53, 57] with the reparametrization of the likelihood function in terms of the Higgs-charm coupling modifier κ_c under the assumption that this coupling only affects the Higgs boson decays. Taking into account only SM decays, setting all other couplings to their SM predictions values and including effects in both the partial and full width then:

$$\mu_{VH}^{H\to c\bar{c}} = \frac{\kappa_c^2}{1 + \mathcal{B}_{H\to c\bar{c}}^{\mathrm{SM}}(\kappa_c^2 - 1)}, \tag{5.9.1}$$

with $\mathcal{B}_{H\to c\bar{c}}^{\mathrm{SM}}$ the branching ratio of the $H \to c\bar{c}$ decay as predicted by the SM. The combination of the channels for the profile-likelihood ratio test statistic allows an observed (expected) constraint for $|\kappa_c| < 8.5$ (12.4) to be set for the 95% CL as shown in Figure 5.34c. The Figure 5.35 presents the invariant mass m_{cc} of the 3 processes of interest: $VH, H \to c\bar{c}$, $VW, W \to cq$ and $VZ, Z \to c\bar{c}$. Finally, the Table 5.10 gathers the uncertainty breakdown for the fitted signal strengths.

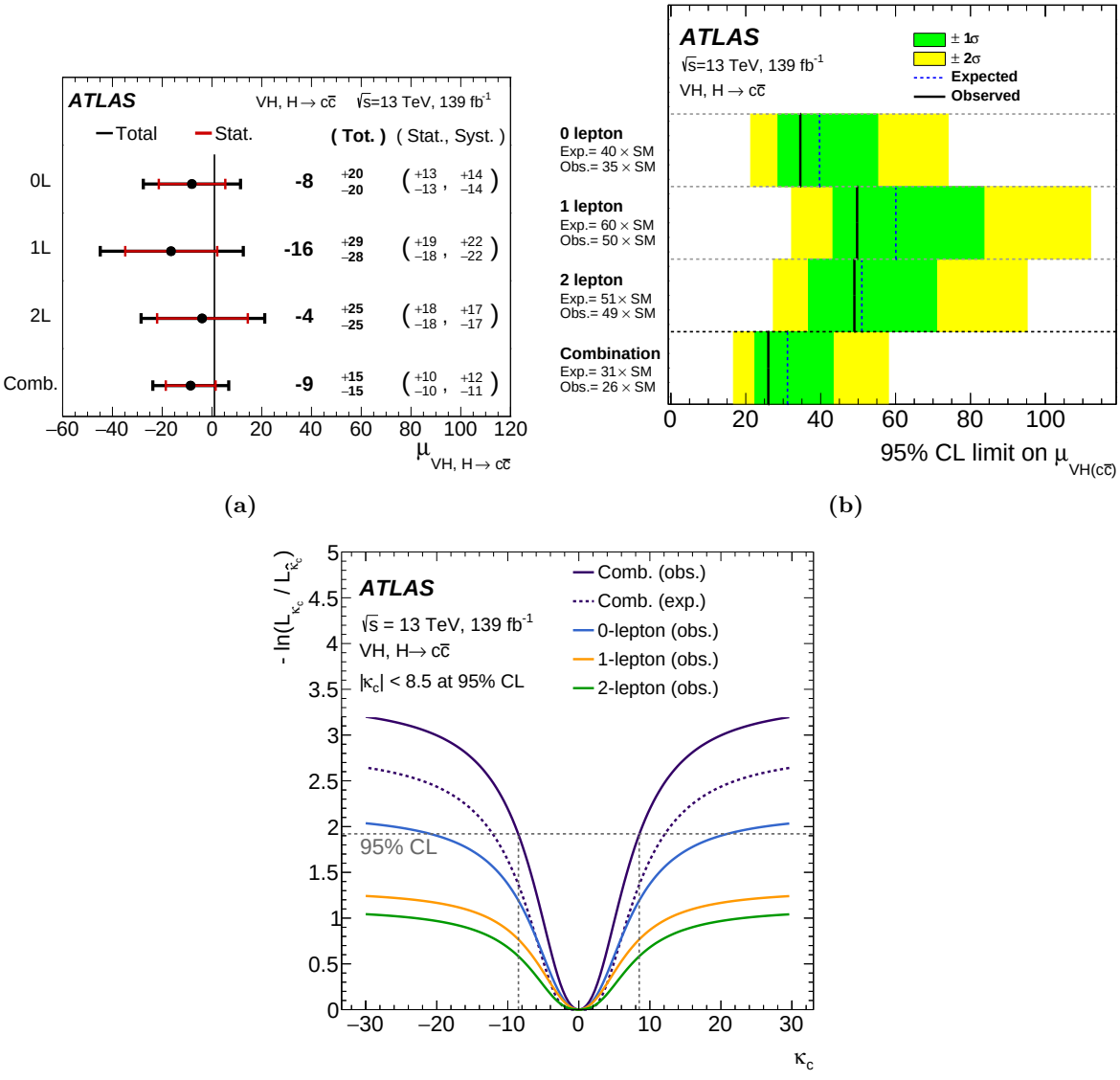


Figure 5.34: (a) Fitted signal strengths ($\mu_{VH}^{H\to c\bar{c}}$), (b) observed and expected 95% CL upper limits on $\mu_{VH}^{H\to c\bar{c}}$, and (c) expected and observed values of the negative profile log-likelihood ratio as a function of κ_c. For the single channel fits, a separate $\mu_{VH}^{H\to c\bar{c}}$ parameter for each channel is used in a five POI fit, while the combined fit (gathering the three leptonic channels) is performed with three POIs the $\mu_{VH}^{H\to c\bar{c}}$ parameter being shared among all channels (taken from Ref. [6]).

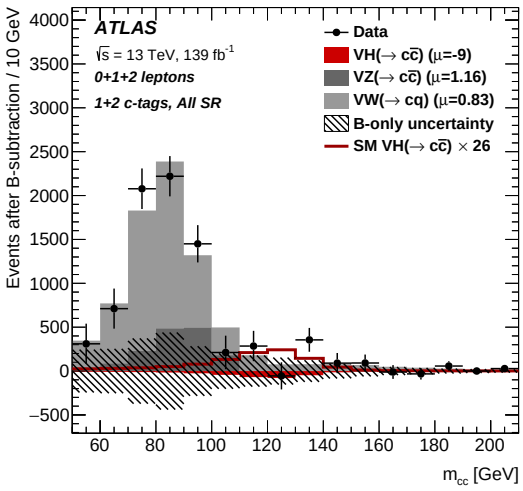


Figure 5.35: The post-fit m_{cc} distribution summed over all 1 c-tagged and 2 c-taggged signal regions after subtracting background contributions. Only the $VH, H \to c\bar{c}$, $VW, W \to cq$ and $VZ \to c\bar{c}$ processes, which are of interest, are left. The red filled histogram corresponds to the $VH, H \to c\bar{c}$ signal for the fitted value of $\mu_{VH}^{H\to c\bar{c}} = -9$, while the open red histogram corresponds to the signal expected at the 95% CL upper limit on $\mu_{VH}^{H\to c\bar{c}}$ ($\mu_{VH}^{H\to c\bar{c}} = 26$). The hatched band shows the uncertainty on the fitted background (taken from Ref. [6]).

Source of uncertainty		$\mu_{VH(c\bar{c})}$	$\mu_{VW(cq)}$	$\mu_{VZ(c\bar{c})}$
Total		15.3	0.24	0.48
Statistical		10.0	0.11	0.32
Systematic		11.5	0.21	0.36
Statistical uncertainties				
Signal normalisation		7.8	0.05	0.23
Other normalisations		5.1	0.09	0.22
Theoretical and modelling uncertainties				
$VH(\to c\bar{c})$		2.1	< 0.01	0.01
Z + jets		7.0	0.05	0.17
Top quark		3.9	0.13	0.09
W+ jets		3.0	0.05	0.11
Diboson		1.0	0.09	0.12
$VH(\to b\bar{b})$		0.8	< 0.01	0.01
Multi-jet		1.0	0.03	0.02
Simulation samples size		4.2	0.09	0.13
Experimental uncertainties				
Jets		2.8	0.06	0.13
Leptons		0.5	0.01	0.01
$E_{\mathrm{T}}^{\mathrm{miss}}$		0.2	0.01	0.01
Pile-up and luminosity		0.3	0.01	0.01
Flavour tagging	c-jets	1.6	0.05	0.16
	b-jets	1.1	0.01	0.03
	light-jets	0.4	0.01	0.06
	τ-jets	0.3	0.01	0.04
Truth-flavour tagging	ΔR correction	3.3	0.03	0.10
	Residual non-closure	1.7	0.03	0.10

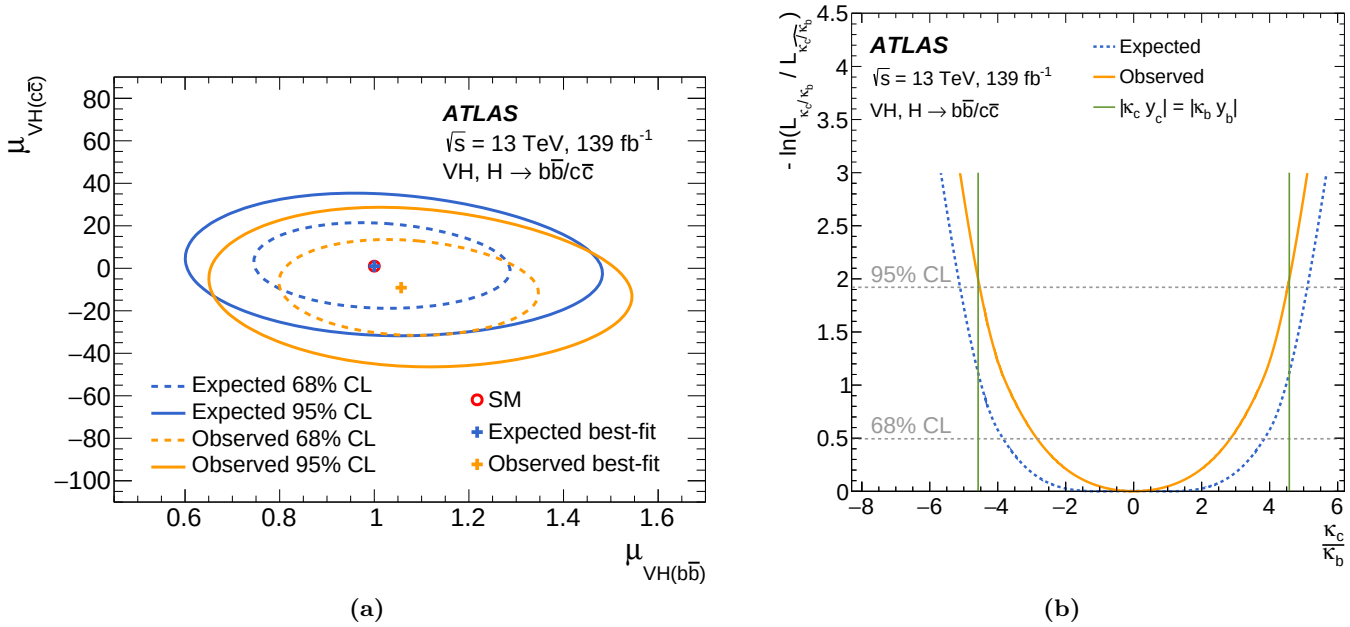
Table 5.10: Uncertainties breakdown for the fitted signal strength values $\mu_{VH}^{H\to c\bar{c}}$, $\mu_{VW}^{W\to cq}$ and $\mu_{VZ}^{Z\to c\bar{c}}$. The sum in quadrature of uncertainties from different sources may differ from the total due to correlations. In cases different values are found for the upward and downward systematic variations, the mean of the absolute values is shown (taken from Ref. [6]).

5.9.2.2 Combination of $VH, H \to b\bar{b}$ and $VH, H \to c\bar{c}$

The (resolved) $VH, H \to c\bar{c}$ [6] (described in the previous section) and resolved only $VH, H \to b\bar{b}$ [65] results have been combined (has the division of phase space is orthogonal between the two analyses) creating a combined likelihood function from the product of the likelihood of each analysis. As a result two parameters of interest are used $\mu_{VH}^{H\to c\bar{c}}$ and $\mu_{VH}^{H\to b\bar{b}}$ for the two Higgs decay processes. Those 2 signal strengths are included in both of the input likelihood functions which is important as enlighten in References [331, 332]. The experimental uncertainties shared among the two analyses are considered as correlated except the flavour tagging ones as the calibration and taggers used for b- and c-tagging are different. On the other hand, the normalization of background are considered uncorrelated. The signal strength found are:

$$\mu_{VH}^{H\to c\bar{c}} = -9 \pm 15 = -9 \pm 10\,(\text{stat.}) \pm 11\,(\text{syst.}),$$
$$\mu_{VH}^{H\to b\bar{b}} = 1.06^{+0.19}_{-0.18} = 1.06 \pm 0.12\,(\text{stat.})^{+0.15}_{-0.13}\,(\text{syst.}),$$

with a correlation coefficient of -12%. The expected and observed best-fit values and their confidence level contours are presented in Figure 5.36a. The best fit values $\mu_{VH}^{H\to c\bar{c}}$ and $\mu_{VH}^{H\to b\bar{b}}$ are then interpreted with the κ-framework parameterizing (formalism introduced in Section 1.4.5) the combined likelihood function with the coupling modifiers κ_b and κ_c while other SM couplings are set to their expected SM values and considering only SM Higgs decays. The observed ratio $|\kappa_c/\kappa_b|$, as shown in Figure 5.36b, is found to be smaller than 4.5 with a 95% CL (for an expected ratio of 5.1) which is less than the ratio of the b- and c-quarks masses $m_b/m_c = 4.578 \pm 0.008$ [333] determined with lattice QCD computations. It is the first confirmation that the Higgs-charm coupling is weaker than the Higgs-bottom coupling with a 95% CL.



(a) (b)

Figure 5.36: (a) Observed and expected best fit values of $\mu_{VH}^{H\to c\bar{c}}$ and $\mu_{VH}^{H\to b\bar{b}}$ with their 68% and 95% CL contours. (b) Expected and observed values of the combined $VH, H \to c\bar{c}$ and $VH, H \to b\bar{b}$ negative profile log-likelihood ratio as a function of κ_c/κ_b, where κ_b is a free parameter. The vertical green lines correspond to the $|\kappa_c/\kappa_b|$ values for which the Higgs-charm and Higgs-bottom couplings are equal corresponding to the situation where the charm and bottom coupling strengths to the Higgs are equal with their value being the ratio m_b/m_c: $|\kappa_c y_c| = |\kappa_b y_b| = m_b/m_c = 4.578 \pm 0.008$ [333] with $\kappa_i y_i$ the coupling strength ($i = b, c$), κ_i the coupling modifier and y_i the Yukawa coupling (taken from Ref. [6]).

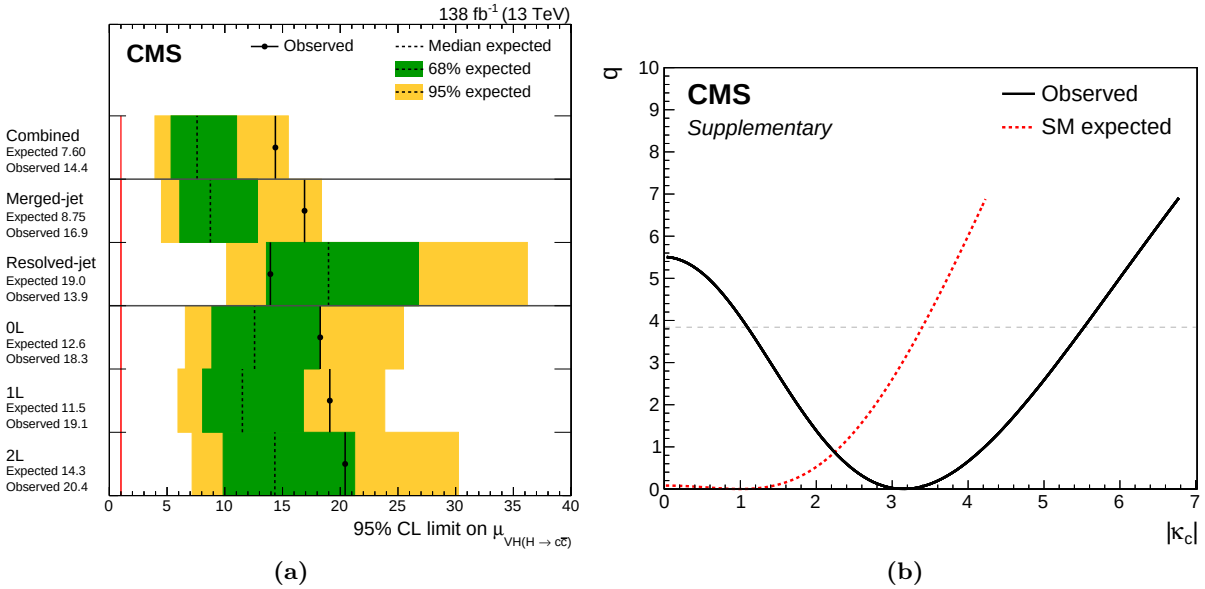
5.9.2.3 Comparison with CMS latest results

CMS published in 2022 its $VH, H \to c\bar{c}$ full Run 2 analysis results [334]. Several main differences exist between ATLAS and CMS. In terms of event reconstruction, CMS exploits both the resolved and boosted topologies whereas only the resolved topology is used by ATLAS. Events are reconstructed with the boosted topology if a large-R jet with a $p_{\mathrm{T}} > 300$ GeV is found in the event. For the resolved topology a BDT classifier is used while in the boosted topology the fits are performed with the mass of the leading large-R jet of the event. Moreover, unlike ATLAS, the CMS $VH, H \to c\bar{c}$ analysis is not orthogonal with the $VH, H \to b\bar{b}$ one hence a combination of the results is not possible. The reader is referred to the references provided in this section for more details.

The best fit value yields a signal strength $\mu_{VH}^{H\to c\bar{c}} = 7.7^{+3.8}_{-3.5}$ combining the resolved and boosted $VH, H \to c\bar{c}$ topologies. Using a CL_s criterion, the observed (expected) upper limit on $\mu_{VH}^{H\to c\bar{c}}$ found is 14.4 ($7.6^{+3.4}_{-2.3}$) as shown in the Figure 5.37. When reinterpreted with the kappa framework, the corresponding observed 95% CL interval is $1.1 < |\kappa_c| < 5.5$ for an expected constraint of $|\kappa_c| < 3.4$. This measurement is the most stringent constraint to date on $|\kappa_c|$. The constraints found are comparable to what had previously been expected at the end of the high-luminosity phase of the LHC.

Moreover, it led to the first observation of the $VZ, Z \to c\bar{c}$ process with a signal strength of $\mu_{VZ}^{Z\to c\bar{c}} = 1.01^{+0.23}_{-0.21}$ and an observed (expected) significance of 5.7 σ (5.9 σ).

Such precise measurements were possible in particular thanks a large improvement of the large-R jet c-tagging performances for the boosted topology as presented in Figure 5.38 where the new ParticleNet algorithm [335] results in an increase by a factor 4 to 7 of the rejection of non c-jets compared to the previous DeepAK15 tagger [336, 337].



(a) (b)

Figure 5.37: (a) The 95% CL upper limits on $\mu_{VH}^{H\to c\bar{c}}$. The vertical red line indicates the SM value $\mu_{VH}^{H\to c\bar{c}} = 1$. (b) Scan of the test statistic q as a function of $|\kappa_c|$. The dashed horizontal line corresponds to the 95% confidence level (taken from Ref. [334]).

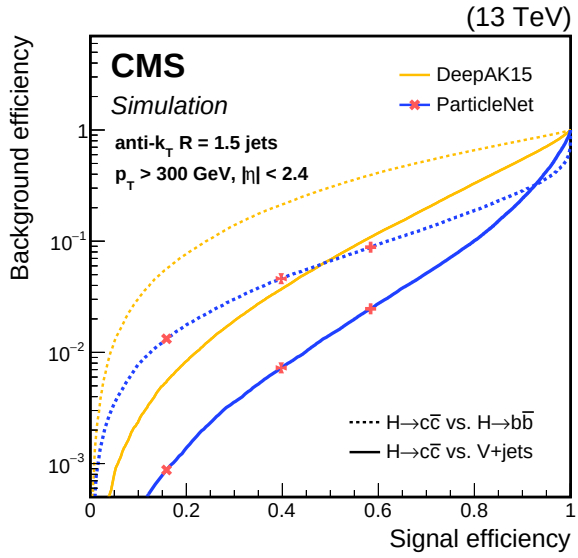


Figure 5.38: Performance of ParticleNet (blue lines) for identifying a $c\bar{c}$ pair for large-R jets with $p_{\mathrm{T}} > 300$ GeV. The solid (dashed) line shows the efficiency to correctly identify $H \to c\bar{c}$ versus the efficiency of misidentifying quarks or gluons from the $V+$jets process (versus $H \to b\bar{b}$). The red crosses represent the three working points used in the merged-jet analysis. The performance of DeepAK15 (yellow lines) used in Ref. [336] is shown for comparison (taken from Ref. [334]).

5.9.3 $VH, H \to b\bar{b}/c\bar{c}$ cross-check analyses

To validate the event selection procedure and ensure that no bias or mismodelling is introduced for the $VH, H \to b\bar{b}/c\bar{c}$ analyses, two possible cross-checks which serve as reference candles are usually performed: the dijet invariant mass check for the resolved topology and the diboson (VZ) check. This section describes those two methods.

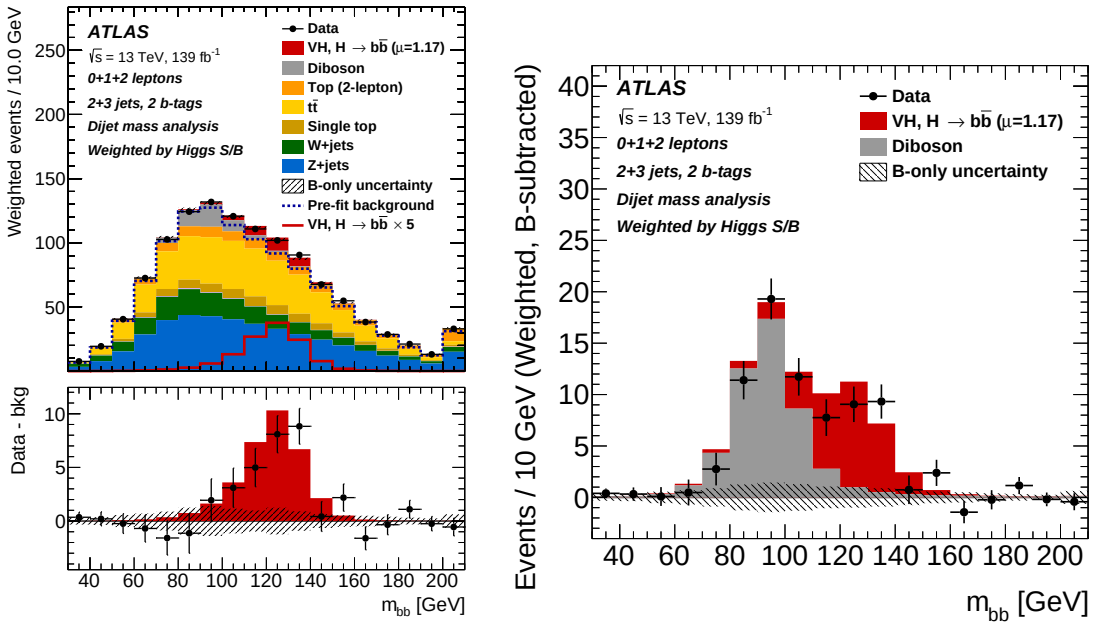
5.9.3.1 Dijet invariant mass cross-check

The dijet cross-check is a cut-based analysis where the fits of the BDT output distributions are replaced by the fits of the invariant mass (m_{bb}) distribution. The structure of the event selections is the same as for the BDT analysis but some additional selections for the Higgs candidate system are applied to increase the signal purity. This independent analysis from BDT analysis is thus an important cross-check to assess the validity of the event selections.

Such cross-check was performed for the resolved only $VH, H \to b\bar{b}$ [65] analysis, the invariant mass (m_{bb}) distribution obtained is presented in Figure 5.39. A fit with all channels combined leads to a signal strength measurement of

$$\mu_{VH}^{H\to b\bar{b}} = 1.17^{+0.25}_{-0.23} = 1.17 \pm 0.16\,(\text{stat.})^{+0.19}_{-0.16}\,(\text{syst.}),$$

compatible with the SM and the signal strength found with the BDT distributions. An observed (expected) significance of 5.5 (4.9) is found.



(a) (b)

Figure 5.39: m_{bb} distribution for the dijet mass cross-check of the $VH, H \to b\bar{b}$ analysis obtained from the weighted sum of all channels, p_{T}^{V}-regions and jet categories. The weights are computed as the ratio S/B of each region with S and B being respectively the total fitted signal and background events in each region. The expected signal contribution (WH and ZH production modes) is scaled by the measured signal strength $\mu = 1.17$. (a) The top panel includes all background while the bottom pane is after subtraction of all backgrounds. (b) Distribution after subtraction of all backgrounds except for the diboson (WZ and ZZ) processes (taken from Ref. [65]).

5.9.3.2 Diboson cross-check

Another relevant cross-check for the $VH, H \to b\bar{b}$ analysis is the so-called diboson ($VZ = ZZ$ and WZ) analysis targeting $Z \to b\bar{b}$ decays and leptonic decays of the other vector boson with minimal event selections changes with respect to the $VH, H \to b\bar{b}$ analysis. The VH signal is thus considered as background in this cross-check and the diboson VZ is considered as signal: in that case the signal strength $\mu_{VZ}^{Z\to b\bar{b}}$ is measured. Such cross-check is not performed independently for the $VH, H \to c\bar{c}$ analysis as in the fit the 3 POIs ($\mu_{VH}^{H\to c\bar{c}}$, $\mu_{VW}^{W\to cq}$ and $\mu_{VZ}^{Z\to c\bar{c}}$) are determined simultaneously as explained in Section 5.9.2.1.

The signal strength $\mu_{VZ}^{Z\to b\bar{b}}$ was measured for the $VH, H \to b\bar{b}$ combination publication [5] and results have been provided in Section 5.9.1.3. However, no public plots of the measurement were made available, so in this section the figures will be taken from the $VH, H \to b\bar{b}$ resolved only [65] and boosted only analyses [275].

For the resolved topology, a dedicated BDT is trained to discriminate VZ processes from background. A signal strength of

$$\mu_{VZ}^{Z\to b\bar{b}} = 0.93^{+0.16}_{-0.13} = 0.93^{+0.07}_{-0.06}\,(\text{stat.})^{+0.14}_{-0.12}\,(\text{syst.}),$$

is measured and is in good agreement with the SM prediction. Fits have also been performed separately for the ZZ and WZ productions modes, the results are shown in Figure 5.40. Some of the BDT fitted distributions are provided in Figure 5.41.

Since the boosted only $VH, H \to b\bar{b}$ analysis was a cut-based analysis relying on the mass m_J of the large-R jet, the VZ cross-check is performed fitting that mass distribution and the leading large-R jet of the selected events is taken as the $Z \to b\bar{b}$ candidate. Indeed, the likelihood fits are

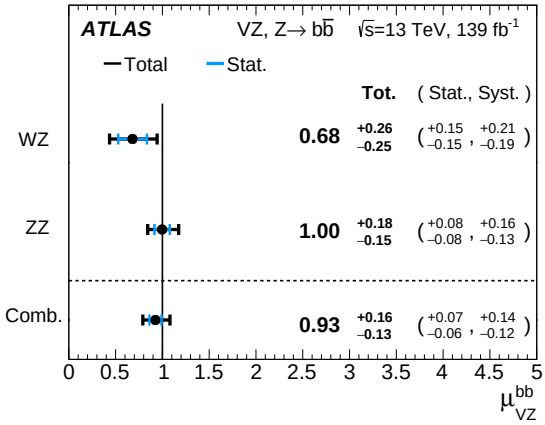


Figure 5.40: Fitted values of the VZ signal strength ($\mu_{VZ}^{Z\to b\bar{b}}$) for the WZ and ZZ production modes and their combination obtained from the resolved regime. The individual $\mu_{VZ}^{Z\to b\bar{b}}$ values for the WZ and ZZ processes are obtained from a simultaneous fit with the WZ and ZZ signal strengths floating independently(taken from Ref. [65]).

performed determining simultaneously the signal strengths $\mu_{VH}^{H\to b\bar{b}}$ POIs in each p_{T}^{V}-region and a global VZ signal strength POI which is measured to be

$$\mu_{VZ}^{Z\to b\bar{b}} = 0.91^{+0.29}_{-0.23} = 0.93 \pm 0.15\,(\text{stat.})^{+0.24}_{-0.17}\,(\text{syst.}),$$

and is in good agreement with the SM prediction. The corresponding inclusive signal strength found in that case is equal to

$$\mu_{VH}^{H\to b\bar{b}} = 0.72^{+0.39}_{-0.36} = 0.72^{+0.29}_{-0.28}\,(\text{stat.})^{+0.26}_{-0.22}\,(\text{syst.}).$$

Fits have also been performed separately for the ZZ and WZ productions modes, the results are shown in Figure 5.42. Some of the mass fitted distributions are provided in Figure 5.43.

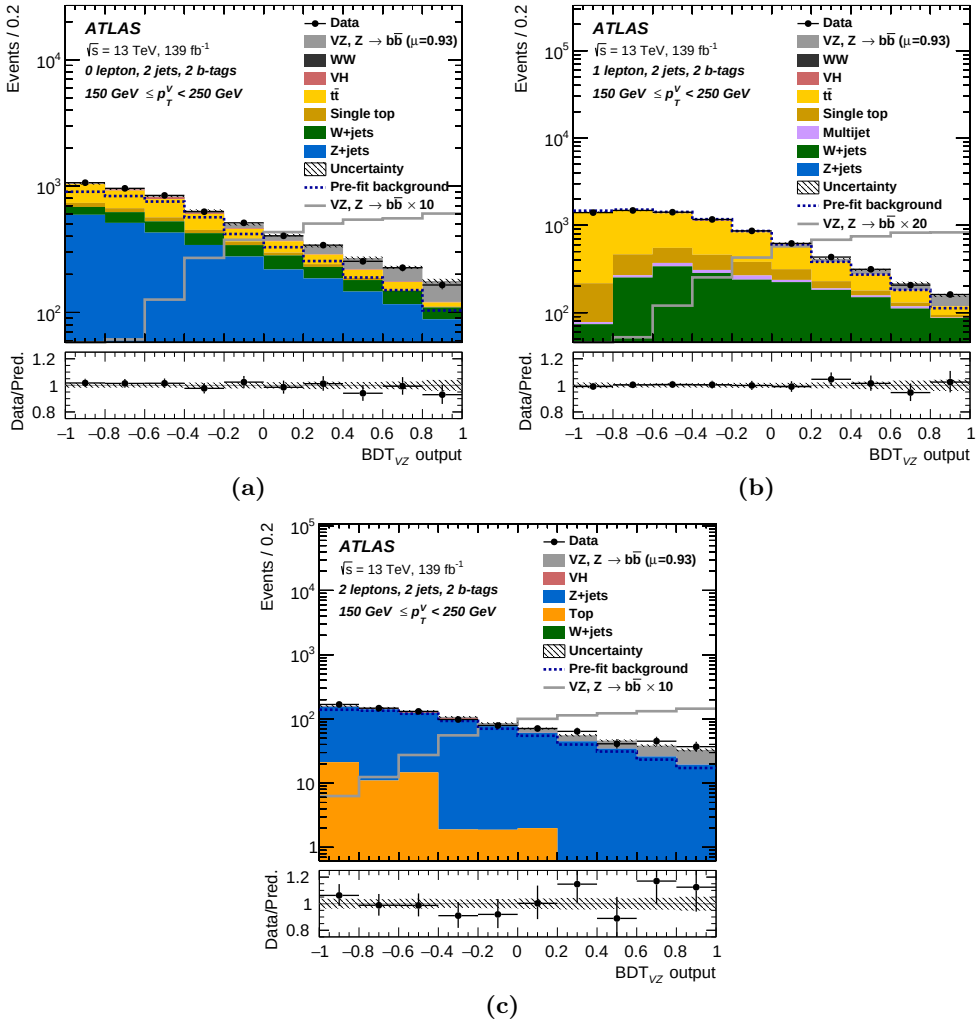


Figure 5.41: Resolved BDT_{VZ} post-fit distributions for the (a) 0-, (b) 1- and (c) 2-lepton channels. The VZ signal shown as a filled histogram on top of the fitted backgrounds is normalized with the VZ signal strength extracted from data ($\mu_{VZ}^{Z\to b\bar{b}} = 0.93$), and unstacked as an unfilled histogram, scaled by the factor indicated in the legend. The dashed histogram shows the total pre-fit background. The size of the combined statistical and systematic uncertainty for the sum of the fitted signal and background is indicated by the hatched band. The ratio of the data to the sum of the VZ fitted signal and background is shown in the lower panel (taken from Ref. [65]).

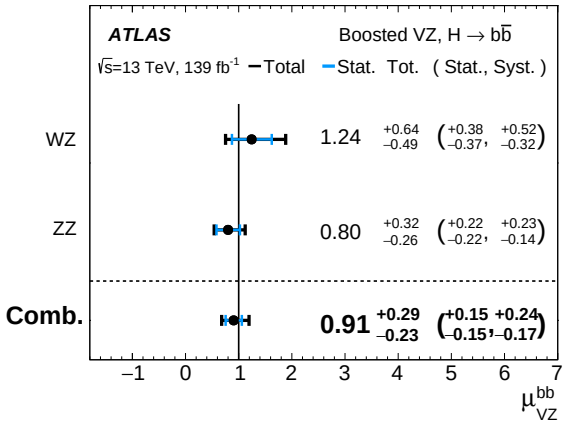


Figure 5.42: Fitted values of the VZ signal strength ($\mu_{VZ}^{Z\to b\bar{b}}$) for the WZ and ZZ production modes and their combination obtained from the boosted regime. The individual $\mu_{VZ}^{Z\to b\bar{b}}$ values for the WZ and ZZ processes are obtained from a simultaneous fit with the WZ and ZZ signal strengths floating independently(taken from Ref. [275]).

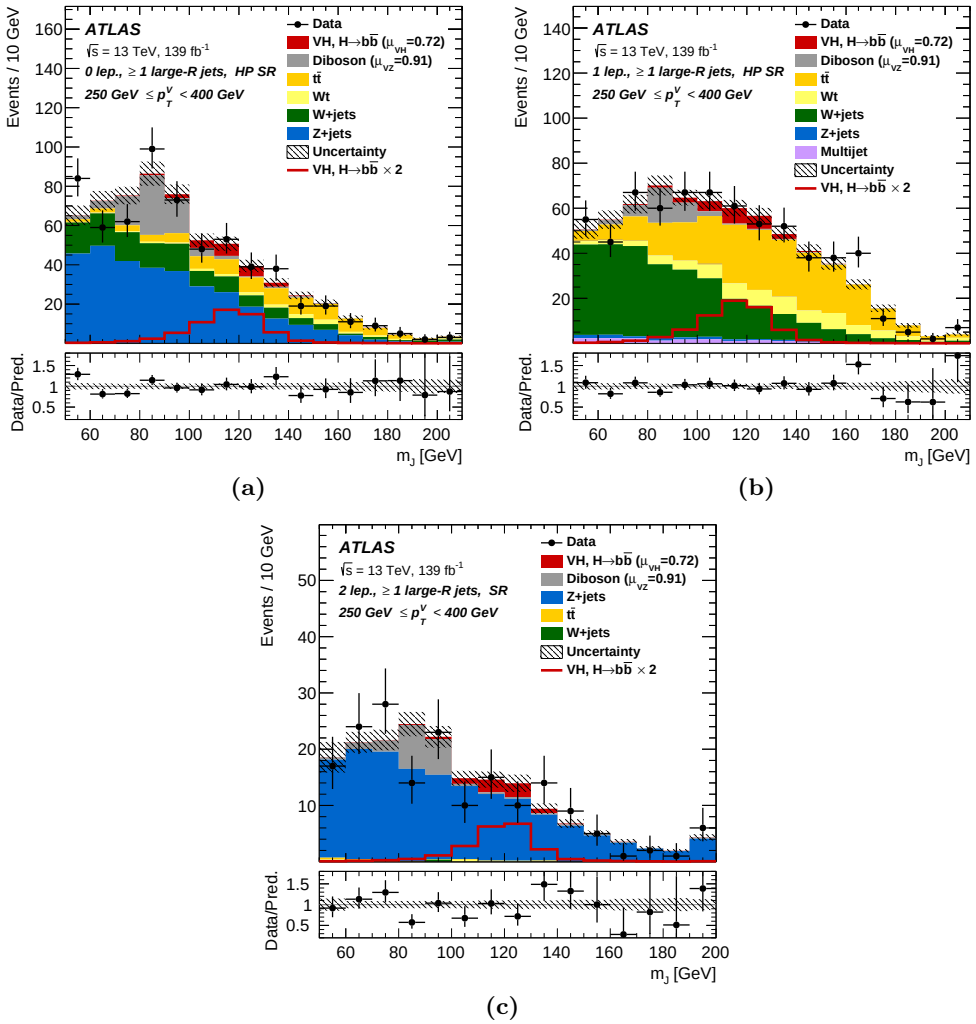


Figure 5.43: The post-fit large-R jet mass distributions in the (a) high purity SR 0-lepton, (b) high purity SR 1-lepton and (c) SR 2-lepton channels (the HP SR, LP SR and top CR are merged in the 2-lepton channel) for the $[250, 400]$ GeV p_{T}^{V}-region . The background contributions after the likelihood fit are shown as filled histograms. The VH signal shown as a filled histogram on top of the fitted backgrounds is normalized with the signal strength ($\mu_{VH}^{H\to b\bar{b}} = 0.72$), and unstacked as an unfilled histogram, scaled by the factor indicated in the legend. The VZ signal shown as a filled histogram is normalized with the VZ signal strength extracted from data ($\mu_{VZ}^{Z\to b\bar{b}} = 0.91$). The size of the combined statistical and systematic uncertainty for the sum of the fitted signal and background is indicated by the hatched band. The highest bin in the distributions contains the overflow. The ratio of the data to the sum of the fitted signal and background is shown in the lower panel (taken from Ref. [275]).

5.10 Expected results for the Legacy Run 2 analysis

For the Legacy Run 2 publication the goal is to perform a simultaneous fit of the $VH, H \to b\bar{b}$ resolved, $VH, H \to b\bar{b}$ boosted and $VH, H \to c\bar{c}$ datasets. However, as I am writing this chapter the implementation of the fit is still ongoing. As a consequence the expected results for the two analyses are presented separately and obtained from two different likelihood fits. Moreover, the analysis targets for the $VH, H \to b\bar{b}$ measurement a more granular STXS binning than previously dividing phase space according to the number of extra jets in the event either 0 or ≥ 1 respectively for the 2-jets and 3-jets (or $\geq$ 3-jets in the 2-lepton channel) categories for the resolved topology. The implementation of such feature was not ready, but the expected STXS measurements presented in this section is anyway more granular, 9-STXS bins against 7 in previous publications, as it includes the $p_{\mathrm{T}}^{V} > 600$ GeV region both for the ZH and WH processes.

The results provided are preliminary as some analysis choices are still under investigation. The analysis is still blinded so either unconditional Asimov dataset fits (no constraint on the POIs) or data conditional fits (*i.e.* the signal strength are set to unity) are performed. In addition, since the modelling studies are still ongoing, the associated uncertainties are estimated based on the previous publication results *i.e.* with the "old" BDT reweighting technique or old shape comparisons of key variables such as m_{bb} and p_{T}^{V}: it might not be fully accurate for the current results. Hence, the expected results allow to have an insight of the sensitivity to $VH, H \to b\bar{b}/c\bar{c}$ processes, but they might change and be a bit different for the Legacy Run 2 publication.

5.10.1 Expected $VH, H \to b\bar{b}$ results

With the full Run 2 dataset, the expected inclusive sensitivity to $VH, H \to b\bar{b}$ processes is found to be of 6.9 standard deviations with a conditional dataset fit *i.e.* using a single POI ($\mu_{VH}^{H\to b\bar{b}}$) across all channels. It corresponds to a 10% improvement of the sensitivity compared to the expected significance, 6.3 standard deviations, of the previous $VH, H \to b\bar{b}$ resolved and boosted topologies combination [5]. The expected uncertainties on the signal strength performing a likelihood fit either with a POI per lepton channel (3 POIs in total) or in the 9 STXS bins (*i.e.* 9 POIs) are reported in Figure 5.44. The associated cross-section measurements and the correlations between the different truth level categories are also shown: a relatively low level of correlations of the order of few percents is expected for most categories up to 20% for consecutive p_{T}^{V}-regions above 250 GeV. The theoretical cross-section predictions and associated uncertainties are derived following the procedure described in Reference [338]. The combination of channels yields an expected signal strength of

$$\mu_{VH}^{H\to b\bar{b}} = 1.00^{+0.15}_{-0.14} = 1.00^{+0.11}_{-0.11}\,(\text{stat.})^{+0.10}_{-0.10}\,(\text{syst.}),$$

hence the precision on the signal strength measurement would be improved by approximately 20% when compared to the observed signal strength $\mu_{VH}^{H\to b\bar{b}} = 1.00^{+0.18}_{-0.17}$ from the previous publication. The impact of group of uncertainties on the inclusive signal strength is summarized in Table 5.11: dominating sources of uncertainties for this measurement are hence the statistical and the modelling uncertainties. The highest impact of individual nuisance parameters and their pulls is shown in Figure 5.45. The improvement of the sensitivity is mainly due to the new DL1r b-tagging algorithm and the use of an MVA approach for the boosted topology: both of them increase the purity in the regions and/or bins with the highest sensitivity to signal.

All the BDT output distributions for signal regions obtained with a conditional data fit for the resolved and boosted regimes can be found in Figures 5.46, 5.47 and 5.48 respetively for the 0-, 1- and 2-lepton channels. An overall good agreement between data and simulation is observed for the non-blinded bins[5]. Finally, the Figure 5.49 presents the invariant mass of the leading large-R jet in

the top control regions for the 0- and 1-lepton channels, distribution which is used to normalize the $t\bar{t}$ background in the boosted topology.

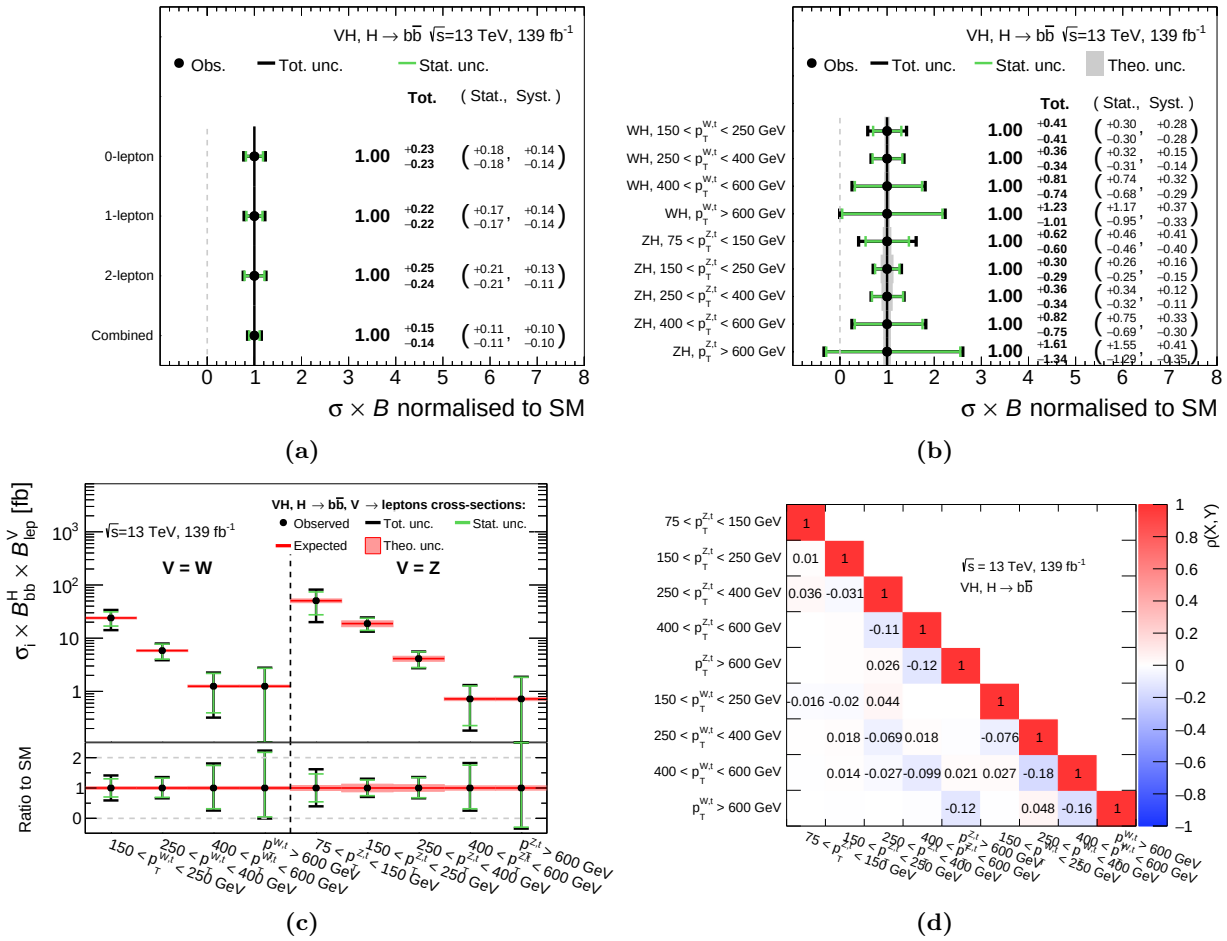


Figure 5.44: (a) Expected $VH, H \to b\bar{b}$ signal strength uncertainties for a single (referred to as combined) or a 3 POIs (one POI per channel) fit. Expected (b) signal strength and (c) cross-section uncertainties, (d) and correlations for a 9 STXS POIs measurement. The theoretical cross-section predictions and associated uncertainties for the truth p_{T}^{V}-bins above 400 GeV are taken from an inclusive computation for $p_{\mathrm{T}}^{V,t} > 400$ GeV and have not yet been calculated separately for $400 < p_{\mathrm{T}}^{V,t} < 600$ GeV and $p_{\mathrm{T}}^{V,t} > 600$ GeV.

5. The $VH, H \to b\bar{b}/c\bar{c}$ blinding procedure consists in blinding the highest BDT score bins in signal regions until at least 60% of the signal event yields is excluded for the BDT distribution considered. The invariant mass distribution in SR of the Higgs candidate is also blinded between 80 and 140 GeV. Finally, for any other distribution, if the ratio of signal over background yields (S/B) is found to be greater than 10% for a given bin then it is blinded.

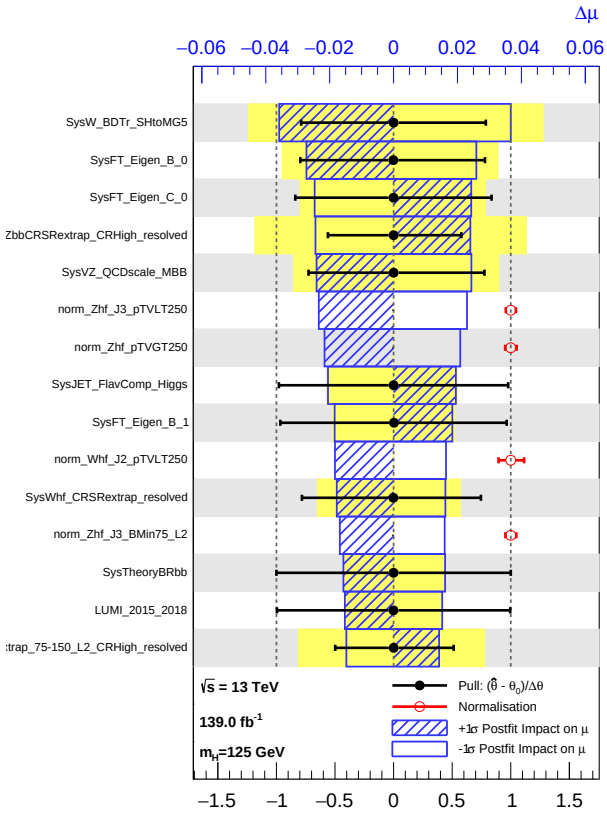


Figure 5.45: Expected nuisance parameters ranking for a single POI fit ($\mu_{VH}^{H\to b\bar{b}}$) performed over all channels. The blue (respectively yellow) boxes show the post-fit (respectively pre-fit) impact of systematic uncertainties on the signal strength $\hat{\mu}$ sorted in decreasing post-fit amplitude order. The boxes correspond to the variations of $\hat{\mu}$ (top x-axis), when fixing the corresponding individual NP to its post-fit (or pre-fit) value modified upwards or downwards by its post-fit (or pre-fit) uncertainty ($\theta_i \pm \sigma_{\theta_i}$), and repeating the fit. The impact of post-fit up- and down- variations is respectively symbolized with a dashed and empty blue boxes. The filled circles show the pull of the nuisance parameters (bottom x-axis) *i.e.* the deviation of the fitted value ($\hat{\theta}$) for the NPs from their pre-fit expected value (θ_0) normalized with respect to their pre-fit uncertainty ($\Delta\theta$). The error bars show the post-fit uncertainties on the NP $\hat{\theta}$ normalized with the pre-fit uncertainty. The open circles show the fitted values and uncertainties of the normalization parameters that are freely floating in the fit.

POI Inclusive $\mu_{VH}^{H\to b\bar{b}}$	Central Value 1.00		
Set of nuisance parameters	Impact on error		
	Signed		Symmetrized
Total	+0.148	−0.144	±0.146
Data Stat	+0.110	−0.108	±0.109
Full Syst	+0.099	−0.094	±0.097
Data stat only	+0.101	−0.100	±0.100
Top-$e\mu$ CR stat	+0.012	−0.012	±0.012
Floating normalizations	+0.031	−0.032	±0.031
Modelling: VH	+0.024	−0.018	±0.021
Modelling: Background	+0.070	−0.069	±0.069
Multijet	+0.000	−0.000	±0.000
Single top	+0.007	−0.007	±0.007
$t\bar{t}$	+0.021	−0.021	±0.021
$W+$jets	+0.042	−0.042	±0.042
$Z+$jets	+0.032	−0.032	±0.032
Diboson	+0.030	−0.029	±0.029
MC stat	+0.025	−0.026	±0.025
Experimental Syst	+0.038	−0.032	±0.035
Lepton	+0.004	−0.004	±0.004
$E_{\mathrm{T}}^{\mathrm{miss}}$	+0.009	−0.009	±0.009
Small-R jet	+0.029	−0.025	±0.027
Large-R jet	+0.009	−0.009	±0.009
Small-R jet b-tagging (b-jet)	+0.000	−0.000	±0.000
Small-R jet b-tagging (c-jet)	+0.000	−0.000	±0.000
Small-R jet b-tagging (l-jet)	+0.000	−0.000	±0.000
Small-R jet b-tagging (extrap. from charm)	+0.000	−0.000	±0.000
VR track-jet b-tagging (b-jet)	+0.003	−0.003	±0.003
VR track-jet b-tagging (c-jet)	+0.004	−0.003	±0.004
VR track-jet b-tagging (l-jet)	+0.001	−0.001	±0.001
VR track-jet b-tagging (extrap. from charm)	+0.000	−0.000	±0.000
Pile-up	+0.007	−0.005	±0.006
Luminosity	+0.017	−0.013	±0.015

Table 5.11: Breakdown of uncertainties with a single POI fit ($\mu_{VH}^{H\to b\bar{b}}$) performed over all channels. Uncertainties are grouped into categories: data statistics (Data Stat), systematic uncertainties together with MC statistical uncertainties (Full Syst), MC statistics only (MC Stat), and different sub-groups of systematic uncertainties.

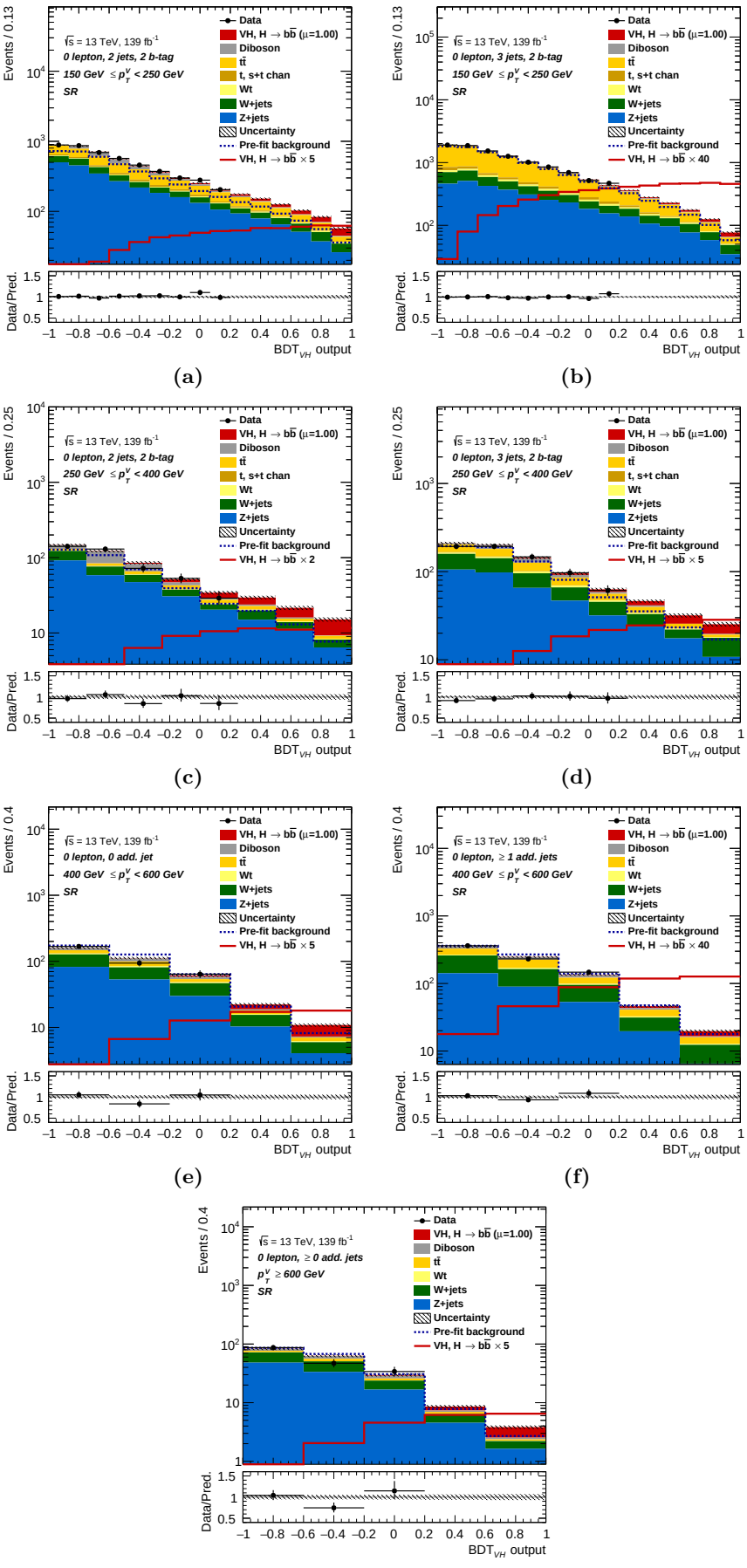


Figure 5.46: Conditional ($\mu = 1$) post-fit BDT output distribution for the SRs of the $VH, H \to b\bar{b}$ (a), (b), (c) and (d) resolved, and (e), (f) and (g) boosted topologies for the 0-lepton channel. (a) and (b) 2-jets and 3-jets SRs for $p_{\mathrm{T}}^{V} \in [150, 250]$ GeV, (c) and (d) 2-jets and 3-jets SRs for $p_{\mathrm{T}}^{V} \in [250, 400]$ GeV, (e) and (f) HP and LP SRs for $p_{\mathrm{T}}^{V} \in [400, 600]$ GeV, and (g) HP-plus-LP SR for $p_{\mathrm{T}}^{V} > 600$ GeV. The ratios of the data to the sum of the fitted signal and background are shown in the lower panels.

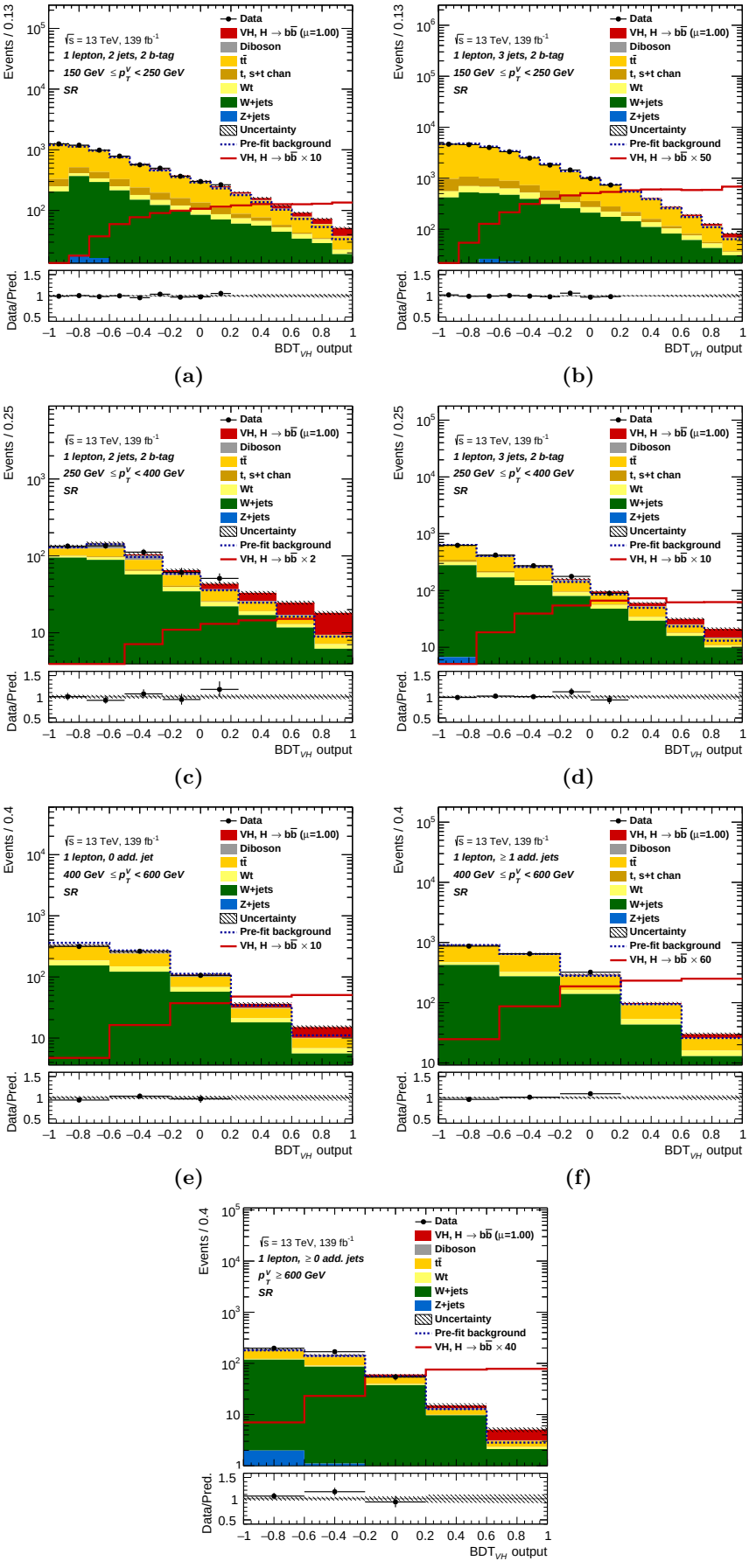


Figure 5.47: Conditional ($\mu = 1$) post-fit BDT output distribution for the SRs of the $VH, H \to b\bar{b}$ (a), (b), (c) and (d) resolved, and (e), (f) and (g) boosted topologies for the 1-lepton channel. (a) and (b) 2-jets and 3-jets SRs for $p_{\mathrm{T}}^{V} \in [150, 250]$ GeV, (c) and (d) 2-jets and 3-jets SRs for $p_{\mathrm{T}}^{V} \in [250, 400]$ GeV, (e) and (f) HP and LP SRs for $p_{\mathrm{T}}^{V} \in [400, 600]$ GeV, and (g) HP-plus-LP SR for $p_{\mathrm{T}}^{V} > 600$ GeV. The ratios of the data to the sum of the fitted signal and background are shown in the lower panels.

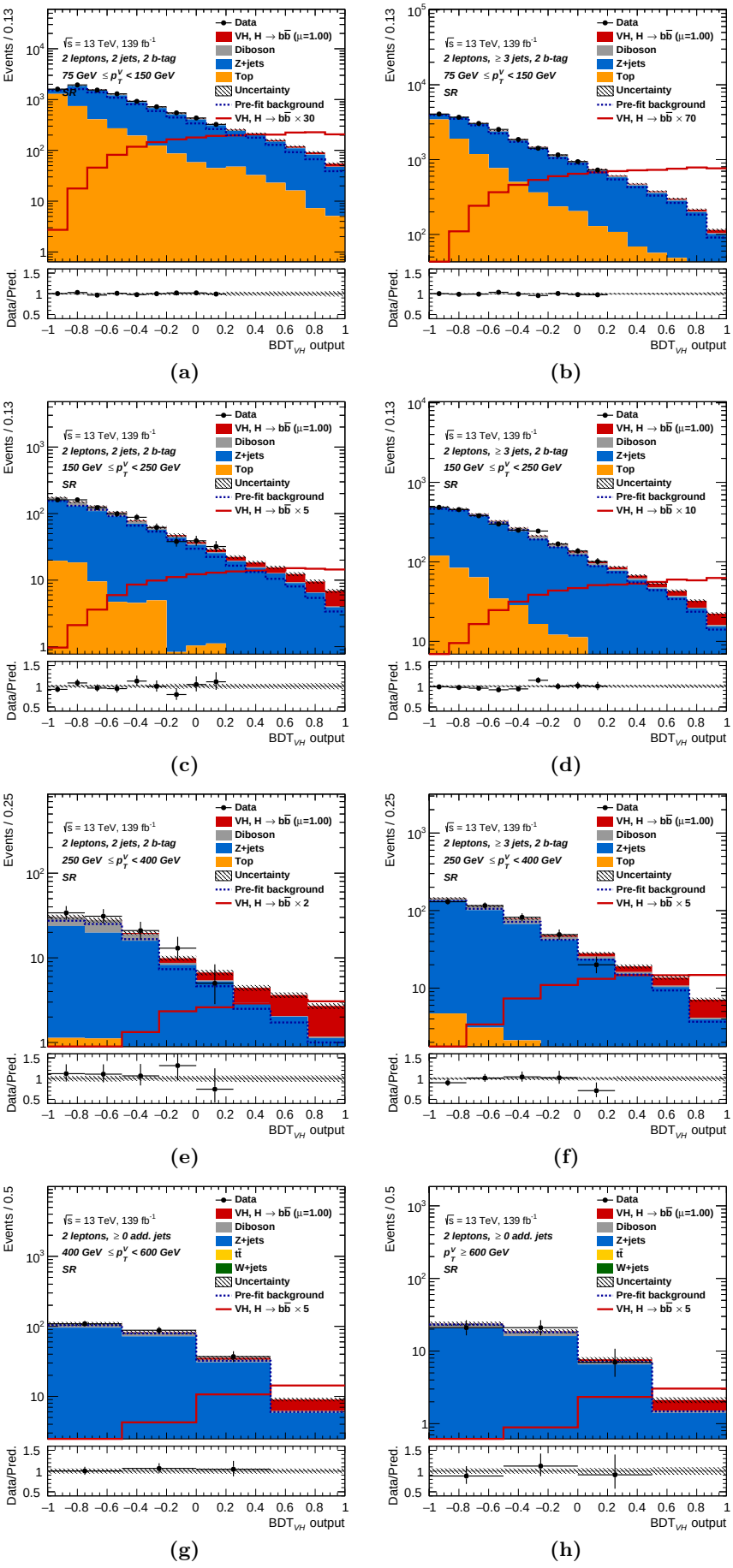


Figure 5.48: Conditional ($\mu = 1$) post-fit BDT output distribution for the SRs of the $VH, H \to b\bar{b}$ (a), (b), (c), (d), (e) and (f) resolved, and (g) and (h) boosted topologies for the 2-lepton channel. (a), (b),2-jets and $\geq$ 3-jets SRs for $p_{\mathrm{T}}^{V} \in [75, 150]$ GeV, (c) and (d) 2-jets and $\geq$ 3-jets SRs for $p_{\mathrm{T}}^{V} \in [150, 250]$ GeV, (e) and (f) 2-jets and $\geq$ 3-jets SRs for $p_{\mathrm{T}}^{V} \in [250, 400]$ GeV, (g) SR for $p_{\mathrm{T}}^{V} \in [400, 600]$ GeV, and (h) SR for $p_{\mathrm{T}}^{V} > 600$ GeV. The ratios of the data to the sum of the fitted signal and background are shown in the lower panels.

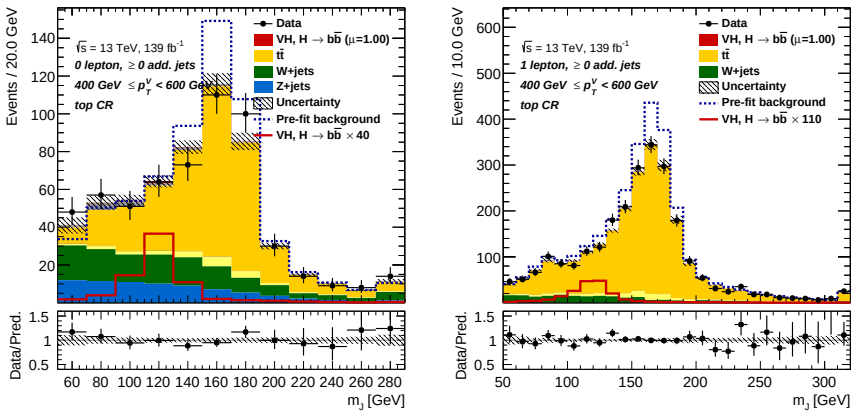


Figure 5.49: Conditional ($\mu = 1$) post-fit leading large-R jet invariant mass (m_J) distribution for the $VH, H \to b\bar{b}$ (a) 0-lepton and (b) 1-lepton channels in the boosted $400 < p_{\mathrm{T}}^{V} < 600$ GeV top CR. The ratios of the data to the sum of the fitted signal and background are shown in the lower panels.

5.10.2 Expected $VH, H \to c\bar{c}$ results

The expected 95% CL upper limit for the $VH, H \to c\bar{c}$ signal strength ($\mu_{VH}^{H\to c\bar{c}}$) is shown in Figure 5.50. The expected constraint of 16 times the SM that is found is almost at the level of the observed upper limit of 14.4 obtained by CMS [334]. Compared to the previous ATLAS publication [6] were the expected CL upper limit was 31, the sensitivity to $VH, H \to c\bar{c}$ processes has been increased by a factor 2 which is a significant improvement. This is mainly due to the new DL1r b- and c-taggers which allows having a higher purity in signal regions and also thanks to the MVA approach which enhances signal from background discrimination.

The Figure 5.51, 5.52 and 5.53 present the post-fit conditional BDT output distributions in the tight-tight and tight-loose signal regions respectively for the 0-, 1- and 2-lepton channels. An overall good agreement between the data and simulation is found in the non-blinded bins. In those figures, *hf* and *mf* stand for heavy flavour (bb and cc components only) and mixed flavour (bc, bl and cl).

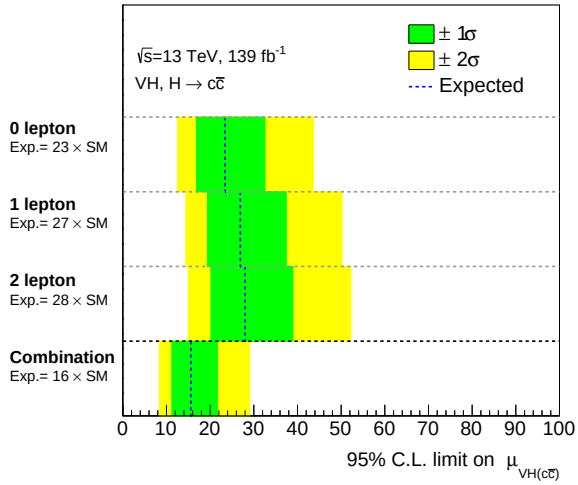


Figure 5.50: Expected 95% CL level upper limit for the $VH, H \to c\bar{c}$ signal strength ($\mu_{VH}^{H\to c\bar{c}}$). For the single channel fits, a separate $\mu_{VH}^{H\to c\bar{c}}$ parameter for each channel is used, while the combined fit (gathering the three leptonic channels) is performed with the $\mu_{VH}^{H\to c\bar{c}}$ parameter being shared among all channels.

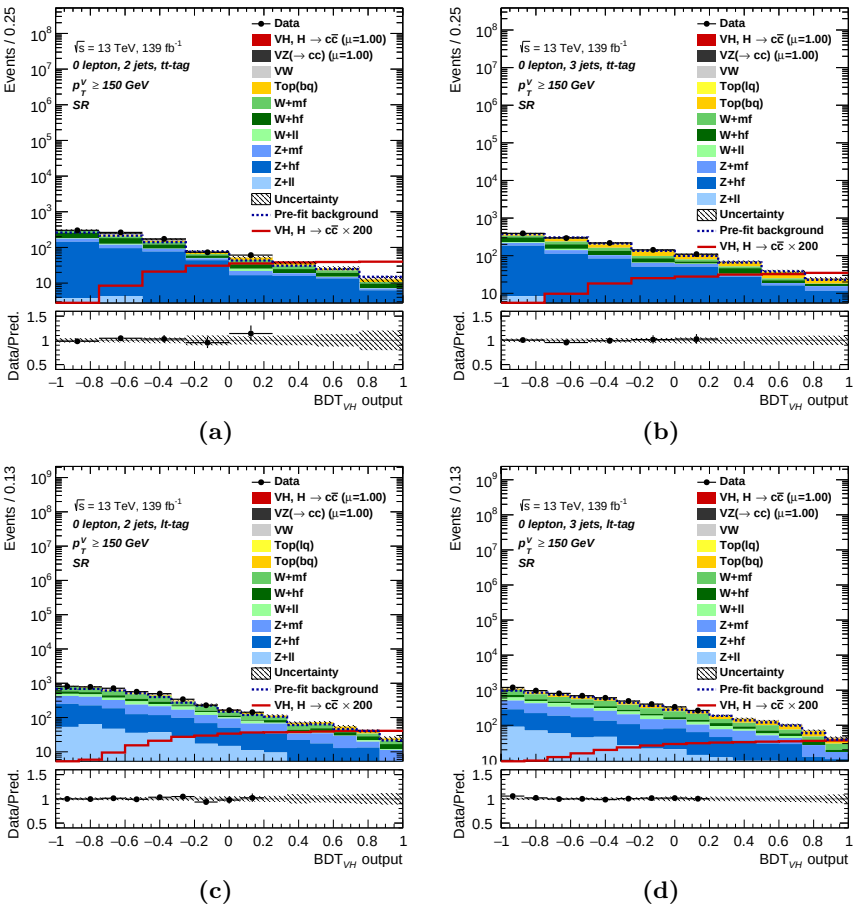


Figure 5.51: Conditional ($\mu = 1$) post-fit BDT output distribution for the (a) 2-jets and (b) 3-jets tight-tight (TT), and (c) 2-jets and (d) 3-jets loose-tight (LT) SRs in the $VH, H \to c\bar{c}$ 0-lepton channel for the $p_{\mathrm{T}}^{V} > 150\,\mathrm{GeV}$ region. The ratios of the data to the sum of the fitted signal and background are shown in the lower panels.

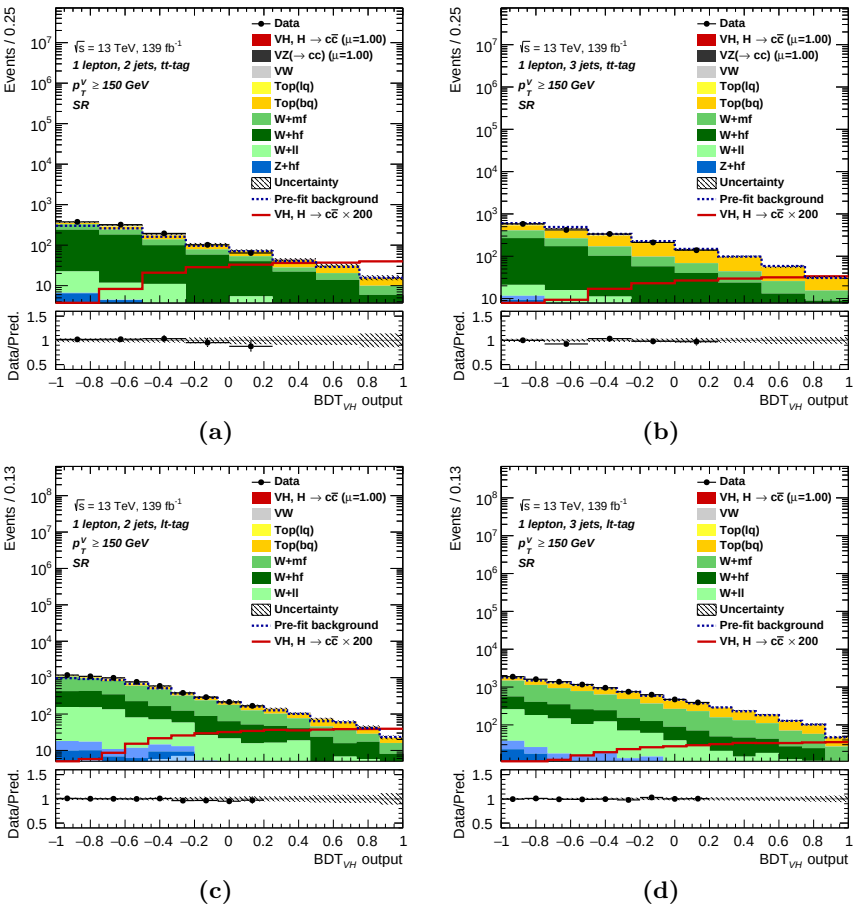


Figure 5.52: Conditional ($\mu = 1$) post-fit BDT output distribution for the (a) 2-jets and (b) 3-jets tight-tight (TT), and (c) 2-jets and (d) 3-jets loose-tight (LT) SRs in the $VH, H \to c\bar{c}$ 1-lepton channel for the $p_{\mathrm{T}}^{V} > 150\,\mathrm{GeV}$ region. The ratios of the data to the sum of the fitted signal and background are shown in the lower panels.

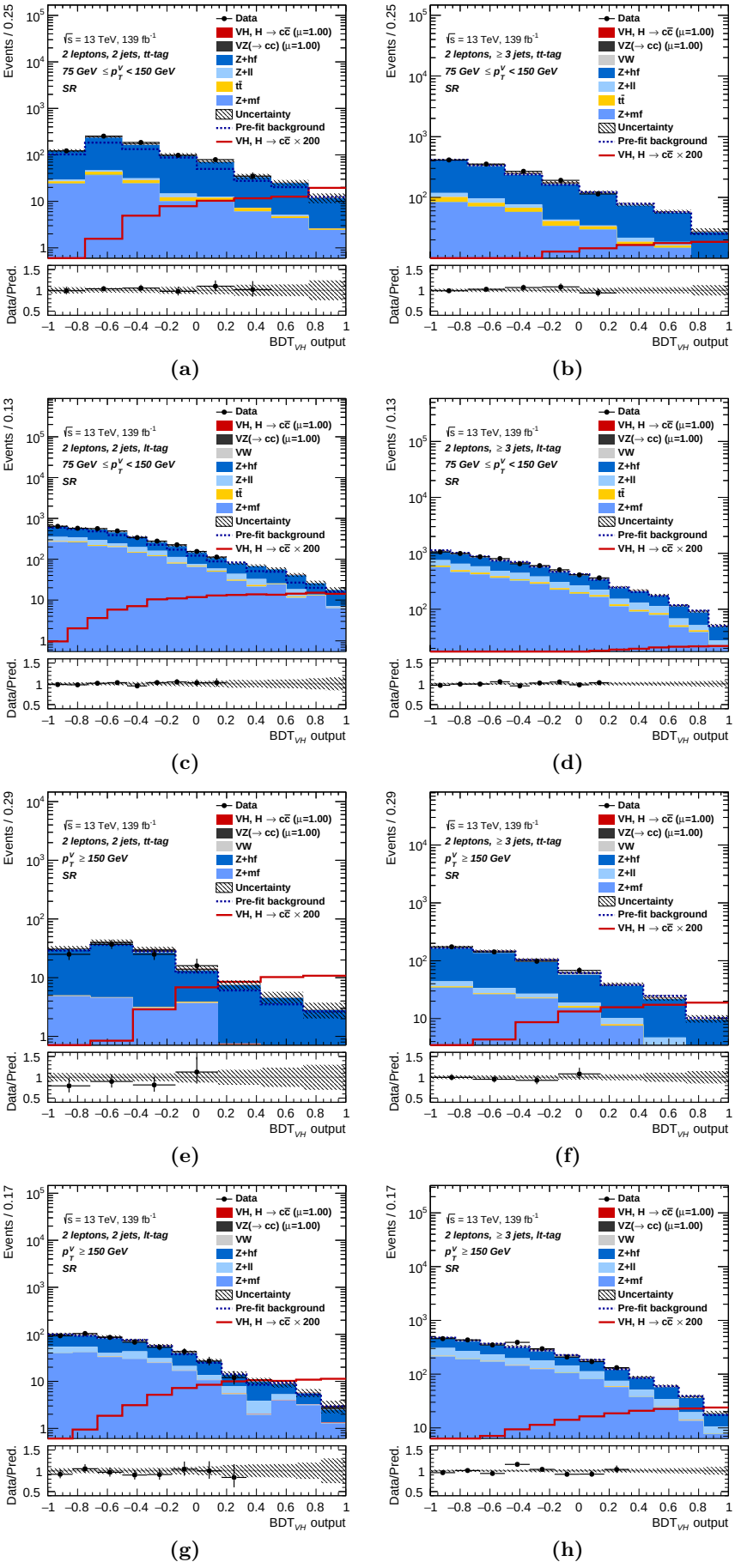


Figure 5.53: Conditional ($\mu = 1$) post-fit BDT output distribution for the (a) and (e) 2-jets, and (b) and (f) $\geq$ 3-jets tight-tight (TT), and (c) and (g) 2-jets, and (d) and (h) $\geq$ 3-jets loose-tight (LT) SRs in the $VH, H \to c\bar{c}$ 2-lepton channel for the (a), (b), (c) and (d) $75 < p_{\mathrm{T}}^{V} < 150$ GeV, and (e), (f), (g) and (h) $p_{\mathrm{T}}^{V} > 150$ GeV regions. The ratios of the data to the sum of the fitted signal and background are shown in the lower panels.

5.11 Future improvements and prospects for the $VH, H \to b\bar{b}/c\bar{c}$ analysis

An important aspect of any analysis is working for future improvements and performing extrapolation of their results in the future. This section discusses the expected possible ameliorations and limits that could be reached by the ATLAS $VH, H \to b\bar{b}/c\bar{c}$ analysis.

5.11.1 Run 3 of the LHC

The key improvements achievable in the coming years, especially that could and/or should be reached for Run 3 physics analyses, are presented. The possible improvements detailed in this section are of course also valid beyond the Run 3.

On the 5th July 2022, one day after the 10th birthday of the Higgs discovery announcement, the Run 3 of the LHC started and will span until 2026. An integrated luminosity of 300 fb^{-1} is expected[6] to be delivered by the LHC at an unprecedented center-of-mass energy of $\sqrt{s} = 13.6$ TeV. As a result the amount of data accumulated will increase by approximately a factor 2 with respect to Run 2 and a factor 3 combining Run 2 and Run 3.

The $VH, H \to b\bar{b}/c\bar{c}$ analysis can benefit from this increase of collected data as currently the statistical uncertainty remains one of the leading source of uncertainty and is a limitation for the boosted $VH, H \to b\bar{b}$ and (resolved) $VH, H \to c\bar{c}$ analyses as reported in Tables 5.9 and 5.10.

Another primordial aspect for the $VH, H \to b\bar{b}/c\bar{c}$ analysis is the ability to identify b- and c-jets and discriminate them from other flavour initiated jets. The switch from the BDT (MV2c10) tagger to the ANN (DL1r) tagger had a non-negligible impact as it increased for instance by around 10% the significance for each channel in the $VH, H \to b\bar{b}$ resolved regime as reported for example in the next chapter in Tables 6.1 and 6.2. The flavour tagging algorithm performances are hence playing a crucial role to enhance the sensitivity to the Higgs-to-bottom and Higgs-to-charm couplings. Their developments and improvements are an active domain both in ATLAS and CMS. For instance the recent $VH, H \to c\bar{c}$ CMS measurement [334], presented in Section 5.9.2.3, achieved a sensitivity comparable to what was thought to be only accessible at the end of the high luminosity phase of the LHC according to projection studies [339]. Such impressive results were reached in particular thanks to a large improvement of the c-tagging performances of CMS using a GNN based on low level information coming from detector measurements for the $VH, H \to c\bar{c}$ boosted topology. On the other hand, the ATLAS DL1r c-tagging algorithm is currently designed using the same input variables as for b-tagging, which is more mature and established within the collaboration. Thus, the c-tagging algorithms need to be improved to match CMS recent improvements. A newly end-to-end GNN b/c-tagger for small-R jets called GN1 tagger [340] was developed by ATLAS. Additional improvements to the c-tagging performances by around 40% in b-jet rejections have been demonstrated for instance thanks to the usage of this GNN algorithm as shown in Figure 5.54. In addition, information based on semi-leptonic decays to muons from heavy hadrons have shown a further boost in the performance by around 15%. Additional consideration shall be to further improve the performances at high energy, performing flavour-tagging for large-R jet similarly to the GNN (ParticleNet) tagger of CMS. However, the GNN advancement investigated so far for tagging small-R jets have not yet been ported to large-R jets. Another algorithm is available within ATLAS for such purpose the Xbb tagger [212] which aims at identifying large-R jets which are the product of a particle (X) decaying to a pair of b-quarks ($X \to b\bar{b}$) such as a Z boson or a Higgs boson. The Xbb tagger could potentially be extended to c-tagging. The Xbb algorithm is a

6. The CERN accelerator complex activity will be reduced by 20% in 2023 due to the energy crisis Europe is currently facing. This situation is to some extent the consequence of the war and invasion conducted by the Russian government against Ukraine. I want here to express my full support to the Ukrainian people, I deeply regret and condemn the actions of the Russian government and of any person supporting those decisions.

feed-forward neural network algorithm trained to discriminate $X \to b\bar{b}$ decays from multijet and top processes. The performances of that tagger are compared to the default DL1r tagger and old MV2 tagger in Figure 5.55 and show a significant improvement. However, the $VH(b\bar{b}/c\bar{c})$ group tested the Xbb tagger and performances for the analysis were found to be similar to the DL1r algorithm possibly because of the other sources of background $W+$jets, $Z+$jets, diboson... Moreover, the timeline for the calibration of the Xbb tagger was not matching with the $VH, H \to b\bar{b}/c\bar{c}$ analysis one, so it was decided not to use this new tagger. Possibly in the future and with improvements, the Xbb tagger could become a key ingredient of the analysis and could enhance the sensitivity of the $VH, H \to b\bar{b}$ boosted analysis.

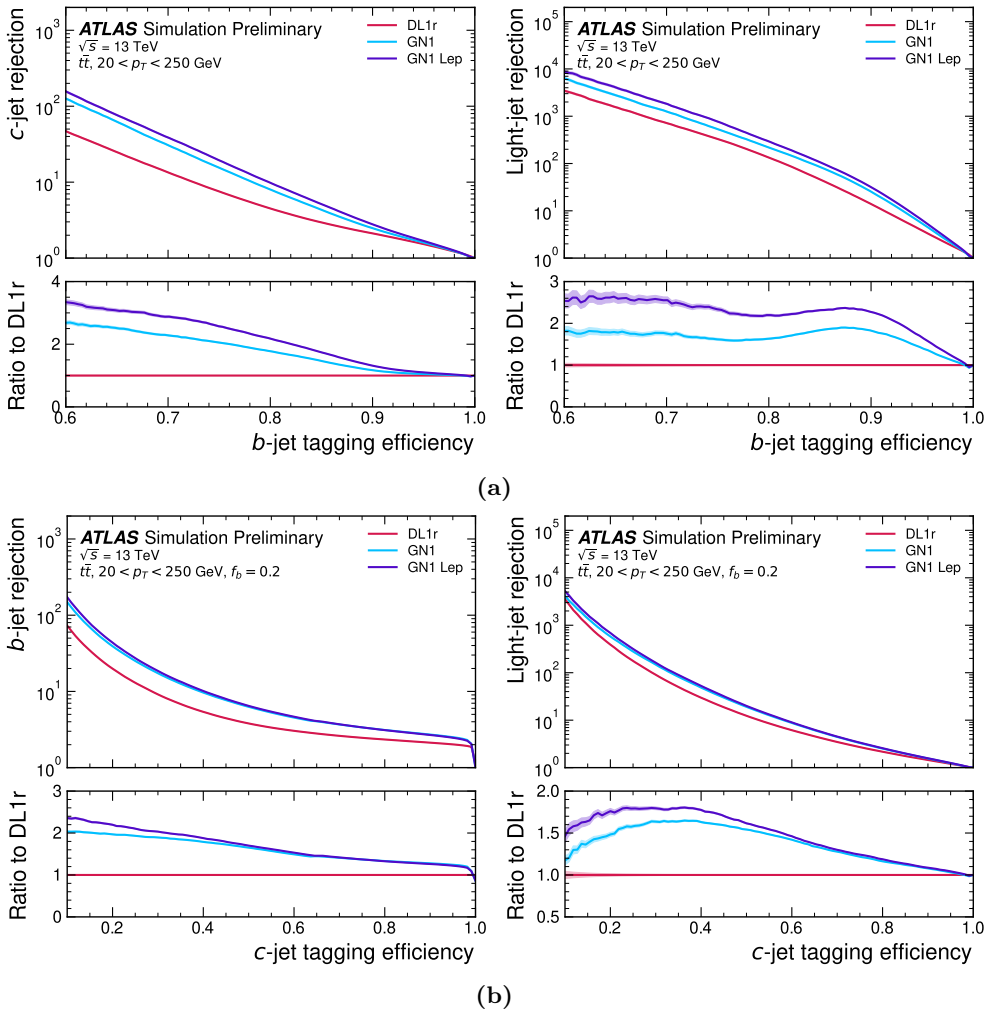


Figure 5.54: GN1 b/c-tagger performances compared to the current ATLAS default tagger (DL1r) evaluated in $t\bar{t}$ reference samples for small-R jets with $20 < p_{\mathrm{T}} < 250$ GeV. (a) The c-jet (left) and light-jet (right) rejections as a function of the b-jet tagging efficiency. (b) The b-jet (left) and light-jet (right) rejections as a function of the c-jet tagging efficiency. The bottom panels show the ratio of the performances with respect to the DL1r tagger (taken from Ref. [340]).

The role of machine learning algorithms to discriminate signal from background is also essential. A 30% improvement is obtained by using BDTs instead of a cut-based analysis for the $VH, H \to b\bar{b}$ resolved and boosted topology, and $VH, H \to c\bar{c}$ as for instance shown in Figure 6.14 of the next chapter. More sophisticated and more performant machine learning classifiers, compared to boosted decisions trees, such as graph neural network or deep neural network exist and could also increase the sensitivity to signal thanks to a better discrimination of signal from background processes. The main limitation for the $VH, H \to b\bar{b}/c\bar{c}$ analysis in that case would be finding a trade-off between the gain brought by those more efficient ML classifiers and the computational time lost. GNN and DNN are more complex algorithms than BDTs and thus their use is more

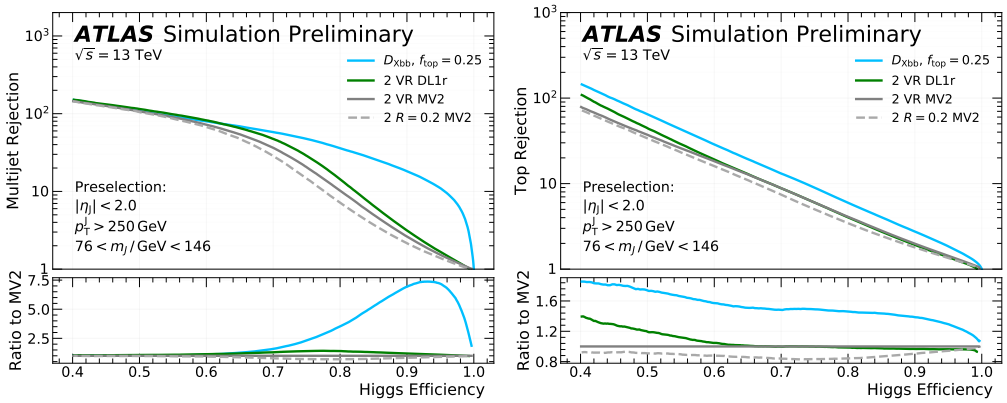


Figure 5.55: (a) Multijet and (b) top jet rejection as a function of the $H \to b\bar{b}$ tagging efficiency, for large-R jets with $p_{\mathrm{T}} > 250$ GeV. Performance of the Xbb algorithm are compared to the DL1r and to two variants of the MV2 tagger, one evaluated on variable-radius (VR, $\rho = 30$ GeV) track-jets, the other on fixed-radius (FR, $R = 0.2$) track-jets. The efficiency and rejection are calculated with respect to jets that have already passed the p_{T}, η, and mass preselection requirements reported in the legend. The bottom panels show the ratio of the performances with respect to the MV2 tagger (taken from Ref. [212]).

time-consuming. For instance, even though it is not a signal-background classifier, the GNN truth tagging algorithm described in Section 5.5 was found to be a factor 2 times slower than direct tagging, and this was a significant issue for our analysis. It was one of the main reason for the implementation of the submission of jobs to the grid that I carried out for the group.

Improvement of the theoretical and modelling uncertainties are also a must for ATLAS and its $VH(b\bar{b}/c\bar{c})$ group. As explained for instance in Section 4.5.3 of the chapter about the b-JES measurement, some comparison between event generators, that are used for the modelling uncertainties, are known to be overestimated (*e.g.* the matrix element uncertainty). More refined and realistic comparison are available (despite not being recommended for Run 2 analyses at the moment) and could in the future reduce associated uncertainties. Moreover, a recent ATLAS study [341] found that the observed difference in the hadronic jet response, *i.e.* the ratio of the reconstructed jet energy to the true jet energy, which varies by about 1-2% depending on the hadronization model, strongly depends on baryon (and kaon) energy fractions and their production rate for the generators. With a careful tuning of those parameters (based on experimental measurements) the discrepancies between different event generators could be reduced: it would induce a reduction of modelling uncertainties for physics analyses. Finally, for the Run 3 ATLAS will make significantly larger use of fast simulations. This could enable to generate alternative samples with more events and hence would reduce statistical fluctuations when estimating modelling uncertainties.

Improvement on the jet calibration and reconstruction algorithms could potentially contribute to an increase of the sensitivity of the analysis. For instance, for small-R jets, the switch from a EMTopo to PFlow reconstruction algorithm was found to slightly improve the m_{bb} invariant mass distribution for $VH, H \to b\bar{b}$ signal events by 1 to 3%, m_{bb} being the most discriminating variable used by the BDT. For large-R jets, a similar to PFlow reconstruction algorithm, referred to as Unified Flow Objects (UFO), is available. It exploits both tracker-based and calorimeter-based information contrarily to the LCTopo (default) reconstruction algorithm which is only based on calorimeter measurements. The UFO algorithm could not be used for the Legacy Run 2 publication because the calibration timeline was not compatible with the analysis. However, the group studied the potential improvements that could be expected by using such UFO large-R jets in the $VH, H \to b\bar{b}$ boosted 0-lepton channel. It was found that it increases by 20% the signal yield (mostly more events pass the $p_{\mathrm{T}}^{J} > 250$ GeV cut for UFO jets), a better JMS and JES but a worse JER with respect to the default LCTopo large-R jets were obtained. Also, ATLAS is developing machine learning algorithms to perform calibration of jets or of the energy clusters composing the jets

which should improve for example pile-up corrections, pile-up dependence and should reduce the energy resolution hence allowing for more precise measurements.

5.11.2 High luminosity phase of the LHC

The high luminosity phase of the LHC (HL-LHC) will occur after the Run 3 and the long-shut down of the LHC which will enable to upgrade the detectors. This high luminosity phase is scheduled to last from 2028 to 2040 and will accumulate 3000 fb^{-1} of data from pp collisions at a center-of-mass-energy $\sqrt{s} = 14\,\text{TeV}$ *i.e.* around 10 times more data will be collected than for the Run 3. The instantaneous luminosity will be a factor 5 to 7 times larger than the LHC nominal value hence requiring several parts of the detectors to be changed and replaced with materials able to sustain the increased radiation rate conditions. The granularity of the detectors will also be increased to overcome the large increase of pile-up collisions. During this phase the number of inelastic collisions per bunch crossing is expected to be ranging between 100 and 200.

Based on the $VH, H \to b\bar{b}/c\bar{c}$ measurements already performed for Run 2 in particular the combination of the resolved only $VH, H \to b\bar{b}$ topology [65] and the $VH, H \to b\bar{b}$ and $VH, H \to c\bar{c}$ resolved combination [6], results of the $VH, H \to b\bar{b}/c\bar{c}$ analysis can be extrapolated to the HL-LHC phase. The experimental and theory uncertainties are included in the extrapolation, but the size of the uncertainties are scaled to account for reductions in their statistical components and potential improvements in the analysis techniques associated with the larger available dataset. The different rescaling of uncertainties are reported in Table 5.12. The extrapolated results for the HL-LHC can be compared to the ones obtained for Run 2 in Section 5.9. Flavour-tagging improvements are assumed for the HL-LHC scenario.

Uncertainties	Scale Factor
$E_{\text{T}}^{\text{miss}}$	0.5
Lepton	1
Jet	1
Flavour tagging c-, b- and τ-jets	0.5
Flavour tagging light-jets (MV2c10 in $VH(bb)$)	0.5
Flavour tagging light-jets (DL1 in $VH(cc)$)	1.0
Luminosity	0.58
Signal modelling	0.5
Background modelling	0.5
MC statistics	0
Truth-tagging uncertainties ($VH, H \to c\bar{c}$ only)	0

Table 5.12: Scale factors applied to systematic uncertainties in the extrapolation. The luminosity uncertainty is reduced from 1.7% to 1.0%, resulting in a scale factor of 0.58 (taken from Ref. [339]).

The signal strengths, of the WH and ZH production modes, extrapolated to the HL-LHC for the $VH, H \to b\bar{b}$ resolved measurement are found to be [339]:

$$\mu_{WH}^{H\to b\bar{b}} = 1.00 \pm 0.08 = 1.00 \pm 0.04\,(\text{stat.}) \pm 0.07\,(\text{syst.}),$$
$$\mu_{ZH}^{H\to b\bar{b}} = 1.00 \pm 0.07 = 1.00 \pm 0.03\,(\text{stat.}) \pm 0.06\,(\text{syst.}),$$

to be compared with the Run 2 measurement [65]

$$\mu_{WH}^{H\to b\bar{b}} = 0.95^{+0.27}_{-0.25} = 0.95 \pm 0.18\,(\text{stat.}) \pm^{+0.19}_{-0.18}\,(\text{syst.}),$$
$$\mu_{ZH}^{H\to b\bar{b}} = 1.08^{+0.25}_{-0.24} = 1.08 \pm 0.17\,(\text{stat.}) \pm^{+0.18}_{-0.15}\,(\text{syst.}).$$

The associated STXS signal strength measurement is provided in Figure 5.56.

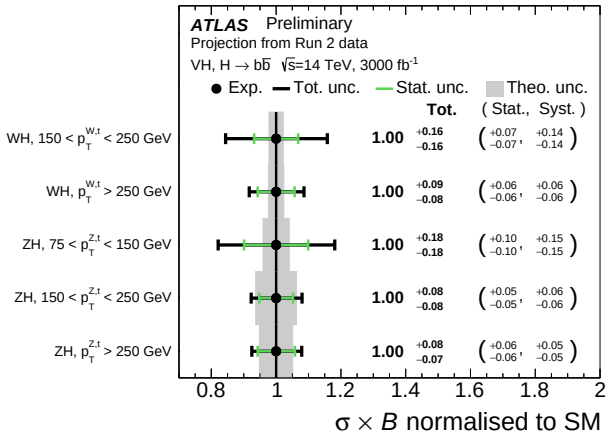


Figure 5.56: Signal strength for the resolved $VH, H \to b\bar{b}$ signal in the reduced 1.2 STXS scheme as extrapolated to the HL-LHC conditions. The theory uncertainties are shown as a grey band and reduced by a factor 1/2 with respect to the Run 2 analysis (taken from Ref. [339]).

For the $VH, H \to c\bar{c}$ measurement, the extrapolation for the inclusive signal strength is:

$$\mu_{VH}^{H\to c\bar{c}} = 1.0 \pm 3.2 = 1.0 \pm 2.0\,(\text{stat.}) \pm {}^{+2.6}_{-2.5}\,(\text{syst.}),$$

to be compared with the Run 2 inclusive signal strength [6]

$$\mu_{VH}^{H\to c\bar{c}} = -9 \pm 16 = -9 \pm 10\,(\text{stat.}) \pm 12\,(\text{syst.}).$$

The expected upper limit on $\mu_{VH}^{H\to c\bar{c}}$ is equal to 6.4 times the SM prediction at a 95% CL for the HL-LHC, as shown in Figure 5.57, against 31 for the Run 2. This HL-LHC extrapolation is comparable to the CMS expected limit (7.6 and observed 14.4) for the Run 2 as shown in Figure 5.37. When

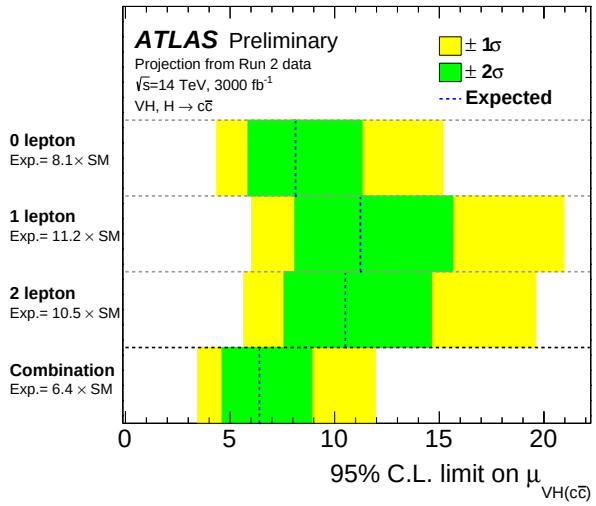


Figure 5.57: $VH, H \to c\bar{c}$ signal strength limits in the different lepton channels from a fit to an Asimov dataset extrapolated to the HL-LHC conditions. The per-channel results are obtained using a 5-POI fit, in which each channel has a separate $VH, H \to c\bar{c}$ parameter of interest (taken from Ref. [339]).

reinterpreted with the coupling strength modifier it leads to an expected constraint of $|\kappa_c| < 3.0$ with a 95% CL while for Run 2 the expected constraints was $|\kappa_c| < 12.4$.

Finally, the combination of the resolved $VH, H \to b\bar{b}$ and $VH, H \to c\bar{c}$ analyses is extrapolated and yields an expected constraint of $|\kappa_c/\kappa_b| < 2.7$ for HL-LHC against 5.1 expected (observed 4.5) for Run 2. The corresponding likelihood scan is reported in Figure 5.58 for the HL-LHC extrapolation which can be compared to the Figure 5.36a for the Run 2.

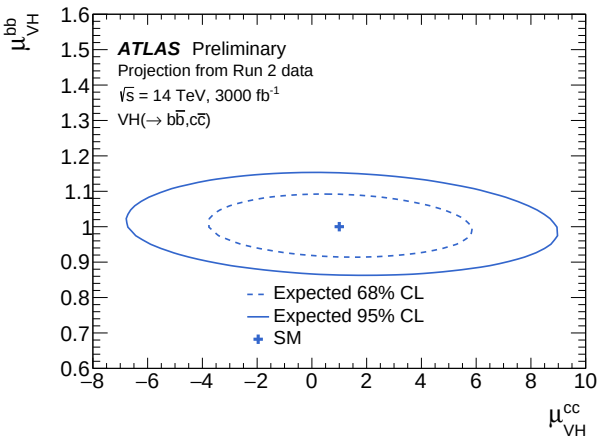


Figure 5.58: Expected profile likelihood scans for the $VH, H \to b\bar{b}/c\bar{c}$ combination extrapolated to HL-LHC conditions (taken from Ref. [339]).

5.11.3 Future Circular Collider

In a more distant future, the Future Circular Collider (FCC) is a project is a proposed project to build at CERN a circular accelerator of 100 km of circumference to reach higher energies (of the order of 100 TeV) and higher integrated luminosities for hadron collisions for instance FCC-hh (hadron-hadron collisions) or be able to collide electrons-positrons for example with the FCC-ee project, or mixed collisions hadrons-electrons FCC-he. This circular collider would follow the HL-LHC phase after 2040 and take over the LHC as main accelerator of the CERN. The large-radius of the FCC is a key feature as collisions with electrons could not be performed with the LHC (or higher energies could not be reached for hadrons at the LHC) because of the large energy loss due to synchrotron radiation as reported in Equation (2.1.6). Ions collisions are also foreseen.

In this section the FCC-ee project from a point of view of VH studies [342] is briefly described. One of the main advantages of lepton colliders with respect to hadron colliders is to not have to deal with the composite nature of hadrons. Lepton colliders allow to fully exploit the conservation of momentum before and after the collision in 3-dimensions *i.e.* not only in the transverse plane but also along the longitudinal axis contrarily to hadron colliders where the momentum fraction carried out by partons inside hadrons is unknown which prevents using conservation of momentum along the beam axis. It would also allow reducing significantly the hadronic QCD pileup which is a limitation for several LHC analyses and is not well modelled in simulations.

For the FCC-ee the main VH production would be ZH events whose leading Feynman diagram is shown in Figure 5.59. For this project e^+e^- collision would occur with a center-of-mass-energy $\sqrt{s} = 240$ GeV. Targeting the $ZH, Z \to \mu^+\mu^-$ decays would lead to the distributions presented in Figure 5.60, where the Higgs recoil mass is easily distinguishable allowing for precision measurement thanks to the high signal to background purity. The recoil mass (m_{recoil}) being defined as the particle/system recoiling against the lepton pair and is determined from the total energy-momentum conservation as

$$m^2_{\text{recoil}} = s + m^2_{ll} - 2\sqrt{s}(E_{l^+} + E_{l^-}). \tag{5.11.1}$$

Determining the recoil mass allows for an inclusive ZH cross-section measurement by only focusing on the $Z \to \mu^+\mu^-$ decays and not based on any selection concerning the recoil system which is here the Higgs boson. This measurement would be expected to achieve a statistical precision of around 1% on the ZH inclusive cross-section and would determine the Higgs mass with a precision of 6 MeV *i.e.* with an increased precision by a factor 20 compared to the current most precise measurement achieved at the LHC. Finally, in case of $ZH, H \to b\bar{b}$ measurement this would be much easier than at the LHC as the hadronic background such as multijet production,

$t\bar{t}$ pairs would be suppressed. As a result a relative precision of respectively 0.3% and 2.2% for $\sigma_{ZH} \times \mathcal{B}_{H\to b\bar{b}}$ and $\sigma_{ZH} \times \mathcal{B}_{H\to c\bar{c}}$ would be expected for the FCC-ee project [342]: lepton colliders truly allow for precision measurements.

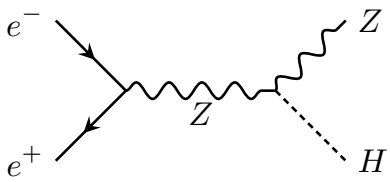
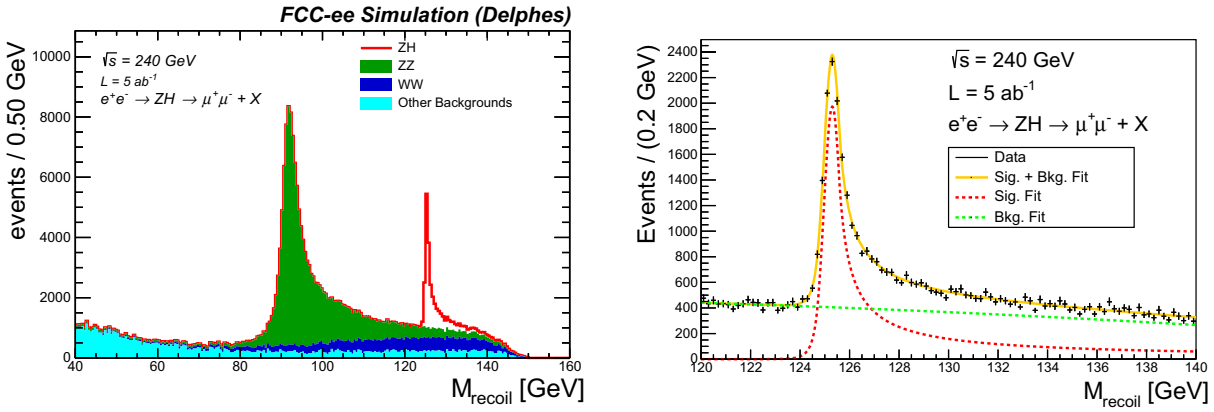


Figure 5.59: Leading Feynman diagram of the ZH production mode for the FCC-ee project ($e^+e^- \to ZH$).



(a) (b)

Figure 5.60: (a) Inclusive m_{recoil} distribution selecting events with a $Z \to \mu^+\mu^-$ decay signature requiring the invariant mass of the di-muon system to be within $40 < m_{\mu^+\mu^-} < 160$ GeV.(b) m_{recoil} mass distribution in the region around m_H. The ZH signal is fitted with a double-sided Crystal Ball function [343], and the simulated background with a second-order polynomia (taken from Ref. [342]).

5.12 Conclusion

Following the JMS study that I performed, it was decided not to apply that calibration to small-R jets since it was observed that the JMS was shifting the invariant mass of jets towards lower values and was increasing the level of disagreement with respect to the truth level predictions.

Thanks in particular to the significant improvement of the b/c-tagging performances because of the new DL1r tagger, and also thanks to the extension of the MVA approach to the $VH, H \to b\bar{b}$ boosted and $VH, H \to c\bar{c}$ analyses, an increased sensitivity is expected for the $VH, H \to b\bar{b}/c\bar{c}$ Legacy Run 2 publication in particular to the $H \to c\bar{c}$ decay mode where the 95% CL upper limit on the signal strength is expected to be reduced by a factor 2 with respect to the previous round of the $VH, H \to c\bar{c}$ analysis. If the $VH, H \to c\bar{c}$ expected limit is observed with a same level of precision, ATLAS would almost compete with the currently most precise observed limit by CMS which determined that the $H \to c\bar{c}$ signal strength is $\mu_{VH}^{H\to c\bar{c}} < 14.4$. For the $VH, H \to b\bar{b}$ analysis, a gain of 20% in sensitivity is expected: the expected precision on the signal strength measurement being equal to $\mu_{VH}^{H\to b\bar{b}} = 1.00^{+0.15}_{-0.14}$. Moreover, a higher STXS granularity introducing the $p_{\text{T}}^V > 600$ GeV region and planning to split the STXS bins as a function of the number of additional jets for $p_{\text{T}}^V < 400$ GeV is foreseen for the STXS measurement. The combination of the resolved and boosted $VH, H \to b\bar{b}$ topologies has also been proven to be of high interest as more

stringent constraints are obtained and degeneracies are lifted for some of the Wilson coefficients, compared to the individual resolved only or boosted only analyses, when the STXS measurements are reinterpreted in the context of an EFT. The ultimate milestone of the $VH, H \to b\bar{b}/c\bar{c}$ analyses for ATLAS and CMS would be now to reach the sensitivity required to measure the Higgs-to-charm coupling which at the moment is thought to be impossible at the LHC with the current state-of-the-art performances and their projection in the future. However, the recent outstanding measurement from CMS which achieved a sensitivity only thought to be reachable at the end of the HL-LHC phase is a game changer. It shows the importance of the flavour tagging for Higgs to heavy flavour decays studies and more generally of the crucial role played by machine learning algorithms in particle physics. The next generation of b- and c-taggers developed are already outperforming the current default ATLAS taggers which is promising for the future. In the same way the $VH, H \to b\bar{b}/c\bar{c}$ analysis could benefit from more complex algorithm than boosted decision trees. Efforts from the detector, the experimental, the theoretical sides and an increase of the collected dataset could largely improve such measurement as the statistical uncertainty is still one of the leading source of uncertainties and systematic uncertainties could still be reduced. Finally, in a more distant future, the FCC project *e.g.* FCC-ee could pave the way towards high precision measurements of the inclusive ZH production mode and dedicated $ZH, H \to b\bar{b}/c\bar{c}$ processes by offering the possibility to collect a large amount of data, by largely reducing the QCD background compared to that of hadron colliders as QCD is to date not well modelled in the low energy limit, and the QCD background experimentally troublesome and challenging for several Higgs analyses.

Chapter 6

MVA studies for the $VH, H \to b\bar{b}$ analysis

6.1 $VH, H \to b\bar{b}$ resolved MVA study in the 0-lepton channel . . . 256
6.1.1 Training, evaluation and input variables . . . 256
6.1.2 Hyperparameters . . . 260
6.1.3 Overtraining checks . . . 260
6.1.4 BDT performances . . . 263
6.1.5 Conclusion of the MVA training . . . 266
6.2 Combination study of the resolved and boosted topologies in the 1-lepton channel . 266
6.2.1 Event selection and combination strategies . . . 266
6.2.1.1 Combination strategies . . . 266
6.2.1.2 Multivariate discriminants . . . 268
6.2.2 Significances and transformation D . . . 268
6.2.3 Combination results . . . 269
6.2.3.1 Results for the different strategies . . . 269
6.2.3.2 Impact of the transformation D parameters . . . 277
6.2.3.3 Impact of the b-tagging strategy for the boosted topology . . . 280
6.2.4 Conclusion of the combination study . . . 287
6.3 Conclusion . . . 287

For the $VH, H \to b\bar{b}/c\bar{c}$ analysis, multivariate analysis techniques play a leading role since it enhances signal from background discrimination. Moreover, the sensitivity to signal is relying on the boosted decision trees whose performances are crucial to optimize.

I have been deeply involved in such optimizations efforts in two aspects. First I have trained and optimized the 0-lepton $VH, H \to b\bar{b}$ BDT for the resolved topology, study which is presented in Section 6.1. This BDT training is going to be used for the Legacy Run 2 publication and has already been intensively used by the group for many studies. I have also been investigating possible improvements of the combination of the $VH, H \to b\bar{b}$ resolved and boosted analyses probing both the performances of the boosted decisions trees trained for those two regimes and in order to gain an insight about their discrimination power. In particular this was of high importance since the boosted $VH, H \to b\bar{b}$ analysis was for the first time using an MVA approach as before the mass of the leading large-R jet was used instead. Moreover, the b-tagging algorithm has been updated both for the resolved and boosted topologies and come with significant improvement in terms of c and light-jet rejections as detailed in Section 5.3.3 and 5.3.4. The b-tagging strategy for the boosted topology was also re-optimized to increase the signal acceptance and sensitivity as explained in Section 5.3.4 adopting the 85% WP leading 3 strategy compared to the 70% leading 2 strategy previously. Following that second study results, the ATLAS $VH(b\bar{b}/c\bar{c})$ analysis team has been able to decide the strategy to adopt for the Legacy Run 2 publication in order to combine the two regimes. This second study is described in Section 6.2. Finally, conclusions are provided in Section 6.3.

The MVA studies performed for the $VH, H \to b\bar{b}/c\bar{c}$ analysis are being documented in an internal ATLAS note, restricted to the members of the ATLAS collaboration, of which I am one of the three main editors. This note is a major update with respect to the previous round since the MVA is now used in all regions of the $VH, H \to b\bar{b}/c\bar{c}$ analysis.

6.1 $VH, H \to b\bar{b}$ resolved MVA study in the 0-lepton channel

This section describes the 0-lepton MVA training for the $VH, H \to b\bar{b}$ resolved topology. The general principles of multivariate analysis algorithm and important notions about boosted decisions trees have been introduced in Section 5.6. Section 6.1.1 presents the BDT training setup and how the training performances are evaluated. Section 6.1.2 provides a description of the BDT hyperparameters. Overtraining checks are shown in the Section 6.1.3. Section 6.1.4 focuses on the performances of the BDTs with their optimal hyperparameters. Finally, the Section 6.1.5 summarizes the results obtained.

6.1.1 Training, evaluation and input variables

For the 0-lepton channel $VH, H \to b\bar{b}$ resolved MVA, the nominal samples, presented in Section 5.2, are used for the training and for the evaluation of performances of boosted decision trees. The signal samples are composed of all $VH, H \to b\bar{b}$ processes while background events consist in all $t\bar{t}$, single top, $W + \text{jets}$, $Z + \text{jets}$ and diboson samples. Moreover, the list of input variables used for the MVA training is provided in Table 5.6. Those input variables selected have all been found to improve the MVA performances.

The BDT training is split with respect to the jet multiplicity resulting in a total of 3 BDTs corresponding to the 2-, 3- and 4-jets categories. The 4-jets BDT training was never performed before, previous rounds of the resolved analysis were only using the 2- and 3-jets regions. The training is performed inclusively over the full p_{T}^{V} spectrum *i.e.* $p_{\text{T}}^{V} > 150$ GeV. The resolved events with $p_{\text{T}}^{V} > 400$ GeV are hence taken into account in the training though the resolved analysis for the legacy publication will be limited to the $150 < p_{\text{T}}^{V} < 400$ GeV region. This was in particular useful for the combination strategy study presented (for the 1-lepton) in Section 6.2. The BDTs are trained over all events from the signal and control regions that are passing the full resolved event selection described in Section 5.3 and which is summarized in Table 5.3 including the b-tagging requirements *i.e.* only the events with exactly 2 b-tagged jets are considered. On the other hand, for the BDTs performance evaluations only events from signal regions are selected. Events from control regions are considered because in the 0-lepton channel it was decided to merge the low δR CR with the SR. This SR+CRs training also allows for more flexibility as for instance the ΔR-p_{T}^{V} continuous side-bands, described in Section 5.3.3, have been updated for the Legacy Run 2 publication and so retraining dedicated MVA was not required since the training is inclusive.

For many input variables such as for instance the dijet invariant mass m_{bb}, the associated background distributions have long upper tails (*e.g.* high m_{bb} values) compared to the signal distributions. To prevent the BDTs from exploring selections on regions of phase space that are almost exclusively populated by background events, upper threshold values are defined for each input variables concerned by long tails in order to contain 99% of the signal events. Any event with one input variable above one of those thresholds is simply ignored for the training part.

The Figure 6.1 shows the distributions of the full set of the BDT input variables for the signal and background events for the 2 b-tagged 2-jets category while Figures 6.2 and 6.3 are presenting respectively few input variables distributions for the 3- and 4-jets categories.

The correlations between input variables for signal and background events can be found in Fig-

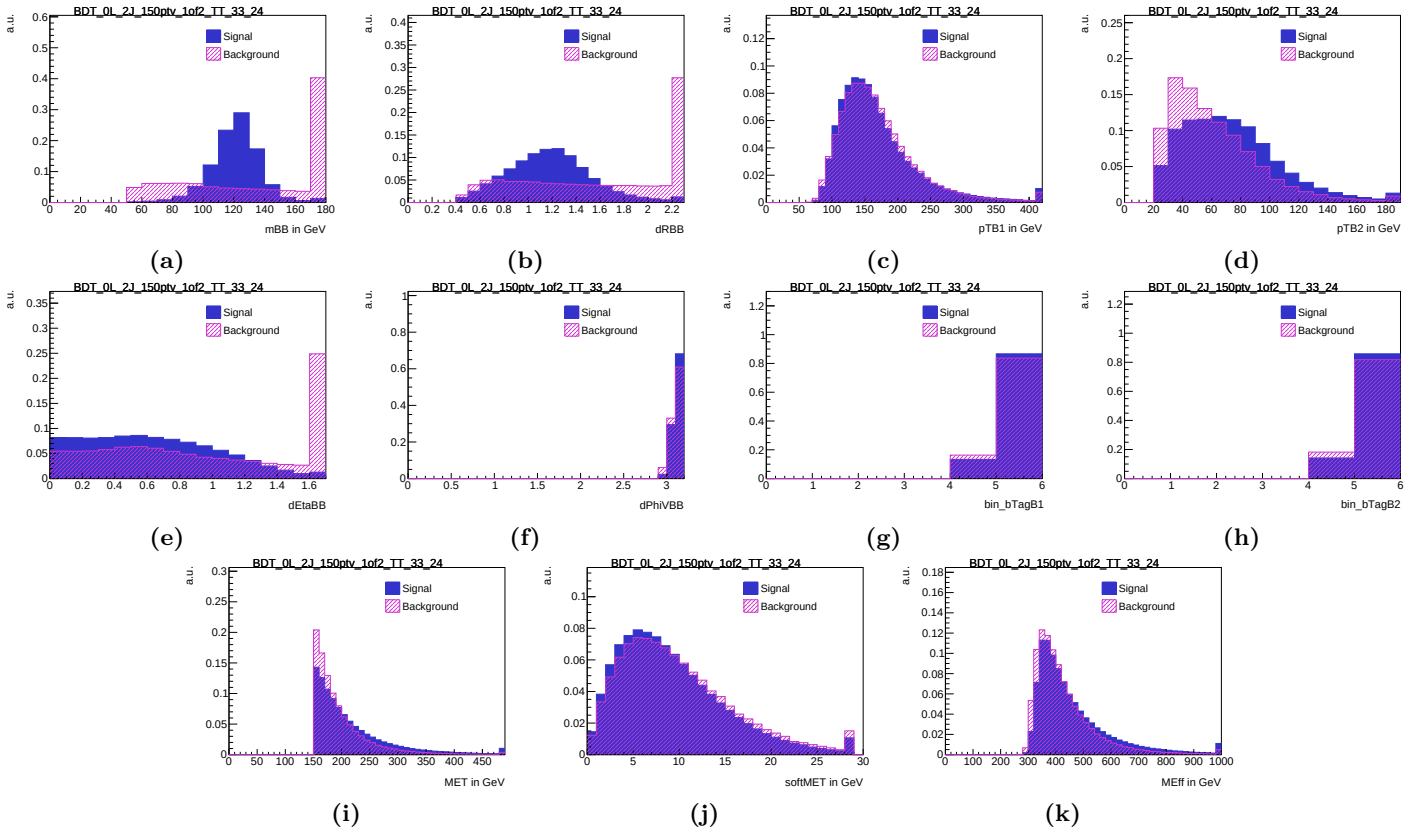


Figure 6.1: Normalized BDT input variable distributions for signal (solid blue) and background (hatched red) in the 2 b-tagged 2-jets region as predicted by the nominal simulation using truth tagging. The upper bin is filled with the overflow content. (a) m_{bb}, (b) $\Delta R(b,b)$, p_{T} of the (c) leading b-tagged and (d) subleading b-tagged jets, (e) $|\Delta\eta(b_1,b_2)|$, (f) $|\Delta\phi(\boldsymbol{V},\boldsymbol{H}_{\mathbf{cand}})|$, DL1r bin for the (g) leading b-tagged (h) subleading b-tagged jets, (i) $E_{\mathrm{T}}^{\mathrm{miss}}$, (j) $p_{\mathrm{T}}^{\mathrm{miss,st}}$ and (k) m_{eff}.

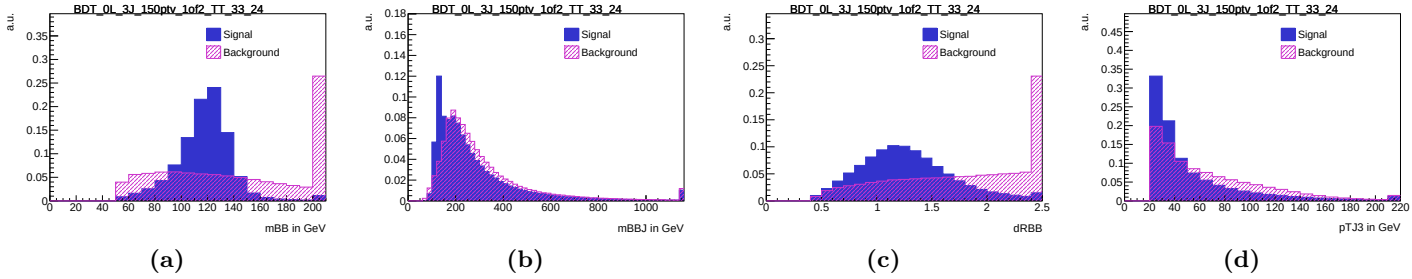


Figure 6.2: Few normalized BDT input variable distributions for signal (solid blue) and background (hatched red) in the 2 b-tagged 3-jets region as predicted by the nominal simulation using truth tagging. The upper bin is filled with the overflow content. (a) m_{bb}, (b) m_{bbj}, (c) $\Delta R(b,b)$, (d) p_{T} of the untagged jet.

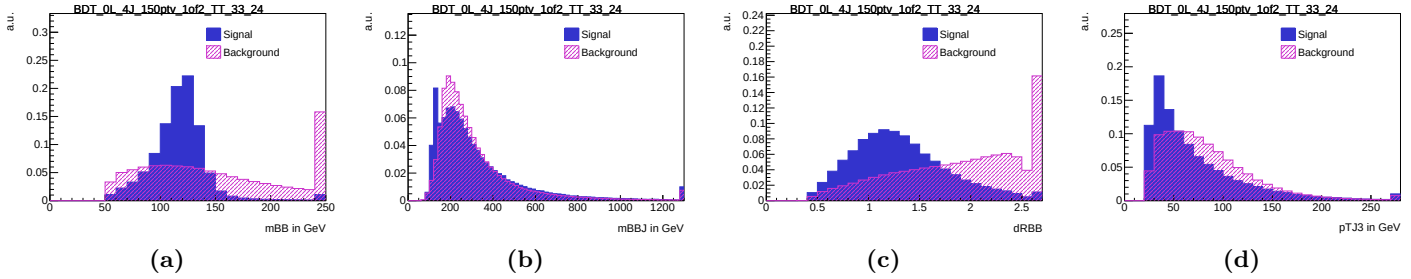
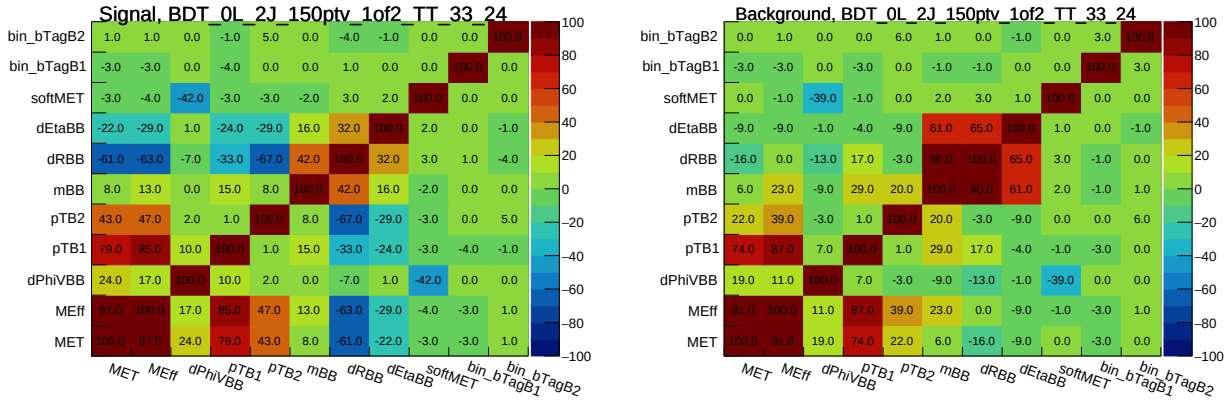


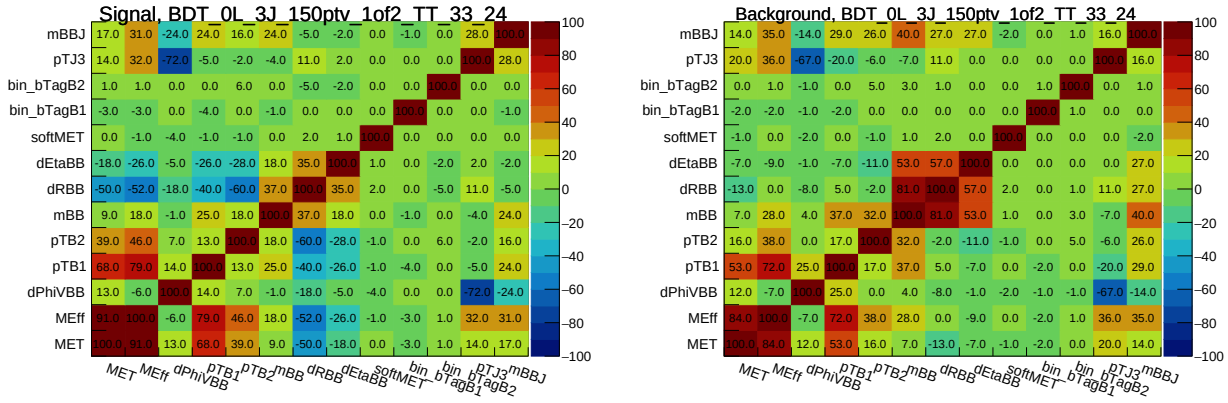
Figure 6.3: Few normalized BDT input variable distributions for signal (solid blue) and background (hatched red) in the 2 b-tagged 4-jets region as predicted by the nominal simulation using truth tagging. The upper bin is filled with the overflow content. (a) m_{bb}, (b) m_{bbj}, (c) $\Delta R(b,b)$, (d) p_{T} of the leading untagged jet.

ures 6.4, 6.5 and 6.6 respectively for the 2-, 3- and 4-jets categories and are in general similar in those three regions of phase space. For instance for the 2-jets region, the correlations between m_{bb} and $\Delta R(b,b)$ is of 42% for signal because of the link between the Higgs mass, the p_{T} of the Higgs and the angular distance between the b-hadrons for a $H \to b\bar{b}$ decay (Equation (5.1.1)) while for the background the correlation reaches 90%. The corresponding 2-dimensional distribution of $\Delta R(b,b)$ displayed as a function of m_{bb} can be found in Figure 6.7. Another example, is the correlation between the $E_{\mathrm{T}}^{\mathrm{miss}}$ and $\Delta R(b,b)$ which is of -61% for signal and only -16% for background again due to kinematic reasons as illustrated in Figure 6.8.



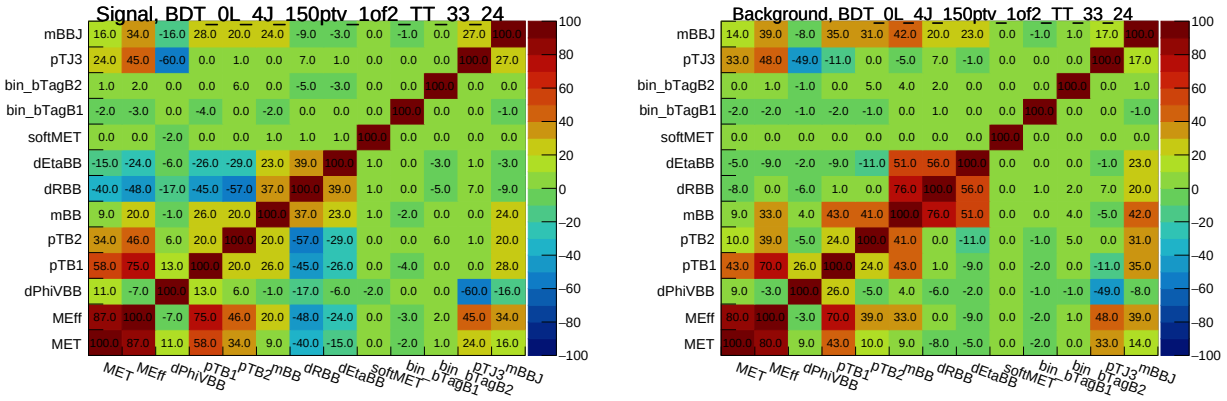
(a) **(b)**

Figure 6.4: BDT input variables correlation matrices for the (a) signal and (b) background events in the 2 b-tagged 2-jets region as predicted by the nominal simulation using truth tagging.



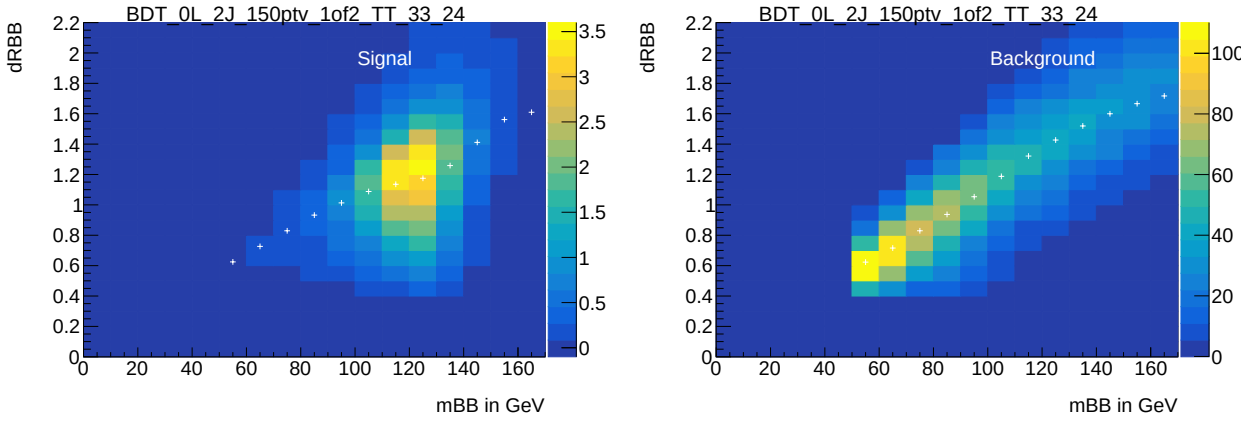
(a) **(b)**

Figure 6.5: BDT input variables correlation matrices for the (a) signal and (b) background events in the 2 b-tagged 3-jets region as predicted by the nominal simulation using truth tagging.



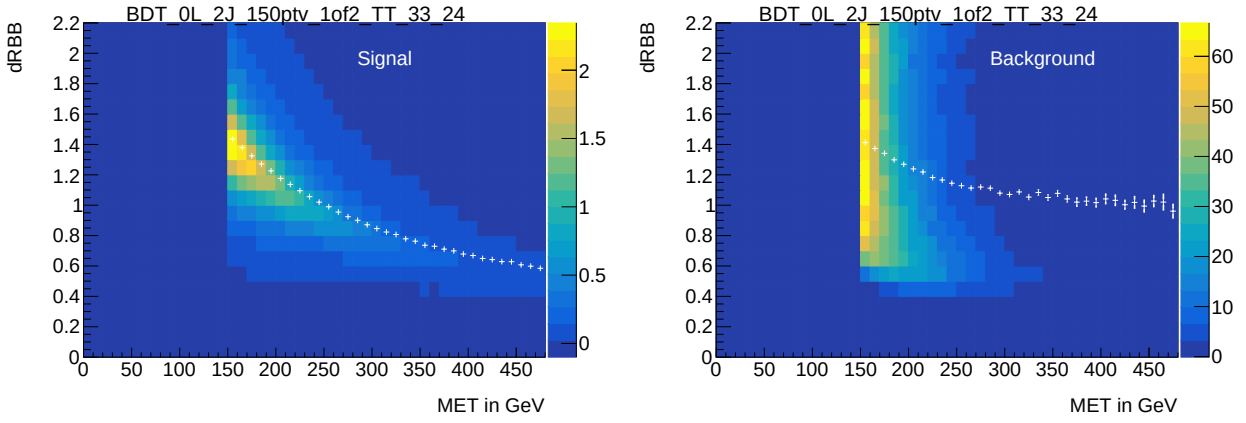
(a) (b)

Figure 6.6: BDT input variables correlation matrices for (a) signal and (b) background events in the 2 b-tagged 4-jets region as predicted by the nominal simulation using truth tagging.



(a) (b)

Figure 6.7: $\Delta R(b,b)$ as a function of m_{bb} for the (a) signal and (b) background events in the 2 b-tagged 2-jets region as predicted by the nominal simulation using truth tagging.



(a) (b)

Figure 6.8: $\Delta R(b,b)$ as a function of $E_{\mathrm{T}}^{\mathrm{miss}}$ for the (a) signal and (b) background events in the 2 b-tagged 2-jets region as predicted by the nominal simulation using truth tagging.

6.1.2 Hyperparameters

The hyperparameters of the BDTs have been optimized per jet category to reduce overtraining and maximize the Asimov significance (presented in Section 5.7.6) after having applied Transformation D to the BDT output distributions (described in Section 5.7.7). For the optimization of hyperparameters, truth tagging was used in order to increase the amount of simulated events available for the training in particular to enrich the training samples in light- and c-component background events. In the near future, a new BDT training using GNN truth tagging instead of truth tagging will be performed in order to combine the advantage of having an increase of statistics with respect to direct tagging and in order to reduce mismodelling that is caused by the use of truth tagging as explained in Section 5.5.

The optimal parameters below were found to be the same for all jet categories. Some of technical terms related to BDTs are discussed in Section 5.6.

- The gradient boost algorithm is used.
- The number of trees (also referred to as number of weak learners) is set to 200.
- The learning rate is equal to $\beta = 0.5$
- The maximum cell tree depth is set to 3. It is the maximal number of nodes *i.e.* successive cuts that each decision tree is allowed to perform in order to further split phase space into signal and background dominated subregions.
- The number of cuts is set to 100. It corresponds to the number of divisions (also called the number of grid points) used when scanning over a variable range to find an optimal cut for a node splitting. The finer the granularity is the better should be the cut at the expense of computation time.
- The minimum node size is fixed to 5%. It is the minimum percentage of training events required in a leaf node. It avoids having parts of the phase space that are too finely split by the BDT which could result in potential overtraining due to low statistic.
- To optimize the decision cuts the *Gini index* method is used

$$\text{Gini} = p(1-p), \tag{6.1.1}$$

  with p the purity of a node: $p = S/(S+B)$, S and B being respectively the weighted sum of all signal and background events contained in that node during the training. The Gini index is maximal when the node contains the same number of signal and background events (*i.e.* $p = 0.5$). For each node, the phase space is divided in 2 child nodes by a cut. The maximum separation is defined as the maximum change in the Gini index between the parent node and the two child nodes:

$$\Delta\text{Gini} = \text{Gini}_{\text{parent}} - \text{Gini}_{\text{child}_1} - \text{Gini}_{\text{child}_2}. \tag{6.1.2}$$

  The cut corresponding to the highest ΔGini is applied, and the iteration continues for the daughter nodes. Once the maximal number of cuts is reached, depending on the purity of the node, all events reaching the node will all be either classified as signal or as background events.
- No pruning is applied. For BDTs with a large maximal depth, the pruning consists in removing statistically insignificant branches. Since the maximal depth for the 0-lepton training remains small there is no reason to use a pruning method.

6.1.3 Overtraining checks

The overtraining for the optimized parameters has been checked for each jet categories. A comparison of the ROC curves and ROC AUC obtained when evaluating the BDT on the training and test

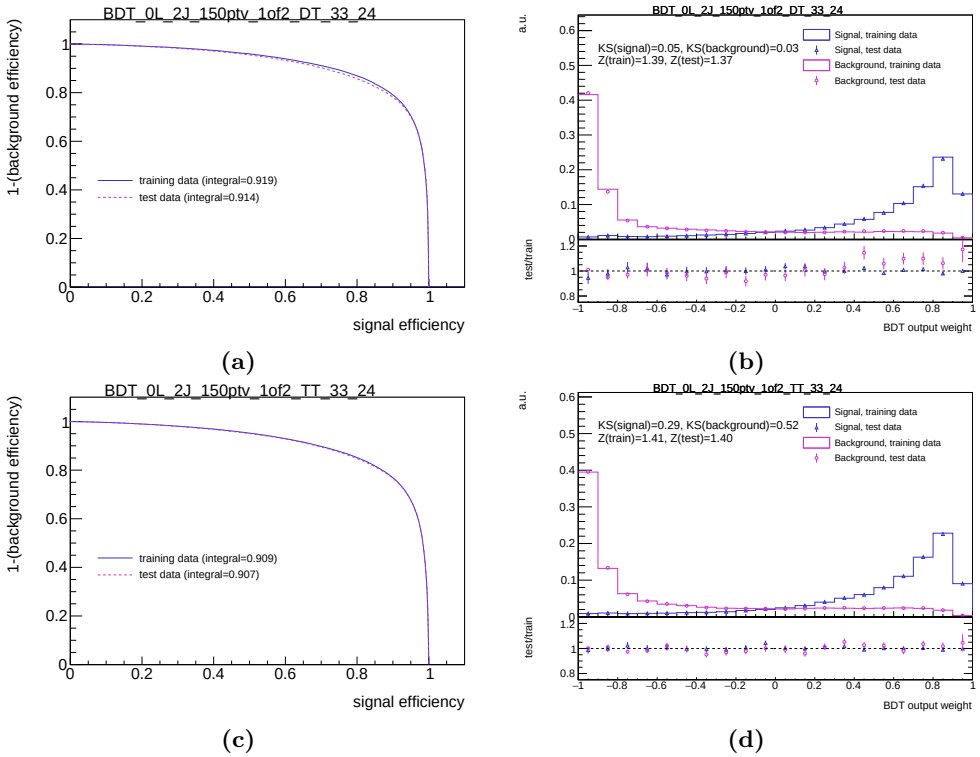


Figure 6.9: Overtraining checks for the 2 b-tagged 2-jet category for BDT trainings performed with (a) and (b) direct tagging, and (c) and (d) truth tagging inputs. (a) and (c) ROC curves (background inefficiency as a function of the signal efficiency) obtained evaluating the BDT on the training and test samples. The ROC AUC are specified inside the parenthesis. (b) and (d) BDT output distributions for signal and background obtained for the training (line) and test (dots) samples. The bottom panel shows the ratio of the test over training BDT output distributions for signal and background.

samples as well as a comparison of the BDT output distribution for signal and background for the test and training samples are shown in Figures 6.9, 6.10 and 6.11 respectively for the 2-, 3- and 4-jets categories. For the ROC curves, differences in ROC AUC between direct and truth tagging BDTs are "meaningless" because the ROC AUC are respectively computed with direct or truth tagging samples. It has been checked that when evaluated on the same direct tagging samples, the BDTs trained on truth tagging inputs achieve similar performances as the direct tagging ones in terms of ROC AUC and significances. The comparison of the ratio of the test and training BDT output distributions shows that the BDTs benefit from the increase in statistics of the truth tagging inputs as reduced discrepancies are observed in high and low BDT score regions compared to direct tagging ratios. For the 4-jets training, an overtraining is observed for low BDT scores only for signal events (Figure 6.11b) which is however improved as a larger overtraining was observed for direct tagging BDT training. Overall the overtraining remains limited to that low BDT region which anyway is not contributing to increase significance and the test and train ROC curves are pretty consistent. For the reasons mentioned above, the truth tagging inputs will be use for the MVA training of the Legacy Run 2 publication.

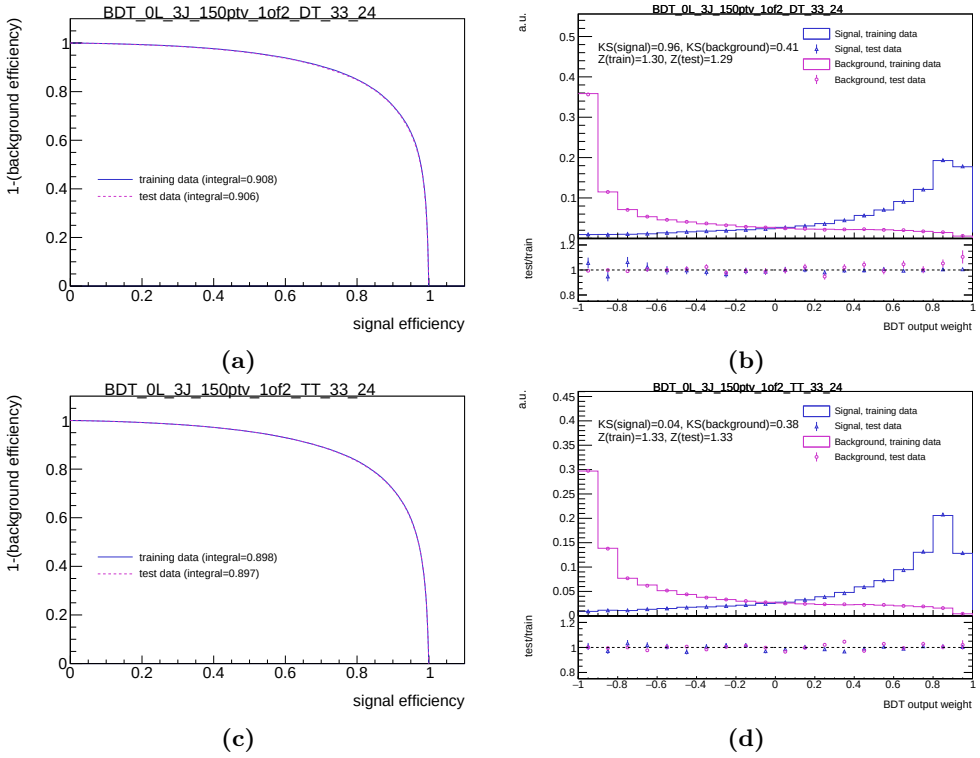


Figure 6.10: Overtraining checks for the 2 b-tagged 3-jet category for BDT trainings performed with (a) and (b) direct tagging, and (c) and (d) truth tagging inputs. (a) and (c) ROC curves (background inefficiency as a function of the signal efficiency) obtained evaluating the BDT on the training and test samples. The ROC AUC are specified inside the parenthesis. (b) and (d) BDT output distributions for signal and background obtained for the training (line) and test (dots) samples. The bottom panel shows the ratio of the test over training BDT output distributions for signal and background.

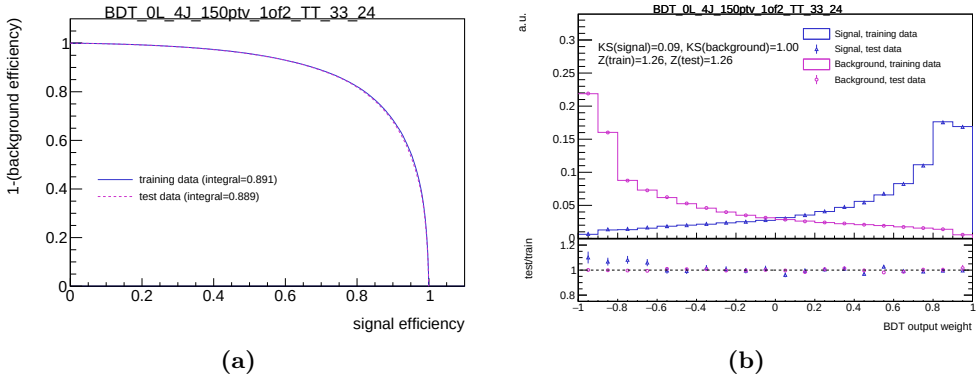


Figure 6.11: Overtraining checks for the 2 b-tagged 4-jet category for BDT trainings performed with truth tagging inputs. (a) ROC curves (background inefficiency as a function of the signal efficiency) obtained evaluating the BDT on the training and test samples. The ROC AUC are specified inside the parenthesis. (b) BDT output distributions for signal and background obtained for the training (line) and test (dots) samples. The bottom panel shows the ratio of the test over training BDT output distributions for signal and background.

6.1.4 BDT performances

The BDTs performances are assessed by computing the Asimov significance and maximizing the ROC AUC.

First a comparison of the BDTs performances when using the old ATLAS default recommendations (MV2c10 tagger and EMTopo jets) versus the new ones (DL1r tagger and PFlow jets) is presented in Tables 6.1 and 6.2. Two separate BDTs are trained per jet category corresponding to those two different setups and their performances are evaluated using the same samples either the MV2c10+EMTopo setup ones or the DL1r+PFlow setup ones. At that time, the strategy for the combination of the $VH, H \to b\bar{b}$ resolved and boosted analysis was not yet decided, so the evaluation is divided in two p_{T}^{V}-regions over the full p_{T}^{V}-spectrum: $p_{\mathrm{T}}^{V} \in [150, 250]$ GeV and $p_{\mathrm{T}}^{V} > 250$ GeV. It appears that the BDTs are robust with respect to the tagging and jet reconstruction algorithms as similar Asimov significances and ROC AUC are obtained for the two BDTs when evaluated on the same samples both in the 2-jets and 3-jets categories. Moreover, the increase of the Asimov significance by approximately 10% with the new recommendation setup was proven to be thanks to the use of the new DL1r tagger which comes with better light- and c-jet discriminations than the MV2c10 tagger. The correlations matrices between the input variables were also found, as expected, to be pretty similar both for signal and background events.

		Significance		ROC AUC	
		MV2+EMTopo	DL1r+PFlow	MV2+EMTopo	DL1r+PFlow
$p_{\mathrm{T}}^{V} \in [150, 250]$ GeV	2-jets	3.801	3.780	0.907	0.906
	3-jets	2.485	2.466	0.897	0.896
$p_{\mathrm{T}}^{V} > 250$ GeV	2-jets	3.287	3.243	0.931	0.930
	3-jets	2.125	2.093	0.920	0.919

Table 6.1: Comparison of the BDTs performances evaluated on the same samples corresponding to the MV2c10 tagger and EMTopo jets setup with direct tagging. The Asimov significance and ROC AUC are computed for the 2-jets and 3-jets SRs. The BDT trained with the MV2c10 tagger and the EMTopo jets (respectively DL1r tagger and the PFlow jets) sample is denoted MV2+EMTopo (respectively DL1r+PFlow).

		Significance		ROC AUC	
		MV2+EMTopo	DL1r+PFlow	MV2+EMTopo	DL1r+PFlow
$p_{\mathrm{T}}^{V} \in [150, 250]$ GeV	2-jets	4.108	4.100	0.911	0.911
	3-jets	2.777	2.784	0.901	0.902
$p_{\mathrm{T}}^{V} > 250$ GeV	2-jets	3.508	3.542	0.933	0.932
	3-jets	2.370	2.371	0.927	0.927

Table 6.2: Comparison of the BDTs performances evaluated on the same samples corresponding to the DL1r tagger and PFlow jets setup with direct tagging. The Asimov significance and ROC AUC are computed for the 2-jets and 3-jets SRs. The BDT trained with the MV2c10 tagger and the EMTopo jets (respectively DL1r tagger and the PFlow jets) sample is denoted MV2+EMTopo (respectively DL1r+PFlow).

An optimization of the BDTs' hyperparameters, concerning the number of trees, maximal depth and number of cuts, has been conducted and led to the choice of the hyperparameters provided in Section 6.1.2 which are achieving good performances and result in negligible overtraining. More "aggressive" parameters such as an increased number of cuts or maximal depth were not bringing significant improvements (increase of the significance smaller than 1%) but on the other hand were increasing the overtraining. The Table 6.3 summarizes the significances achieved with the latest BDT training (*i.e.* with the most up-to-date event selection and p_{T}^{V}-regions for the resolved analysis) with truth tagging input samples for the 2-, 3- and 4-jets categories when they are evaluated on

direct tagging input samples in order to avoid any potential mismodelling introduced by truth tagging. It is important to mention that the 4-jets category, whose BDT training has only been explored for the Legacy Run 2 publication, could add approximately a 2σ sensitivity (statistic significance only) according to the Table 6.3 in the 0-lepton channel and is currently under study to be included for fits.

The Figure 6.12 shows the ranking of the input variables ordered from the most to the least discriminating variable. As expected the invariant mass m_{bb} is the most discriminating quantity, followed by $\Delta R(b, b)$ (or m_{bbj} in the 3- and 4-jets categories) and then the transverse momenta of the second or third jet of the event. However, one has to keep in mind that this ranking is inclusive and could be different for some regions of the phase space.

		Significance	ROC AUC
$p_{\mathrm{T}}^{V} \in [150, 250]$ GeV	2-jets	4.067	0.840
	3-jets	2.726	0.850
	4-jets	1.456	0.854
$p_{\mathrm{T}}^{V} \in [250, 400]$ GeV	2-jets	3.180	0.863
	3-jets	2.167	0.861
	4-jets	1.239	0.856

Table 6.3: Latest BDTs performances (Asimov significance and ROC AUC) trained on truth tagging samples and evaluated on direct tagging samples in the 2-, 3- and 4-jets SRs with the DL1r tagger and PFlow jets.

In addition to the optimization of parameters, other potential discriminating variables have been investigated. Below is a list of some variables that have been tested:

- the pseudorapidity (η_{b_1}) of the leading b-tagged jet.
- The p_{T} asymmetry of the two b-tagged jets: either $\left|p_{\mathrm{T}}^{b_1} - p_{\mathrm{T}}^{b_2}\right|$ or $\left|p_{\mathrm{T}}^{b_1} - p_{\mathrm{T}}^{b_2}\right| / p_{\mathrm{T}}^{H_{\mathrm{cand}}}$ were $p_{\mathrm{T}}^{H_{\mathrm{cand}}}$ is the transverse momentum of the two b-tagged jets system constituting the Higgs candidate.
- The ratio $\Delta R(b, b) p_{\mathrm{T}}^{H_{\mathrm{cand}}} / m_{bb}$ which should be approximately equal to 2 for decays such as $H \to b\bar{b}$ or $Z \to b\bar{b}$ desintegrations according to Equation (5.1.1) and on the other hand whose distribution is not expected to peak around a given value for non-resonant background.
- The ratio $E_{\mathrm{T}}^{\mathrm{miss}} / p_{\mathrm{T}}^{H_{\mathrm{cand}}}$ which is expected to be around 1 for $ZH \to \nu\nu b\bar{b}$ decays.

However, no improvement were found even trying to re-optimize the hyperparameters including one or several of the variables listed.

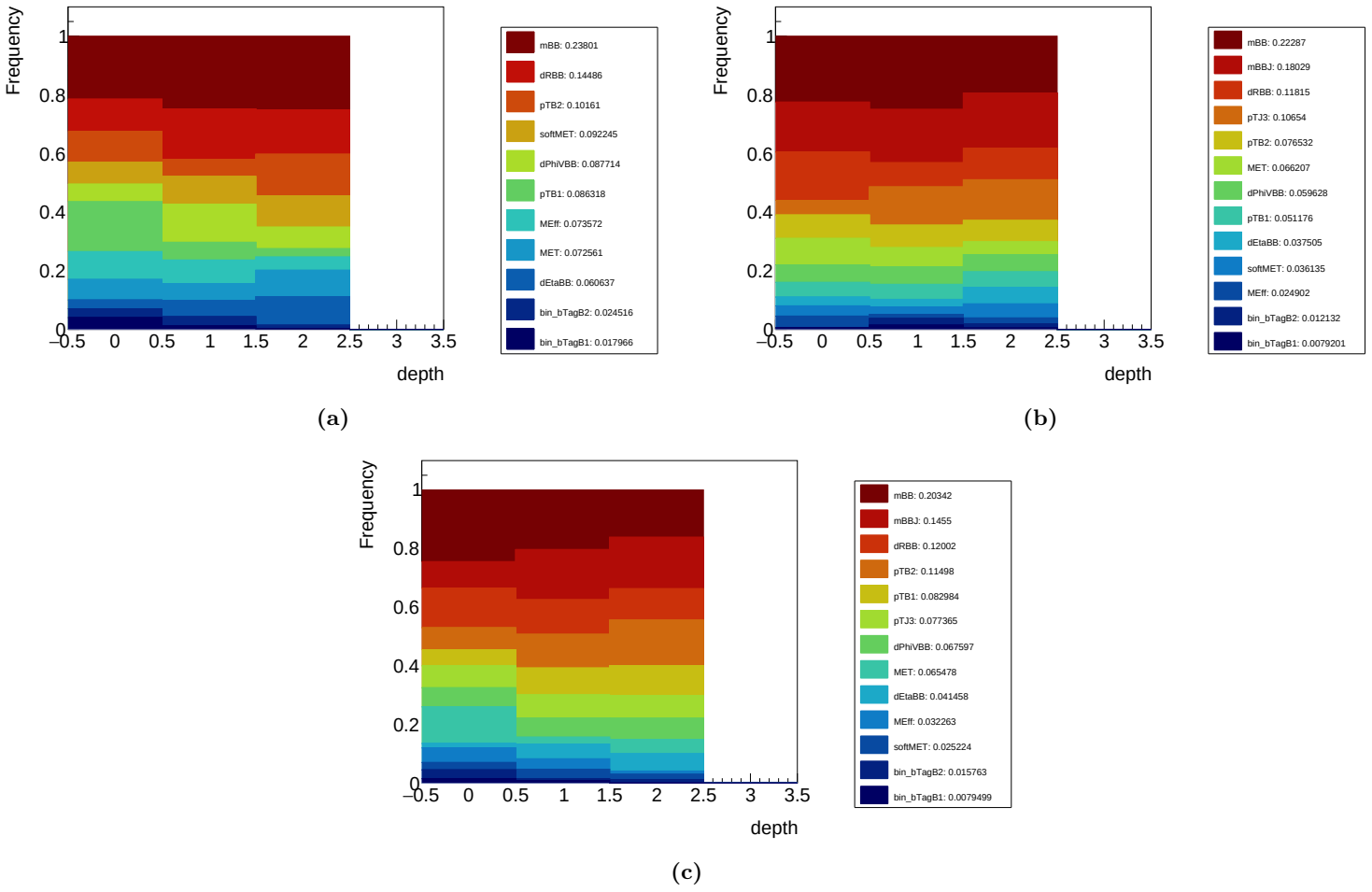


Figure 6.12: Rates at which each input variable is used by nodes of decision trees to perform cuts for each depth (starting at 0) in the (a) 2-jets, (b) 3-jets and (c) 4-jets BDT trainings with truth tagging input samples. The overall frequency of each variable is reported (as a percentage) on the right of each figure. Variables are ranked from the most used variable to the least used one.

6.1.5 Conclusion of the MVA training

As a conclusion, the BDTs hyperparameters have been re-optimized in each jet category and overtraining was found to be negligible thanks to the larger dataset (especially for the background) provided by the use of truth tagging, which reduces the chances of the training to optimize the BDT weights towards a statistical fluctuation in the input samples. For the Legacy Run 2 publications these BDTs, or updated ones retrained with the most up-to-date event selection and samples obtained after applying truth tagging or the GNN truth tagging technique, will be used. An extension to the 4-jets category BDT training has been carried out for the first time resulting in a potential 2σ significance. This additional jet category could also benefit for the granular STXS measurements as a function of the number of jets which is foreseen. Including this jet category for fits is hence currently investigated. An improvement of the signal sensitivity is achieved mainly thanks to the DL1r tagger and its better c- and light-jet rejections. Finally, searches for new discriminating variables have been performed, but no significant improvement have been found.

6.2 Combination study of the resolved and boosted topologies in the 1-lepton channel

For the $VH, H \to b\bar{b}$ analyses, a combination of the resolved and boosted regimes as a function of the transverse momentum of the vector boson was performed [5] as described in Section 5.9.1. Such results are of high interest as for example more stringent limits for the EFT interpretation were obtained in comparison to the individual $VH, H \to b\bar{b}$ analyses: the resolved [65] and the boosted [275] ones as detailed in Section 5.9.1.4.

Since those published results, many improvements discussed in the previous Chapter 5 have been developed as also reminded at the beginning of the current chapter.

The current study focuses on optimizing the combination of the two topologies in the 1-lepton channel by means of different event reconstruction strategies that are presented in Section 6.2.1 exploiting the important improvement in terms of signal sensitivity compared to the already published analyses. The Section 6.2.2 emphasizes the details of the study in terms of significance computation. The 1-lepton combination results are presented in Section 6.2.3. Finally, the conclusions of this study are provided in Section 6.2.4.

6.2.1 Event selection and combination strategies

The current study uses the nominal samples described in Section 5.2. The only exception is the multi-jet background which is omitted as it should have a minor impact on the results obtained since its contribution is reduced to the percent level. The events are required to fulfil the full event selection described in Section 5.3 when reconstructed with the resolved or the boosted topologies including the SRs and b-tagging requirements.

6.2.1.1 Combination strategies

There are several strategies in order to (orthogonally) combine the resolved and boosted topologies. The aim is to optimize the combination of events in order to improve the overall sensitivity to the signal for the following four p_{T}^{V}-regions: $[150, 250], [250, 400], [400, 600]$ and > 600 GeV. Below 400 GeV events are mainly expected to be consistent with the resolved topology. Moreover, the boundary at 400 GeV align with the fiducial definition of the VH STXS categories in the

reduced 1.2 STXS scheme and other VH analyses. The strategies are summarized in Figure 6.13 and described below:

- the ***split at 400 GeV*** strategy (abbreviated *split400* or *split@400* for simplicity): corresponds to a split of the 2 topologies depending on the p_{T}^{V} of the events. For $p_{\mathrm{T}}^{V} < 400$ GeV, events are reconstructed with the resolved topology *i.e.* they are required to pass the full resolved event selection. For $p_{\mathrm{T}}^{V} > 400$ GeV, events are reconstructed with the boosted topology *i.e.* they are required to pass the full boosted event selection. The split400 strategy was the strategy adopted for the previous combination paper [5].
- The ***split at 600 GeV*** strategy (abbreviated *split600* or *split@600*): is the same strategy as the split400 one but with the separation between the resolved and boosted topology being at $p_{\mathrm{T}}^{V} = 600$ GeV.
- The ***resolved priority*** strategy: gives priority to the resolved topology. Events passing the resolved selection without testing the b-tagging related requirements are then tested to pass the full resolved selection including b-tagging criteria. If they satisfy this latest requirement they fall in the resolved category otherwise they are rejected. On the other hand, the events not passing the resolved selection without testing the b-tagging requirements are then tested to pass the boosted event selection. If they satisfy this latest requirement they fall in the boosted category otherwise they are rejected.
- The ***boosted priority*** strategy: it is the exact opposite of the resolved priority strategy inverting resolved by boosted and vice-versa.
- The ***"ideal"*** strategy: events only passing the resolved (respectively the boosted) selection fall in the resolved (respectively the boosted) category. For events passing both the resolved and boosted selections then the split600 strategy is used. This strategy was motivated by the fact that for the significances extracted from the Higgs mass distributions, the resolved topology was outperforming the boosted one in the 0- and 1-lepton channels for $p_{\mathrm{T}}^{V} \in [400, 600]$ GeV. Hence, the split at 600 GeV. Such strategy is called "ideal" because it would indeed require taking into account correlations between the resolved and boosted tagging strategies as both exploit information related to the tracker which is not trivial to account for in terms of correlations.

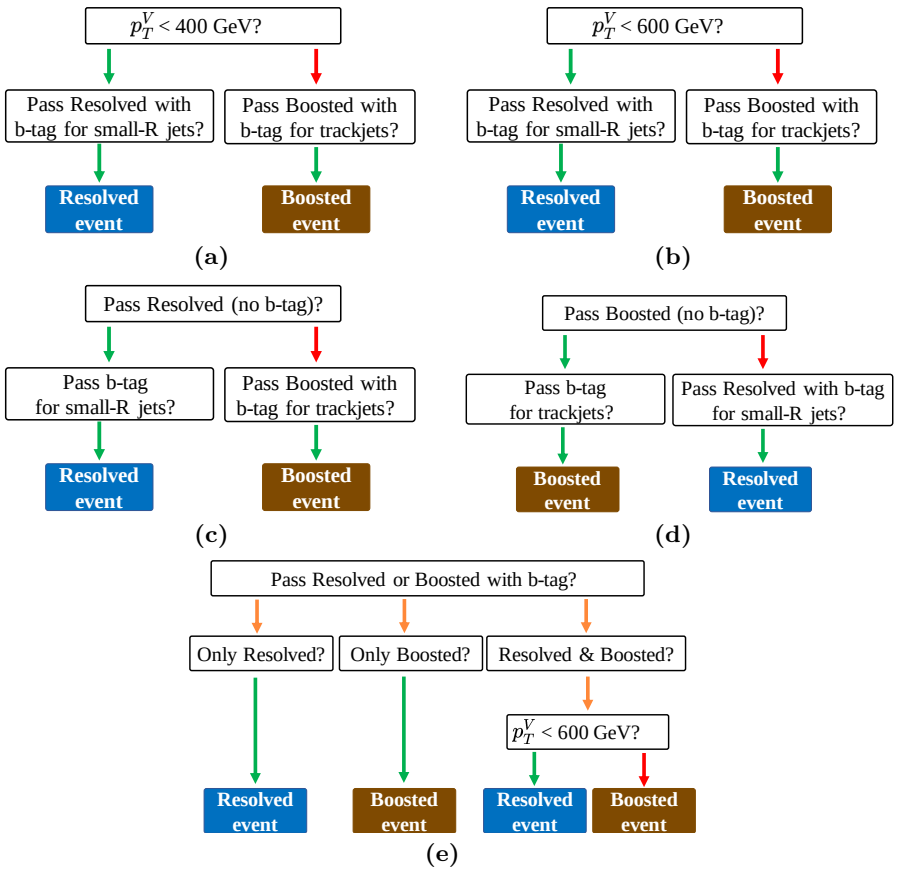


Figure 6.13: Combination strategies: (a) split at 400 GeV, (b) split at 600 GeV, (c) resolved priority, (d) boosted priority and (e) ideal strategy.

6.2.1.2 Multivariate discriminants

For the strategies described in the previous section, several BDTs have been trained:

- for the resolved topology, two BDTs are trained for $p_{\mathrm{T}}^V > 150$ GeV corresponding to the 2- and 3-jets categories.
- For the boosted topology, two BDTs are trained corresponding to the two p_{T}^V-regions: $250 < p_{\mathrm{T}}^V < 400$ GeV and $p_{\mathrm{T}}^V > 400$ GeV.

The BDTs hyperparameters have been optimized in order to improve ROC curves and the statistical significance defined in Equation (5.7.27) while at the same time avoiding overtraining. For the $[150, 250]$ GeV p_{T}^V-bin, events are only reconstructed in the resolved topology, the reconstruction for the boosted topology is only starting for $p_{\mathrm{T}}^V \geq 250$ GeV as almost no boosted signal event are expected below 250 GeV.

6.2.2 Significances and transformation D

In the current study significances are extracted either from the mass of the Higgs candidate or the BDTs output distributions by computing the Asimov significance from Equation (5.7.27). For the MVA distributions, the transformation D presented in Section 5.7.7 is applied before computing the significance. The significances are obtained by only taking into account the nominal MC predictions without considering systematics uncertainties. Only the statistical uncertainties on the significance are taken into account, Equation (5.7.29). Moreover, only the signal regions are exploited: for the resolved topology the 2 b-tagged 2- and 3-jets SRs and for the boosted topology

the high and low purity SRs. The distributions used are hence pre-fit ones and simulated processes are not normalized to data by means of the control regions. Moreover, the b-tagging scale factors are not applied as events can be reconstructed either as resolved or boosted depending on the strategy used. Those scale factors can differ depending on the topology: the aim here is to have a fair comparison between the two topologies using the same event weight for both regimes.

In the following the output of the BDT algorithm are respectively denoted MVA(Reso.) and MVA(Boost.) in figures. Unless specified otherwise, the 70% b-tagging WP is used for the resolved topology and the 85% WP leading 3 strategy is used for the boosted topology. Moreover, the mass distributions m_{bb} and m_J have 20 bins between 50 and 250 GeV in order to keep a reasonable number of bins while still avoiding too large statistical fluctuations due to the decrease of signal and background events selected for increasing p_{T}^V-regions. For the same reason the transformation D parameters are set to be equal to:

- $z_s = 10$ and $z_b = 5$ for $p_{\mathrm{T}}^V \in [150, 250]$ GeV.
- $z_s = 5$ and $z_b = 3$ for $p_{\mathrm{T}}^V \in [250, 400]$ GeV.
- $z_s = 3$ and $z_b = 2$ for $p_{\mathrm{T}}^V \in [400, 600]$ GeV and $p_{\mathrm{T}}^V > 600$ GeV.

6.2.3 Combination results

The Section 6.2.3.1 presents the results obtained with the different strategies using both the mass and BDT output distributions. In the Section 6.2.3.2 the impact on the significance of the hyperparameters used for the transformation D is discussed. The impact of the b-tagging strategy used for the boosted regime is shown in Section 6.2.3.3.

6.2.3.1 Results for the different strategies

The Figure 6.14 presents the obtained significances extracted from the mass or BDT output distributions for the 1-lepton channel as a function of p_{T}^V for the different strategies. The Figures 6.15 and 6.16 (respectively 6.17 and 6.18) are showing the mass m_{bb} and resolved BDT output (m_J and boosted regime BDT output) for the reconstructed Higgs candidate using the resolved (boosted) topology in the different p_{T}^V-regions of interest. The Table 6.4 provides the significance in each subregion of the phase space: 2-jets, 3-jets, HP and LP signal regions. Finally, the Table 6.5 details the signal and background yields in the different regions of phase space with as well the signal efficiency and the signal to background ratio.

Several conclusions can be extracted from the Figure 6.14:

- In all p_{T}^V-regions using the BDTs yields a significant improvement of the sensitivity compared to simply relying on the mass of the Higgs candidate. This means the BDTs benefit from the other input variables provided and do not only rely on the mass of the Higgs candidate even though it is the most discriminating variable (as shown for instance in Figure 6.12).
- For the $[250, 400]$ GeV p_{T}^V-region, the boosted priority strategy performances are worse both for the mass and BDT output. This can be explained by kinematic reasons as $p_{\mathrm{T}}^V \geq 250$ GeV is the loosest bound for which reconstructing the $H \to b\bar{b}$ decays with a large radius jet starts to become possible in ATLAS. In the 1-lepton channel, this region is dominated by background events in particular $t\bar{t}$ and W + jets events as shown in Figures 6.17 and 6.18. The boosted topology is not adequate for this region and reconstructing events with the resolved topology should be preferred.
- The $[400, 600]$ GeV p_{T}^V-bin is the most interesting one. For the Higgs mass significance, the resolved topology (split600 and resolved priority strategies) is outperforming the boosted topol-

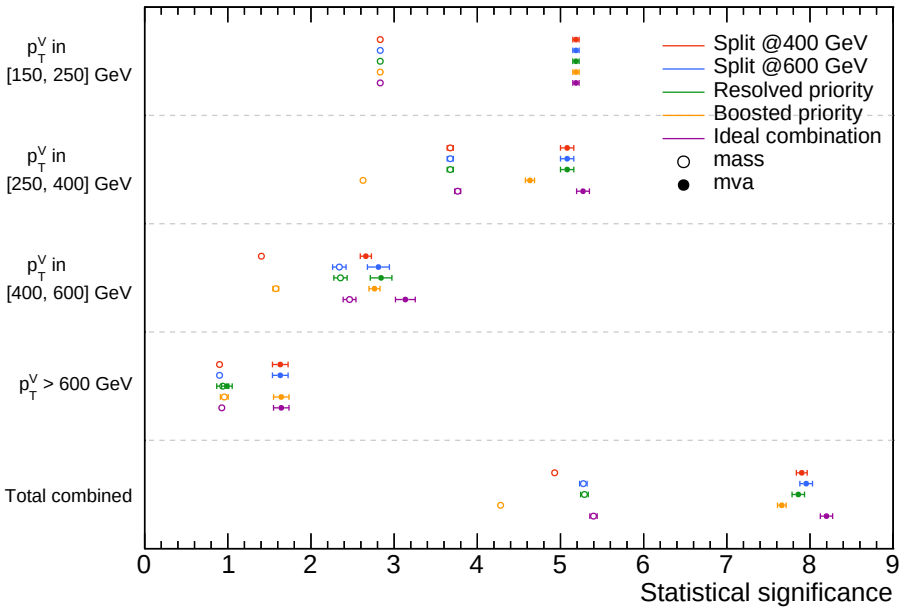


Figure 6.14: Summary of the significances extracted from the mass and BDT output for the different p_{T}^{V}-regions and strategies for the 1-lepton channel. The total combined significance is obtained summing in quadrature all the significances for the different p_{T}^{V}-regions. The error bars are statistical only error bars computed with the Equation (5.7.28).

ogy (split400 and boosted priority strategies). However, when using the BDT distributions, the boosted related strategies catch up with the resolved ones as the significances are compatible within statistical uncertainties. It implies the BDTs trained for the boosted topology are able to grasp differences between signal and background events based on the internal sub-structure and characteristics of the Higgs large-R jet candidates (Table 5.6). Moreover, those information have a large discriminating power as when only using the mass, the boosted topology is far from competing with the resolved event reconstruction. For the resolved related strategies, passing from the mass to the BDT extracted significances yields a relatively small improvement proving that for this p_{T}^{V}-regime the mass m_{bb} is almost the only discriminating variable. The angular distance $\Delta R(b, b)$ which is one of the most discriminating variable in the inclusive ranking (Figure 6.12) might not be helpful for this region due to the proximity between the two b-hadrons. In addition, the signal to background ratio (Table 6.5) is a factor 6 higher for the resolved topology compared to the boosted one. However, the BDTs significances are similar proving that the boosted event reconstruction seems more appropriate. Finally, the resolved event reconstruction acceptance is a factor 2 and 12 lower for signal and background events in SRs compared to the boosted one in that p_{T}^{V}-region. In the context of fits, and to reduce statistical errors it is thus better to use the boosted event reconstruction.

- ➢ For the $p_{\mathrm{T}}^{V} > 600$ GeV region, the resolved priority strategy does not improve when passing from the mass only significance to the BDT significance. This can be explained as only boosted events are expected in that region of phase space. The statistic is though very limited even when reconstructing events with the boosted topology.
- ➢ Finally, in the $[250, 400]$ GeV and $[400, 600]$ GeV p_{T}^{V}-bins, the ideal combination strategy is slightly performing better than the other strategies. However, this strategy complicates treatment of events, in particular applying the correct b-tagging scale factors and correlations would need to be taken into account. For these reasons, despite the small improvement in sensitivity, the ideal strategy is not considered further.

For all the reason cited above the best strategy for the 1-lepton channel seems hence to be the split

at 400 GeV which is a simple yet an effective strategy. The division of phase space for this strategy corresponds to the expected kinematic characteristics of the Higgs decay to a pair of b-hadrons, and the split400 comes with relatively good performances, a good signal acceptance and background rejection in the signal regions when relying on the boosted decision trees with respect to more sophisticated strategies such as the resolved priority or ideal strategies.

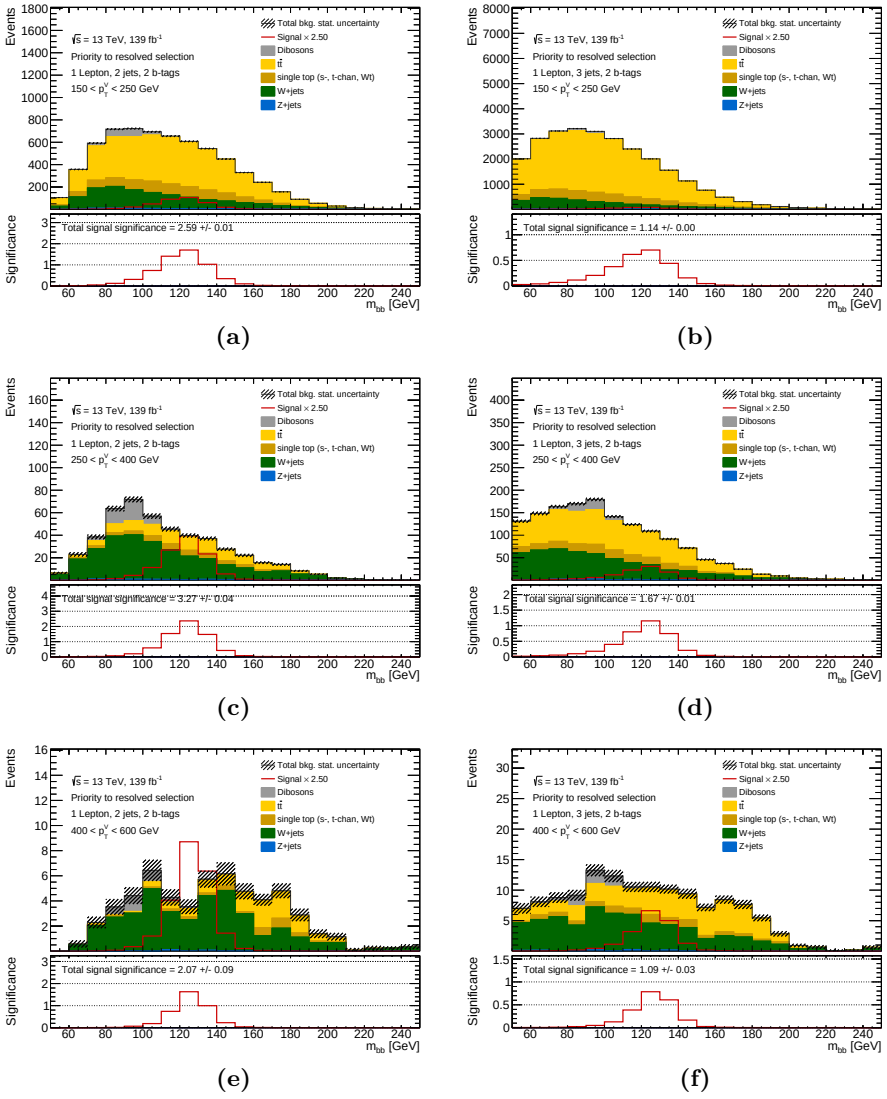


Figure 6.15: Mass for the resolved topology of the Higgs candidate in the (a), (c) and (e) 2 b-tagged 2-jets, and (b), (d) and (f) and 2 b-tagged 3-jets SRs in the (a) and (b) $[150, 250]$ GeV, (c) and (d) $[250, 400]$ GeV, and (e) and (f) $[400, 600]$ GeV p_{T}^{V}-regions.

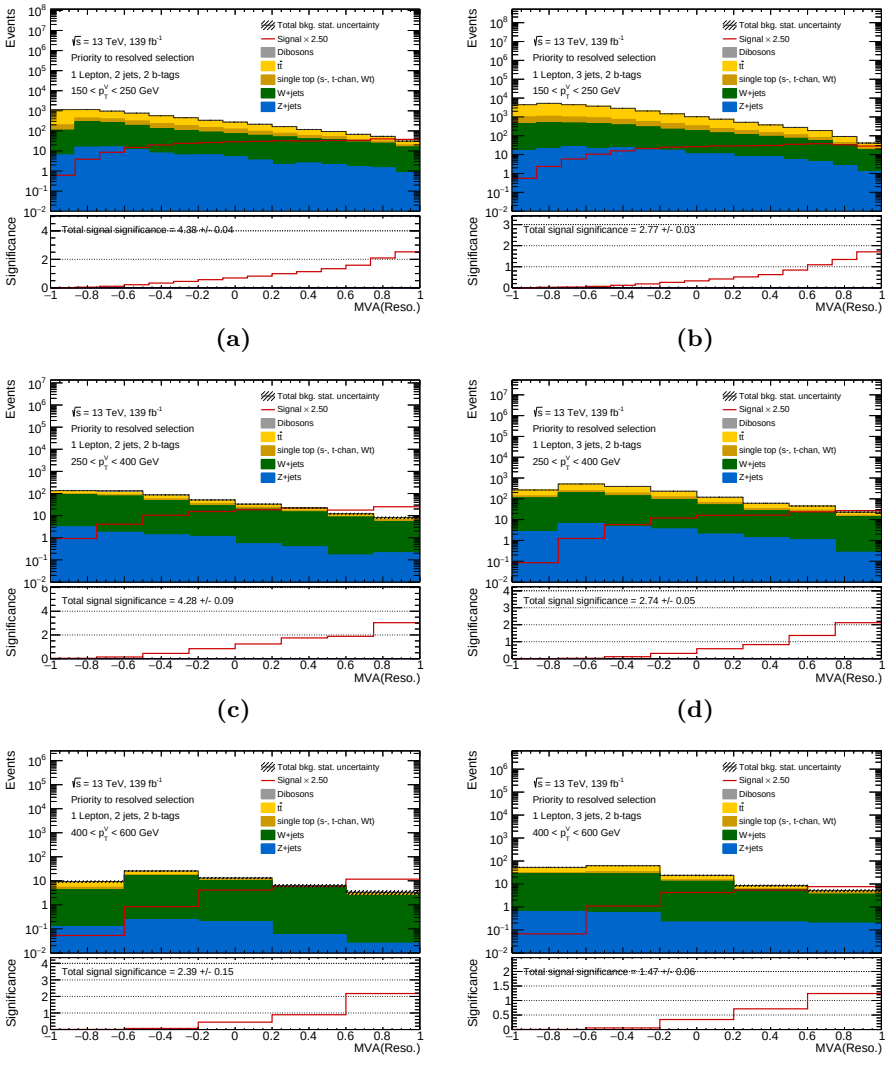


Figure 6.16: BDT output for the resolved topology of the Higgs candidate in the (a), (c) and (e) 2 b-tagged 2-jets, and (b), (d) and (f) and 2 b-tagged 3-jets SRs in the (a) and (b) $[150, 250]$ GeV, (c) and (d) $[250, 400]$ GeV, and (e) and (f) $[400, 600]$ GeV p_{T}^{V}-regions.

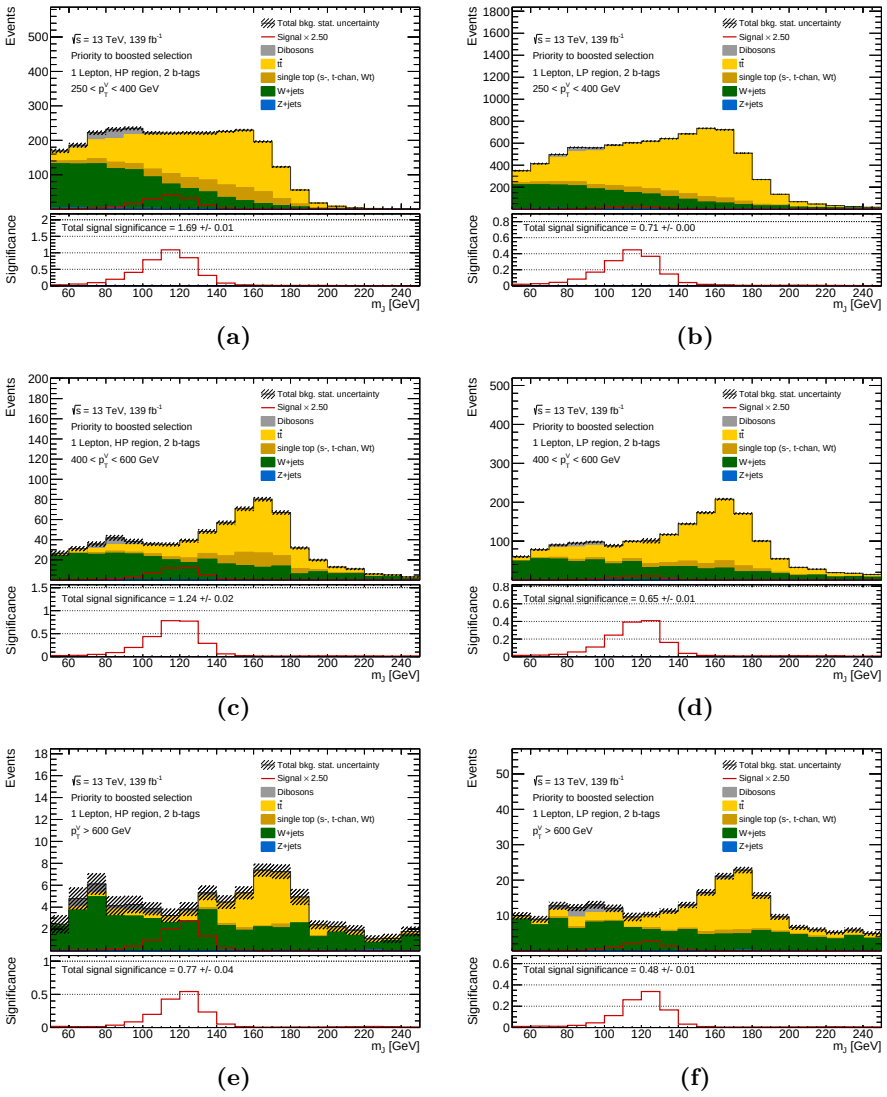


Figure 6.17: Mass for the boosted topology of the Higgs candidate in the (a), (c) and (e) HP and (b), (d) and (f) LP SRs in the (a) and (b) $[250, 400]$ GeV, (c) and (d) $[400, 600]$ GeV, and (e) and (f) > 600 GeV p_{T}^{V}-regions.

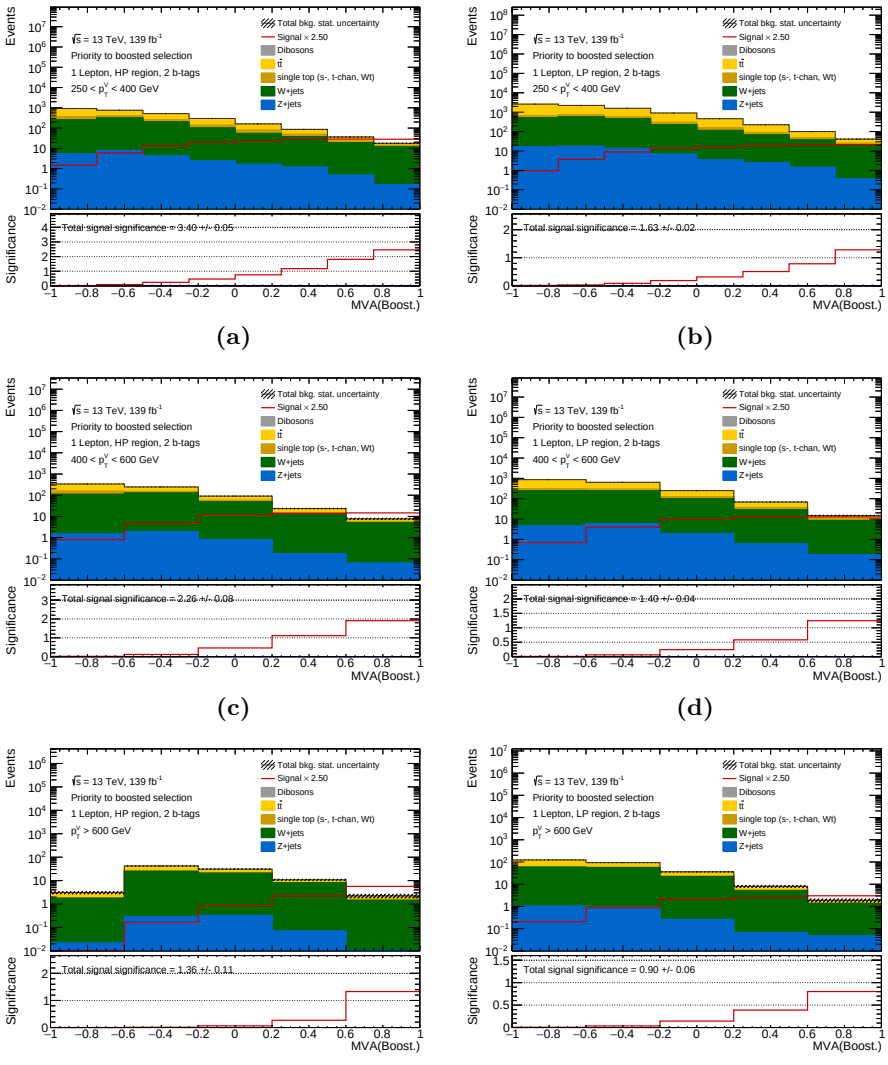


Figure 6.18: BDT output for the boosted topology of the Higgs candidate in the (a), (c) and (e) HP and (b), (d) and (f) LP SRs in the (a) and (b) $[250, 400]$ GeV, (c) and (d) $[400, 600]$ GeV, and (e) and (f) > 600 GeV p_{T}^{V}-regions.

		Strategies																			
		Split 400				Split 600				Resolved priority				Boosted priority				Ideal			
		mass		mva		mass		mva		mass		mva		mass		mva		mass		mva	
p_T^V	Region	S	δS	S	δS	S	δS	S	δS	S	δS	S	δS	S	δS	S	δS	S	δS	S	δS
[150, 250] GeV	2j	2.593	0.009	4.381	0.042	2.593	0.009	4.381	0.042	2.593	0.009	4.381	0.042	2.593	0.009	4.381	0.042	2.593	0.009	4.381	0.042
	3j	1.14	0.003	2.774	0.026	1.14	0.003	2.774	0.026	1.14	0.003	2.774	0.026	1.14	0.003	2.774	0.026	1.14	0.003	2.774	0.026
[250, 400] GeV	2j	3.273	0.041	4.277	0.091	3.273	0.041	4.277	0.091	3.273	0.041	4.277	0.091	1.639	0.029	2.17	0.071	3.273	0.041	4.277	0.091
	3j	1.672	0.013	2.74	0.047	1.672	0.013	2.74	0.047	1.672	0.013	2.74	0.047	0.919	0.01	1.594	0.046	1.672	0.013	2.74	0.047
	HP	–	–	–	–	–	–	–	–	0.054	0.002	0.073	0.006	1.693	0.01	3.4	0.054	0.64	0.005	0.974	0.013
	LP	–	–	–	–	–	–	–	–	0.021	0.001	0.032	0.003	0.705	0.003	1.631	0.023	0.507	0.002	1.014	0.012
[400, 600] GeV	2j	–	–	–	–	2.073	0.09	2.395	0.151	2.073	0.09	2.395	0.151	0.648	0.076	0.6	0.05	2.073	0.09	2.395	0.151
	3j	–	–	–	–	1.089	0.026	1.473	0.056	1.089	0.026	1.473	0.056	0.311	0.017	0.452	0.041	1.089	0.026	1.473	0.056
	HP	1.243	0.018	2.262	0.076	–	–	–	–	0.245	0.014	0.404	0.036	1.243	0.018	2.262	0.076	0.58	0.01	0.972	0.032
	LP	0.652	0.01	1.398	0.037	–	–	–	–	0.064	0.006	0.119	0.013	0.652	0.01	1.398	0.037	0.504	0.007	0.991	0.023
[600, +∞(GeV	2j	–	–	–	–	–	–	–	–	0.574	0.098	0.426	0.042	0.297	0.115	0.161	0.026	0.175	0.029	0.144	0.015
	3j	–	–	–	–	–	–	–	–	0.346	0.035	0.347	0.034	0.133	0.018	0.116	0.013	0.139	0.016	0.141	0.015
	HP	0.765	0.038	1.358	0.105	0.765	0.038	1.358	0.105	0.604	0.04	0.812	0.073	0.765	0.038	1.358	0.105	0.765	0.038	1.358	0.105
	LP	0.475	0.013	0.903	0.06	0.475	0.013	0.903	0.06	0.262	0.093	0.132	0.014	0.475	0.013	0.903	0.06	0.475	0.013	0.903	0.06
[150, 250] GeV	all	2.833	0.009	5.186	0.038	2.833	0.009	5.186	0.038	2.833	0.009	5.186	0.038	2.833	0.009	5.186	0.038	2.833	0.009	5.186	0.038
[250, 400] GeV	all	3.675	0.037	5.08	0.08	3.675	0.037	5.08	0.08	3.675	0.037	5.08	0.08	2.626	0.02	4.633	0.055	3.764	0.036	5.271	0.077
[400, 600] GeV	all	1.404	0.017	2.66	0.067	2.341	0.081	2.811	0.132	2.355	0.08	2.843	0.13	1.577	0.035	2.764	0.066	2.464	0.077	3.135	0.119
[600, +∞(GeV	all	0.901	0.033	1.631	0.094	0.901	0.033	1.631	0.094	0.939	0.072	0.989	0.064	0.958	0.047	1.643	0.093	0.928	0.032	1.643	0.093
[150, +∞(GeV	all	4.93	0.029	7.901	0.065	5.275	0.045	7.954	0.076	5.288	0.046	7.859	0.075	4.28	0.021	7.661	0.052	5.397	0.044	8.198	0.074

Table 6.4: Detail of the significances (S) and their statistical uncertainty (δS) extracted in the different signal regions: 2-jets (2j), 3-jets (3j), high purity (HP) and low purity (LP) for the different strategies and p_{T}^{V}-regions.

			Strategies														
			Split 400			Split 600			Resolved priority			Boosted priority			Ideal		
p_T^V	Region	Process	yield	sig./bkg.	sig. eff.	yield	sig./bkg.	sig. eff.	yield	sig./bkg.	sig. eff.	yield	sig./bkg.	sig. eff.	yield	sig./bkg.	sig. eff.
[150, 250] GeV	2j	sig.	145.89	2.28 %	31.77 %	145.89	2.28 %	31.77 %	145.89	2.28 %	31.77 %	145.89	2.28 %	31.77 %	145.89	2.28 %	31.77 %
		bkg.	6393.08			6393.08			6393.08			6393.08			6393.08		
	3j	sig.	132.72	0.48 %	28.91 %	132.72	0.48 %	28.91 %	132.72	0.48 %	28.91 %	132.72	0.48 %	28.91 %	132.72	0.48 %	28.91 %
		bkg.	27490.00			27490.00			27490.00			27490.00			27490.00		
[250, 400] GeV	2j	sig.	45.76	9.52 %	24.10 %	45.76	9.52 %	24.10 %	45.76	9.52 %	24.10 %	16.98	6.39 %	8.94 %	45.76	9.52 %	24.10 %
		bkg.	480.58			480.58			480.58			265.63			480.58		
	3j	sig.	40.99	2.48 %	21.59 %	40.99	2.48 %	21.59 %	40.99	2.48 %	21.59 %	17.41	1.59 %	9.17 %	40.99	2.48 %	21.59 %
		bkg.	1654.03			1654.03			1654.03			1097.84			1654.03		
	HP	sig.	–	–	–	–	–	–	0.41	0.29 %	0.21 %	59.91	2.15 %	31.56 %	21.66	0.91 %	11.41 %
		bkg.	–			–			141.02			2792.24			2385.04		
	LP	sig.	–	–	–	–	–	–	0.15	0.15 %	0.08 %	41.88	0.52 %	22.06 %	30.41	0.38 %	16.02 %
		bkg.	–			–			103.78			8130.55			7961.02		
[400, 600] GeV	2j	sig.	–	–	–	9.04	15.74 %	20.03 %	9.04	15.74 %	20.03 %	1.23	10.51 %	2.72 %	9.04	15.74 %	20.03 %
		bkg.	–			57.46			57.46			11.70			57.46		
	3j	sig.	–	–	–	7.37	4.84 %	16.32 %	7.37	4.84 %	16.32 %	1.00	2.06 %	2.22 %	7.37	4.84 %	16.32 %
		bkg.	–			152.14			152.14			48.50			152.14		
	HP	sig.	18.09	2.57 %	40.05 %	–	–	–	1.01	1.69 %	2.24 %	18.09	2.57 %	40.05 %	8.12	1.31 %	17.98 %
		bkg.	702.99			–			59.62			702.99			621.99		
	LP	sig.	15.64	0.85 %	34.62 %	–	–	–	0.19	0.70 %	0.43 %	15.64	0.85 %	34.62 %	12.24	0.68 %	27.10 %
		bkg.	1836.01			–			27.76			1836.01			1802.89		
[600, +∞(GeV	2j	sig.	–	–	–	–	–	–	0.72	13.55 %	7.98 %	0.17	16.51 %	1.83 %	0.21	9.71 %	2.32 %
		bkg.	–			–			5.34			1.01			2.17		
	3j	sig.	–	–	–	–	–	–	0.68	3.66 %	7.50 %	0.16	2.59 %	1.73 %	0.22	1.95 %	2.44 %
		bkg.	–			–			18.57			6.05			11.31		
	HP	sig.	3.56	4.00 %	39.32 %	3.56	4.00 %	39.32 %	1.52	11.84 %	16.72 %	3.56	4.00 %	39.32 %	3.56	4.00 %	39.32 %
		bkg.	89.00			89.00			12.80			89.00			89.00		
	LP	sig.	3.62	1.38 %	39.98 %	3.62	1.38 %	39.98 %	0.23	6.70 %	2.58 %	3.62	1.38 %	39.98 %	3.62	1.38 %	39.98 %
		bkg.	262.67			262.67			3.49			262.67			262.67		
[150, 250] GeV	all	sig.	278.61	0.82 %	60.68 %	278.61	0.82 %	60.68 %	278.61	0.82 %	60.68 %	278.61	0.82 %	60.68 %	278.61	0.82 %	60.68 %
		bkg.	33883.08			33883.08			33883.08			33883.08			33883.08		
[250, 400] GeV	all	sig.	86.75	4.06 %	45.69 %	86.75	4.06 %	45.69 %	87.31	3.67 %	45.99 %	136.18	1.11 %	71.73 %	138.83	1.11 %	73.12 %
		bkg.	2134.62			2134.62			2379.41			12286.25			12480.68		
[400, 600] GeV	all	sig.	33.72	1.33 %	74.67 %	16.41	7.83 %	36.35 %	17.62	5.93 %	39.01 %	35.95	1.38 %	79.61 %	36.77	1.40 %	81.42 %
		bkg.	2539.00			209.60			296.99			2599.20			2634.47		
[600, +∞(GeV	all	sig.	7.19	2.04 %	79.30 %	7.19	2.04 %	79.30 %	3.15	7.84 %	34.79 %	7.51	2.09 %	82.87 %	7.62	2.09 %	84.06 %
		bkg.	351.67			351.67			40.19			358.73			365.15		

Table 6.5: Signal and background yields and signal efficiencies for the different strategies in the different signal regions 2-jets (2j), 3-jets (3j), high purity (HP) and low purity (LP) and p_{T}^{V}-regions. The signal efficiencies are computed per p_{T}^{V}-region and correspond to the ratio: number of signal events passing all selections for a given signal region over the total number of events (resolved and boosted combined) passing the kinematic selections (SRs and CRs combined without b-tagging requirement) described in Table 5.3. In this table *sig.* and *bkg.* stand for signal and background.

6.2.3.2 Impact of the transformation D parameters

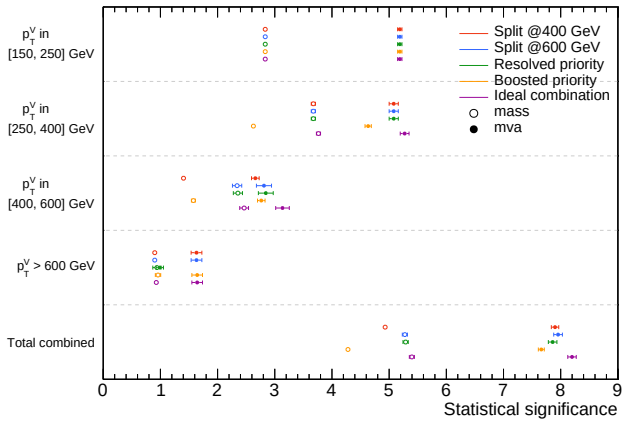
As explained in Section 5.7.7, the sum $z_s + z_b$ of the parameters for the transformation D is equal to the number of bins of the obtained transformed distribution. Increasing those two hyperparameters increases the number of bins and should increase the significance since the signal events are mainly concentrated in the highest bins.

The aim is to focus on the $[400, 600]$ GeV p_{T}^{V}-bin and assess the impact of those parameters. For that particular bin, the default parameters are $z_s = 3$ and $z_b = 2$ and the significances obtained are summarized by the Figure 6.14 and Table 6.4. In the $[400, 600]$ GeV p_{T}^{V}-region, the statistical uncertainty of the significance obtained with the resolved topology (split600 strategy) is larger than the boosted one (split400 strategy). Comparing BDT output distributions in Figure 6.16 and 6.18 for that p_{T}^{V}-bin, it appears that the number of bins of the transformed distributions for the boosted topology could be increased as the signal and background acceptance is larger (Table 6.5) than for the resolved one hence the smaller statistical uncertainty for the split400 strategy.

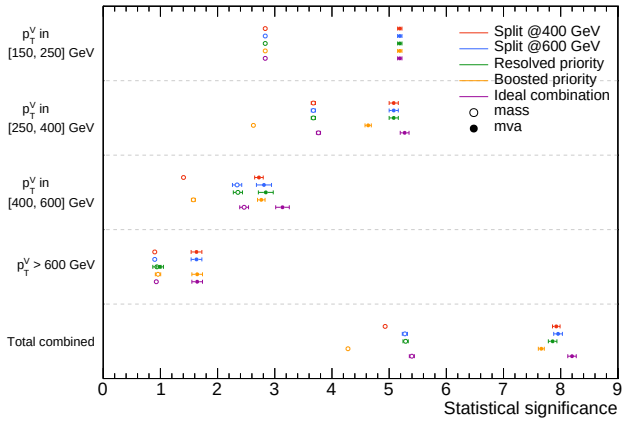
It was decided to modify the z_s, z_b hyperparameters only for the split400 strategy (*i.e.* for the boosted strategy) in the $[400, 600]$ GeV p_{T}^{V}-bin, using alternative hyperparameters, $(z_s = 4, z_b = 3)$ and $(z_s = 6, z_b = 4)$, to see how the related significance and its statistical uncertainty evolve with respect to the significance computed with the default parameters for the split600 strategy (resolved strategy). For the other strategies and p_{T}^{V}-regions, he z_s and z_b hyperparameters are kept to their default values.

As illustrated in Figure 6.19, with the default parameters, the boosted (split400) significance is smaller than the resolved (split600) significance. For an increased binning for the boosted topology, it is the other way around though results are always compatible within statistical uncertainties. Doubling the number of bins with respect to the default binning has a reasonably small impact on the significance. This test confirms that the boosted topology event reconstruction is competitive with the resolved topology one in the $[400, 600]$ GeV p_{T}^{V}-bin and that the slightly lower significance with the default hyperparameters for the boosted topology can simply be due to binning effects.

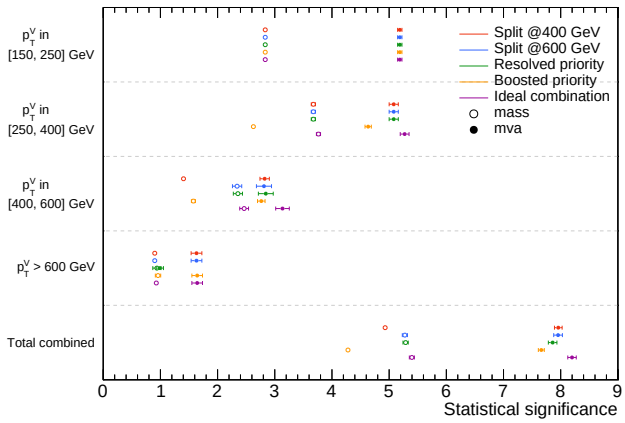
The impact of the z_s and z_b parameters on the BDT distributions after transformation D is shown in Figure 6.20 for the boosted topology (split400) which can be compared with the statistics obtained in Figure 6.21 for the resolved topology (split600).



(a)



(b)



(c)

Figure 6.19: Evolution of the significance of the split400 strategy in the $[400, 600]$ GeV p_{T}^{V}-bin using for the transformation D: (a) the default hyperparameters ($z_s = 3, z_b = 2$), (b) ($z_s = 4, z_b = 3$) and (c) ($z_s = 6, z_b = 4$).

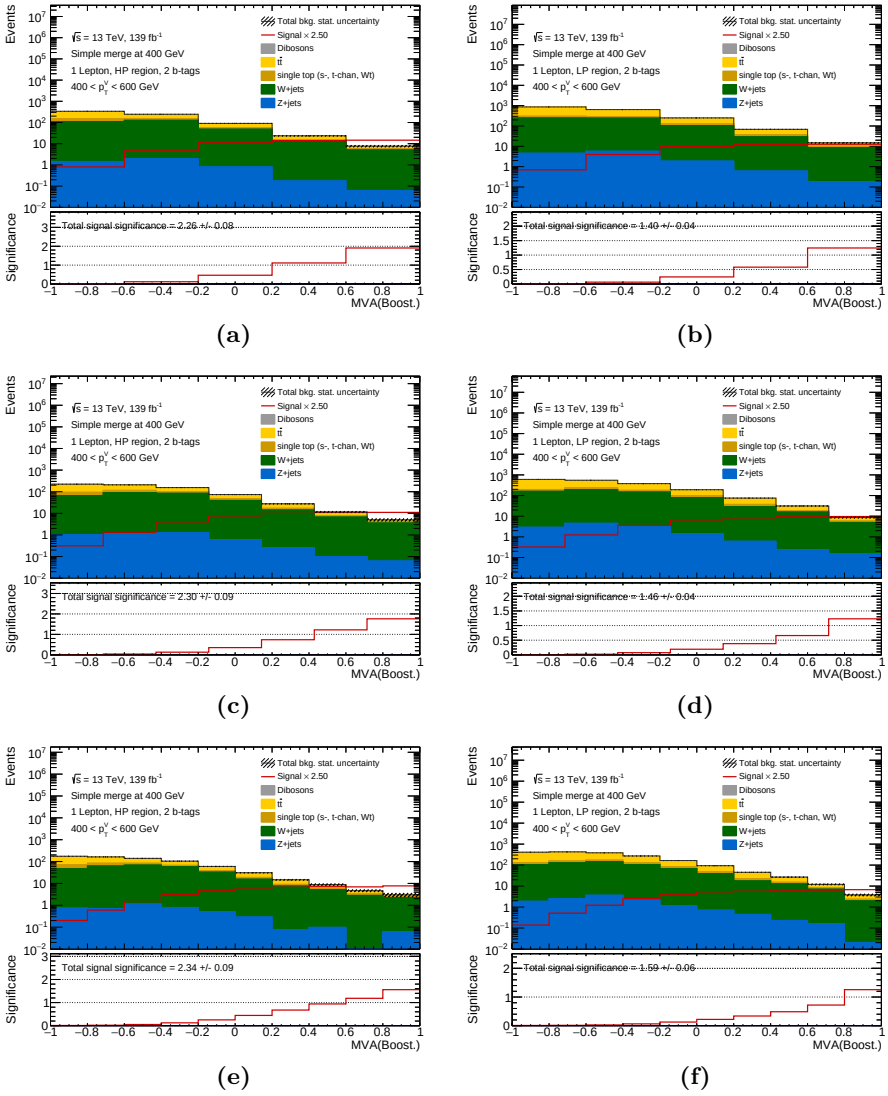


Figure 6.20: Evolution of the boosted BDT distributions in the (a), (c) and (e) HP and (b), (d) and (f) LP SRs in the $[400, 600]$ GeV p_{T}^{V}-bin after applying the transformation D: (a) and (b) with the default hyperparameters $(z_s = 3, z_b = 2)$, (c) and (d) with $(z_s = 4, z_b = 3)$ and (e) and (f) with $(z_s = 6, z_b = 4)$.

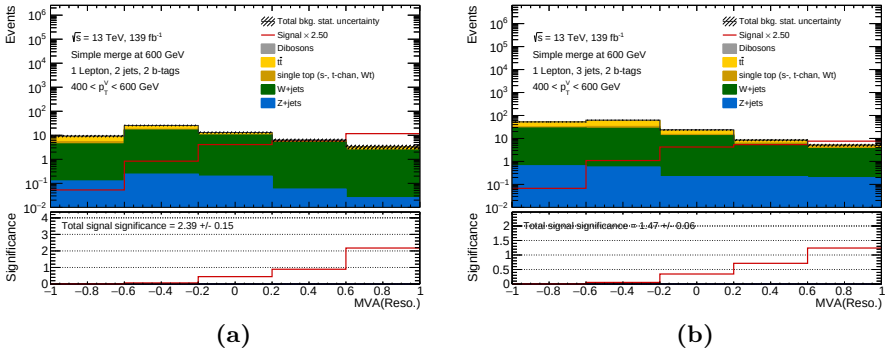


Figure 6.21: BDT output distributions for the resolved topology in the $[400, 600]$ GeV p_{T}^{V}-bin after applying the transformation D with the default hyperparameters $(z_s = 3, z_b = 2)$ in the (a) 2-jets and (a) 3-jets SRs.

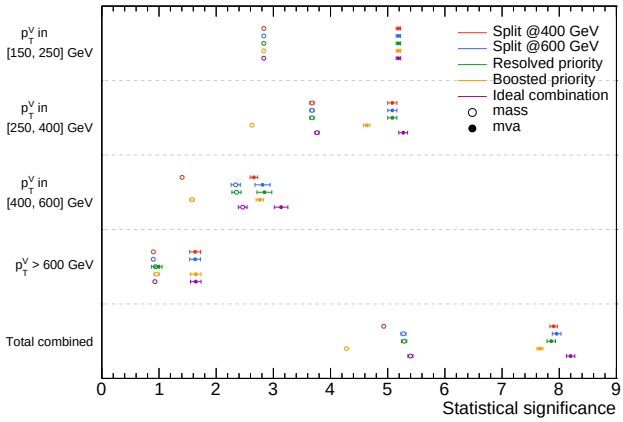
6.2.3.3 Impact of the b-tagging strategy for the boosted topology

For the boosted topology, as explained in Section 5.3.4, the new default 85% WP leading 3 strategy is used as b-tagging requirement whose results have been presented in Figure 6.14 for the significance summary, and Tables 6.4 and 6.5 respectively for the details of the significances in each subregion of the phase space and details about the signal and background yields in the different region with as well the signal efficiency and the signal to background ratio. Instead, the 85% WP leading 2 strategy or the old default 70% WP leading 2 strategy from the combination paper [5] could be used.

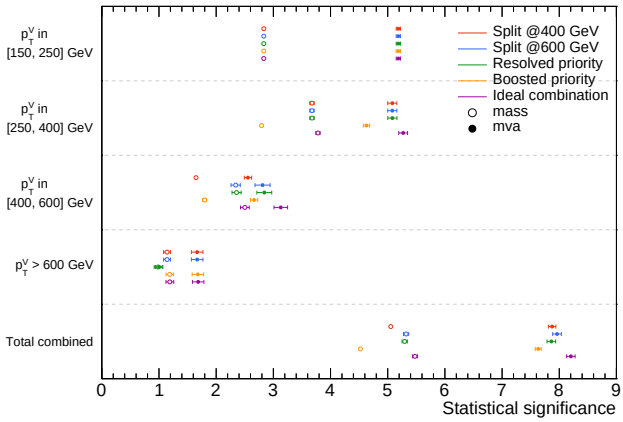
The Figure 6.22 compares the significances obtained with those three b-tagging strategies for the boosted topology event selection to see their impact. For each of these boosted b-tagging strategy, specific BDTs have been trained and optimized (by the $VH(b\bar{b}/c\bar{c})$ analysis team) since the events selected are affected by the tagging strategy. The Tables 6.6 and 6.7 provide the significance in each subregion of the phase space for the 85% WP leading 2 and 70% WP leading 2 strategies. The Tables 6.8 and 6.9 details the signal and background yields in the different region of phase space with as well the signal efficiency and the signal to background ratio for the 85% WP leading 2 and 70% WP leading 2 strategies.

Three conclusions can be drawn:

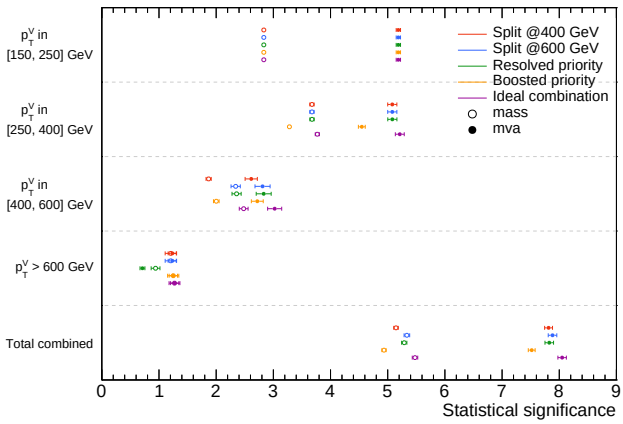
- The mass related significances are lower for the 85% WP leading 3 strategy than for the 70% WP leading 2 strategies. This can be explained by the large increase in acceptance especially the large increase of the $t\bar{t}$ and $W + \text{jets}$ contamination as shown in Figure 6.23. Moreover, the leading 3 strategy gives lower significances than the leading 2 strategy for the exact same reason. The 70% WP leading 2 strategy is the tightest selection hence it has the lowest acceptance for background and signal while the 85% WP leading 3 is the loosest selection. The mass significances obtained for the tightest b-tagging strategy (70% WP leading 2) is the largest while the loosest one (85% WP leading 3 strategy) results in the lowest significance for the $p_{\text{T}}^{V} \in [400, 600]$ GeV.
- The BDT output significances are pretty much equivalent for the three strategies taking into account the statistical uncertainties. However, the 85% WP b-tagging efficiency strategy results in a lower statistical error thanks to the increase in statistics for that strategy both for signal and background.
- Despite the lower signal to background ratio (Tables 6.5, 6.8 and 6.9) for the 85% b-tagging efficiency strategy compared to the 70% WP there is no difference for the BDT significances whereas the mass significances are lower for the leading 3 strategies. That switch of tagging strategy resulted in a large increase of a factor 5.5 for the background against only 1.7 for the signal between the 70% leading 2 and the 85% leading 3 strategies in the $[400, 600]\,\text{GeV}$ p_{T}^{V}-bin. This means that the BDTs trained for the boosted event reconstruction are truly able to discriminate the $t\bar{t}$ and $W + \text{jets}$ events from the signal events as they are the 2 main backgrounds of the 1-lepton channel (Figure 6.23) based on the sub-structure and internal information related to the large-R jet.



(a)



(b)



(c)

Figure 6.22: Summary of the significances extracted from the mass and BDT output for the different p_T^V-regions and strategies for the 1-lepton channel. Using for the boosted topology (a) the new default 85%WP leading 3 strategy, (b) the 85%WP leading 2 strategy and (c) the old default 70%WP leading 2 strategy.

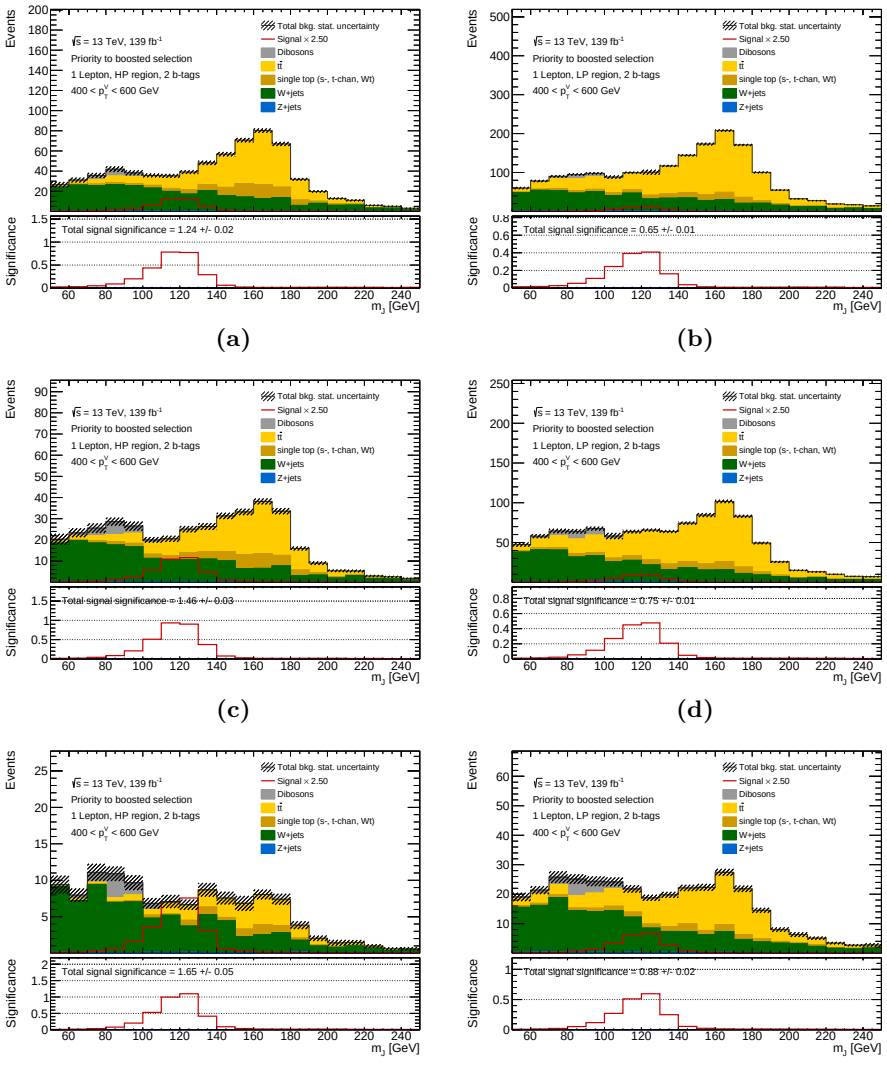


Figure 6.23: Mass for the boosted topology of the Higgs candidate in the (a), (c) and (e) HP, and (b), (d) and (f) LP SRs in the $[400, 600]$ GeV p_{T}^{V}-bin for (a) and (b) the new default 85%WP leading 3 strategy, (c) and (d) the 85%WP leading 2 strategy, and (e) and (f) the 70%WP leading 2 strategy.

		Strategies																			
		Split 400				Split 600				Resolved priority				Boosted priority				Ideal			
		mass		mva		mass		mva		mass		mva		mass		mva		mass		mva	
p_T^V	Region	S	δS	S	δS	S	δS	S	δS	S	δS	S	δS	S	δS	S	δS	S	δS	S	δS
$[150, 250]$ GeV	2j	2.593	0.009	4.381	0.042	2.593	0.009	4.381	0.042	2.593	0.009	4.381	0.042	2.593	0.009	4.381	0.042	2.593	0.009	4.381	0.042
	3j	1.14	0.003	2.774	0.026	1.14	0.003	2.774	0.026	1.14	0.003	2.774	0.026	1.14	0.003	2.774	0.026	1.14	0.003	2.774	0.026
$[250, 400]$ GeV	2j	3.273	0.041	4.277	0.091	3.273	0.041	4.277	0.091	3.273	0.041	4.277	0.091	1.639	0.029	2.17	0.071	3.273	0.041	4.277	0.091
	3j	1.672	0.013	2.74	0.047	1.672	0.013	2.74	0.047	1.672	0.013	2.74	0.047	0.919	0.01	1.593	0.046	1.672	0.013	2.74	0.047
	HP	–	–	–	–	–	–	–	–	0.071	0.003	0.092	0.007	1.908	0.013	3.375	0.054	0.69	0.006	0.955	0.014
	LP	–	–	–	–	–	–	–	–	0.028	0.002	0.044	0.004	0.793	0.004	1.67	0.026	0.559	0.003	1.028	0.014
$[400, 600]$ GeV	2j	–	–	–	–	2.073	0.09	2.395	0.151	2.073	0.09	2.395	0.151	0.648	0.076	0.6	0.05	2.073	0.09	2.395	0.151
	3j	–	–	–	–	1.089	0.026	1.473	0.056	1.089	0.026	1.473	0.056	0.311	0.017	0.452	0.041	1.089	0.026	1.473	0.056
	HP	1.465	0.027	2.129	0.073	–	–	–	–	0.264	0.018	0.405	0.037	1.465	0.027	2.129	0.073	0.671	0.015	0.962	0.037
	LP	0.752	0.008	1.414	0.039	–	–	–	–	0.075	0.006	0.103	0.011	0.752	0.008	1.414	0.039	0.575	0.007	0.982	0.024
$[600, +\infty($ GeV	2j	–	–	–	–	–	–	–	–	0.574	0.098	0.426	0.042	0.297	0.115	0.164	0.027	0.288	0.148	0.176	0.018
	3j	–	–	–	–	–	–	–	–	0.346	0.035	0.347	0.034	0.133	0.018	0.121	0.013	0.177	0.024	0.163	0.017
	HP	0.988	0.068	1.4	0.113	0.988	0.068	1.4	0.113	0.69	0.053	0.829	0.075	0.988	0.068	1.4	0.113	0.988	0.068	1.4	0.113
	LP	0.572	0.023	0.906	0.063	0.572	0.023	0.906	0.063	0.217	0.046	0.13	0.014	0.572	0.023	0.906	0.063	0.572	0.023	0.906	0.063
$[150, 250]$ GeV	all	2.833	0.009	5.186	0.038	2.833	0.009	5.186	0.038	2.833	0.009	5.186	0.038	2.833	0.009	5.186	0.038	2.833	0.009	5.186	0.038
$[250, 400]$ GeV	all	3.675	0.037	5.08	0.08	3.675	0.037	5.08	0.08	3.676	0.037	5.081	0.08	2.793	0.02	4.629	0.055	3.78	0.036	5.27	0.077
$[400, 600]$ GeV	all	1.647	0.024	2.556	0.064	2.341	0.081	2.811	0.132	2.357	0.08	2.842	0.13	1.797	0.036	2.664	0.063	2.502	0.076	3.13	0.119
$[600, +\infty($ GeV	all	1.141	0.06	1.668	0.101	1.141	0.06	1.668	0.101	0.986	0.07	1.003	0.065	1.187	0.064	1.68	0.1	1.19	0.068	1.685	0.1
$[150, +\infty($ GeV	all	5.054	0.032	7.875	0.065	5.321	0.046	7.961	0.076	5.297	0.046	7.861	0.075	4.523	0.026	7.631	0.052	5.477	0.045	8.203	0.074

Table 6.6: Detail of the significances (S) and their statistical uncertainty (δS) extracted in the different signal regions using for the boosted topology the 85% WP leading 2 strategy

		Strategies																			
		Split 400				Split 600				Resolved priority				Boosted priority				Ideal			
		mass		mva		mass		mva		mass		mva		mass		mva		mass		mva	
p_T^V	Region	S	δS	S	δS	S	δS	S	δS	S	δS	S	δS	S	δS	S	δS	S	δS	S	δS
[150, 250] GeV	2j	2.593	0.009	4.381	0.042	2.593	0.009	4.381	0.042	2.593	0.009	4.381	0.042	2.593	0.009	4.381	0.042	2.593	0.009	4.381	0.042
	3j	1.14	0.003	2.774	0.026	1.14	0.003	2.774	0.026	1.14	0.003	2.774	0.026	1.14	0.003	2.774	0.026	1.14	0.003	2.774	0.026
[250, 400] GeV	2j	3.273	0.041	4.277	0.091	3.273	0.041	4.277	0.091	3.273	0.041	4.277	0.091	1.639	0.029	2.17	0.071	3.273	0.041	4.277	0.091
	3j	1.672	0.013	2.74	0.047	1.672	0.013	2.74	0.047	1.672	0.013	2.74	0.047	0.919	0.01	1.593	0.046	1.672	0.013	2.74	0.047
	HP	–	–	–	–	–	–	–	–	0.091	0.006	0.099	0.009	2.499	0.03	3.298	0.064	0.53	0.01	0.687	0.02
	LP	–	–	–	–	–	–	–	–	0.039	0.002	0.054	0.004	0.995	0.007	1.599	0.026	0.646	0.005	0.935	0.014
[400, 600] GeV	2j	–	–	–	–	2.073	0.09	2.395	0.151	2.073	0.09	2.395	0.151	0.648	0.076	0.6	0.05	2.073	0.09	2.395	0.151
	3j	–	–	–	–	1.089	0.026	1.473	0.056	1.089	0.026	1.473	0.056	0.311	0.017	0.452	0.041	1.089	0.026	1.473	0.056
	HP	1.648	0.051	2.231	0.122	–	–	–	–	0.265	0.025	0.313	0.031	1.648	0.051	2.231	0.122	0.524	0.027	0.654	0.045
	LP	0.88	0.015	1.36	0.046	–	–	–	–	0.068	0.005	0.103	0.011	0.88	0.015	1.36	0.046	0.632	0.012	0.899	0.032
[600, +∞(GeV	2j	–	–	–	–	–	–	–	–	0.574	0.098	0.426	0.042	0.297	0.115	0.164	0.027	0.337	0.085	0.249	0.025
	3j	–	–	–	–	–	–	–	–	0.346	0.035	0.347	0.034	0.133	0.018	0.121	0.013	0.219	0.027	0.202	0.02
	HP	1.019	0.113	0.94	0.08	1.019	0.113	0.94	0.08	0.642	0.066	0.439	0.051	1.019	0.113	0.94	0.08	1.019	0.113	0.94	0.08
	LP	0.644	0.033	0.8	0.064	0.644	0.033	0.8	0.064	0.144	0.027	0.103	0.011	0.644	0.033	0.8	0.064	0.644	0.033	0.8	0.064
[150, 250] GeV	all	2.833	0.009	5.186	0.038	2.833	0.009	5.186	0.038	2.833	0.009	5.186	0.038	2.833	0.009	5.186	0.038	2.833	0.009	5.186	0.038
[250, 400] GeV	all	3.675	0.037	5.08	0.08	3.675	0.037	5.08	0.08	3.676	0.037	5.081	0.08	3.281	0.027	4.547	0.06	3.769	0.036	5.211	0.078
[400, 600] GeV	all	1.868	0.045	2.613	0.107	2.341	0.081	2.811	0.132	2.357	0.08	2.831	0.131	2.002	0.049	2.719	0.104	2.481	0.077	3.023	0.123
[600, +∞(GeV	all	1.205	0.097	1.234	0.074	1.205	0.097	1.234	0.074	0.94	0.077	0.711	0.043	1.249	0.098	1.251	0.073	1.271	0.095	1.275	0.072
[150, +∞(GeV	all	5.145	0.039	7.813	0.069	5.335	0.049	7.882	0.075	5.289	0.046	7.825	0.075	4.935	0.037	7.518	0.06	5.477	0.048	8.05	0.074

Table 6.7: Detail of the significances (S) and their statistical uncertainty (δS) extracted in the different signal regions using for the boosted topology the 70% WP leading 2 strategy.

			Strategies														
			Split 400			Split 600			Resolved priority			Boosted priority			Ideal		
p_T^V	Region	Process	yield	sig./bkg.	sig. eff.	yield	sig./bkg.	sig. eff.	yield	sig./bkg.	sig. eff.	yield	sig./bkg.	sig. eff.	yield	sig./bkg.	sig. eff.
[150, 250] GeV	2j	sig.	145.89	2.28 %	31.84 %	145.89	2.28 %	31.84 %	145.89	2.28 %	31.84 %	145.89	2.28 %	31.84 %	145.89	2.28 %	31.84 %
		bkg.	6393.08			6393.08			6393.08			6393.08			6393.08		
	3j	sig.	132.72	0.48 %	28.97 %	132.72	0.48 %	28.97 %	132.72	0.48 %	28.97 %	132.72	0.48 %	28.97 %	132.72	0.48 %	28.97 %
		bkg.	27490.00			27490.00			27490.00			27490.00			27490.00		
[250, 400] GeV	2j	sig.	45.76	9.52 %	24.69 %	45.76	9.52 %	24.69 %	45.76	9.52 %	24.69 %	16.98	6.39 %	9.16 %	45.76	9.52 %	24.69 %
		bkg.	480.58			480.58			480.58			265.57			480.58		
	3j	sig.	40.99	2.48 %	22.11 %	40.99	2.48 %	22.11 %	40.99	2.48 %	22.11 %	17.41	1.59 %	9.39 %	40.99	2.48 %	22.11 %
		bkg.	1654.03			1654.03			1654.03			1097.81			1654.03		
	HP	sig.	–	–	–	–	–	–	0.47	0.38 %	0.25 %	55.35	3.00 %	29.86 %	18.71	1.23 %	10.09 %
		bkg.	–			–			122.25			1844.90			1520.58		
	LP	sig.	–	–	–	–	–	–	0.19	0.25 %	0.10 %	38.20	0.74 %	20.61 %	26.96	0.54 %	14.55 %
		bkg.	–			–			75.96			5133.36			4974.28		
[400, 600] GeV	2j	sig.	–	–	–	9.04	15.74 %	21.00 %	9.04	15.74 %	21.00 %	1.23	10.54 %	2.85 %	9.04	15.74 %	21.00 %
		bkg.	–			57.46			57.46			11.66			57.46		
	3j	sig.	–	–	–	7.37	4.84 %	17.11 %	7.37	4.84 %	17.11 %	1.00	2.07 %	2.32 %	7.37	4.84 %	17.11 %
		bkg.	–			152.14			152.14			48.47			152.14		
	HP	sig.	16.21	4.06 %	37.64 %	–	–	–	0.93	1.96 %	2.16 %	16.21	4.06 %	37.64 %	6.84	2.02 %	15.89 %
		bkg.	399.20			–			47.65			399.20			339.66		
	LP	sig.	14.02	1.34 %	32.56 %	–	–	–	0.20	1.03 %	0.46 %	14.02	1.34 %	32.56 %	10.76	1.06 %	24.98 %
		bkg.	1044.53			–			18.99			1044.53			1017.34		
[600, +∞(GeV	2j	sig.	–	–	–	–	–	–	0.72	13.55 %	8.54 %	0.17	17.00 %	1.96 %	0.25	8.67 %	3.01 %
		bkg.	–			–			5.34			0.98			2.93		
	3j	sig.	–	–	–	–	–	–	0.68	3.66 %	8.02 %	0.16	2.54 %	1.84 %	0.30	2.12 %	3.49 %
		bkg.	–			–			18.57			6.14			13.91		
	HP	sig.	3.17	6.77 %	37.48 %	3.17	6.77 %	37.48 %	1.43	15.34 %	16.88 %	3.17	6.77 %	37.48 %	3.17	6.77 %	37.48 %
		bkg.	46.87			46.87			9.32			46.87			46.87		
	LP	sig.	3.20	2.28 %	37.76 %	3.20	2.28 %	37.76 %	0.22	7.74 %	2.64 %	3.20	2.28 %	37.76 %	3.20	2.28 %	37.76 %
		bkg.	140.07			140.07			2.89			140.07			140.07		
[150, 250] GeV	all	sig.	278.61	0.82 %	60.81 %	278.61	0.82 %	60.81 %	278.61	0.82 %	60.81 %	278.61	0.82 %	60.81 %	278.61	0.82 %	60.81 %
		bkg.	33883.08			33883.08			33883.08			33883.08			33883.08		
[250, 400] GeV	all	sig.	86.75	4.06 %	46.80 %	86.75	4.06 %	46.80 %	87.41	3.75 %	47.15 %	127.94	1.53 %	69.02 %	132.42	1.53 %	71.44 %
		bkg.	2134.62			2134.62			2332.83			8341.64			8629.47		
[400, 600] GeV	all	sig.	30.23	2.09 %	70.20 %	16.41	7.83 %	38.12 %	17.54	6.35 %	40.73 %	32.46	2.16 %	75.38 %	34.02	2.17 %	78.99 %
		bkg.	1443.73			209.60			276.23			1503.87			1566.60		
[600, +∞(GeV	all	sig.	6.37	3.41 %	75.25 %	6.37	3.41 %	75.25 %	3.06	8.46 %	36.09 %	6.69	3.45 %	79.05 %	6.92	3.40 %	81.74 %
		bkg.	186.94			186.94			36.12			194.06			203.79		

Table 6.8: Signal and background yields and signal efficiencies for the different strategies in the different signal regions using for the boosted topology the 85% WP leading 2 strategy. The signal efficiencies are computed per p_{T}^{V}-region and correspond to the ratio: number of signal events passing all selections for a given signal region over the total number of events (resolved and boosted combined) passing the kinematic selections (SR and CRs combined without b-tagging requirement) described in Table 5.3. In this table *sig.* and *bkg.* stand for signal and background.

			Strategies														
			Split 400			Split 600			Resolved priority			Boosted priority			Ideal		
p_T^V	Region	Process	yield	sig./bkg.	sig. eff.	yield	sig./bkg.	sig. eff.	yield	sig./bkg.	sig. eff.	yield	sig./bkg.	sig. eff.	yield	sig./bkg.	sig. eff.
[150, 250] GeV	2j	sig.	145.89	2.28 %	32.21 %	145.89	2.28 %	32.21 %	145.89	2.28 %	32.21 %	145.89	2.28 %	32.21 %	145.89	2.28 %	32.21 %
		bkg.	6393.08			6393.08			6393.08			6393.08			6393.08		
	3j	sig.	132.72	0.48 %	29.31 %	132.72	0.48 %	29.31 %	132.72	0.48 %	29.31 %	132.72	0.48 %	29.31 %	132.72	0.48 %	29.31 %
		bkg.	27490.00			27490.00			27490.00			27490.00			27490.00		
[250, 400] GeV	2j	sig.	45.76	9.52 %	28.12 %	45.76	9.52 %	28.12 %	45.76	9.52 %	28.12 %	16.98	6.39 %	10.43 %	45.76	9.52 %	28.12 %
		bkg.	480.58			480.58			480.58			265.57			480.58		
	3j	sig.	40.99	2.48 %	25.19 %	40.99	2.48 %	25.19 %	40.99	2.48 %	25.19 %	17.41	1.59 %	10.70 %	40.99	2.48 %	25.19 %
		bkg.	1654.03			1654.03			1654.03			1097.81			1654.03		
	HP	sig.	–	–	–	–	–	–	0.32	0.65 %	0.20 %	36.75	7.02 %	22.59 %	5.74	1.98 %	3.53 %
		bkg.	–			–			49.32			523.55			290.18		
	LP	sig.	–	–	–	–	–	–	0.27	0.28 %	0.17 %	26.23	1.61 %	16.12 %	16.56	1.10 %	10.17 %
		bkg.	–			–			96.17			1631.21			1510.80		
[400, 600] GeV	2j	sig.	–	–	–	9.04	15.74 %	25.93 %	9.04	15.74 %	25.93 %	1.23	10.54 %	3.52 %	9.04	15.74 %	25.93 %
		bkg.	–			57.46			57.46			11.66			57.46		
	3j	sig.	–	–	–	7.37	4.84 %	21.13 %	7.37	4.84 %	21.13 %	1.00	2.07 %	2.87 %	7.37	4.84 %	21.13 %
		bkg.	–			152.14			152.14			48.47			152.14		
	HP	sig.	10.25	8.57 %	29.39 %	–	–	–	0.55	2.73 %	1.59 %	10.25	8.57 %	29.39 %	2.54	3.24 %	7.29 %
		bkg.	119.68			–			20.26			119.68			78.51		
	LP	sig.	9.28	2.69 %	26.60 %	–	–	–	0.17	0.92 %	0.50 %	9.28	2.69 %	26.60 %	6.54	2.01 %	18.75 %
		bkg.	345.00			–			18.76			345.00			324.99		
[600, +∞(GeV	2j	sig.	–	–	–	–	–	–	0.72	13.55 %	12.27 %	0.17	17.00 %	2.82 %	0.36	10.50 %	6.14 %
		bkg.	–			–			5.34			0.98			3.45		
	3j	sig.	–	–	–	–	–	–	0.68	3.66 %	11.52 %	0.16	2.54 %	2.65 %	0.37	2.44 %	6.33 %
		bkg.	–			–			18.57			6.14			15.33		
	HP	sig.	1.80	11.93 %	30.58 %	1.80	11.93 %	30.58 %	0.80	26.27 %	13.51 %	1.80	11.93 %	30.58 %	1.80	11.93 %	30.58 %
		bkg.	15.11			15.11			3.03			15.11			15.11		
	LP	sig.	1.91	3.90 %	32.44 %	1.91	3.90 %	32.44 %	0.14	6.33 %	2.30 %	1.91	3.90 %	32.44 %	1.91	3.90 %	32.44 %
		bkg.	49.04			49.04			2.14			49.04			49.04		
[150, 250] GeV	all	sig.	278.61	0.82 %	61.52 %	278.61	0.82 %	61.52 %	278.61	0.82 %	61.52 %	278.61	0.82 %	61.52 %	278.61	0.82 %	61.52 %
		bkg.	33883.08			33883.08			33883.08			33883.08			33883.08		
[250, 400] GeV	all	sig.	86.75	4.06 %	53.31 %	86.75	4.06 %	53.31 %	87.34	3.83 %	53.67 %	97.37	2.77 %	59.83 %	109.05	2.77 %	67.01 %
		bkg.	2134.62			2134.62			2280.10			3518.15			3935.60		
[400, 600] GeV	all	sig.	19.53	4.20 %	55.99 %	16.41	7.83 %	47.06 %	17.14	6.89 %	49.14 %	21.76	4.15 %	62.38 %	25.50	4.16 %	73.10 %
		bkg.	464.68			209.60			248.62			524.81			613.09		
[600, +∞(GeV	all	sig.	3.72	5.79 %	63.02 %	3.72	5.79 %	63.02 %	2.33	8.03 %	39.60 %	4.04	5.67 %	68.48 %	4.45	5.37 %	75.49 %
		bkg.	64.15			64.15			29.08			71.27			82.93		

Table 6.9: Signal and background yields and signal efficiencies for the different strategies in the different signal regions using for the boosted topology the 70% WP leading 2 strategy. The signal efficiencies are computed per p_{T}^{V}-region and correspond to the ratio: number of signal events passing all selections for a given signal region over the total number of events (resolved and boosted combined) passing the kinematic selections (SR and CRs combined without b-tagging requirement) described in Table 5.3. In this table *sig.* and *bkg.* stand for signal and background.

6.2.4 Conclusion of the combination study

The split at 400 GeV appears to be a simple yet effective strategy for the 1-lepton channel: events are reconstructed with the resolved topology for $p_{\mathrm{T}}^{V} < 400$ GeV and boosted topology above. The expected kinematic properties of the Higgs decay and the obtained sensitivity associated to the resolved and boosted topologies are in adequacy with that strategy. More complex strategies either favoring the resolved or boosted regimes based on kinematic criteria have been explored but raised a negligible increase of the expected significance and are hence not of interest.

A similar combination study in the 0-lepton channel [344] led to similar results. As a consequence, the ATLAS $VH(b\bar{b}/c\bar{c})$ analysis team decided to adopt the split at 400 GeV strategy for the Legacy Run 2 publication in the three lepton channels.

That combination study also emphasizes the large improvement obtained when using the multivariate approach in the boosted topology compared to the mass approach which was used for that regime in the previous publications. The boosted decision trees are benefitting from the information of the internal structure of the large-R jets to discriminate signal from background events. In particular in the 1-lepton channel the MVA is able to grasp the differences between $VH, H \to b\bar{b}$ signal events and the $t\bar{t}$ and $W + \mathrm{jets}$ background events.

It has also been shown that the impact of the transformation D hyperparameters used is minor. However, they should be kept reasonably low to guarantee having enough statistics in each bin of the transformed BDT output distribution to reduce statistical uncertainties. Moreover, the switch to the looser 85% WP leading 3 strategy for the boosted event reconstruction has increased the signal and background acceptances without a loss of performances in terms of sensitivity with respect to the old 70% WP leading 2 strategy. Such increase in statistics is of high importance for fits in order to properly normalize all background processes by means of control regions and reduce statistical uncertainties.

6.3 Conclusion

The conclusions of the two studies presented in this chapter can be found in Section 6.1.5 for the resolved 0-lepton MVA training and in Section 6.2.4 for the combination of the resolved and boosted topologies optimization in the 1-lepton channel. The work that I have performed both helped the $VH(b\bar{b}/c\bar{c})$ analysis team to decide which combination strategy was going to be used for the Legacy Run 2 publication and the resolved $VH, H \to b\bar{b}$ 0-lepton BDT training was and is still currently used by several persons of the analysis as for instance for the combination strategy that was carried out in the 0-lepton channel.

Here I would like to emphasize one aspect of the MVA that has not been mentioned yet which is really important. The input variables provided to the MVA should be well modelled in simulation and agree with data otherwise the response of the BDT in data would be different to simulation and could bias the results. All the input variables used by the BDTs for the $VH, H \to b\bar{b}$ analysis have been checked by means of data versus simulation comparison figures to be in good agreement.

Finally, in the future potentially better performances could be achieved using more advanced machine learning techniques such as deep neural networks.

General conclusion

In this thesis the physics of b-jets with the full ATLAS Run 2 was studied focusing on two aspects: the calibration of the energy scale of b-jets and the study of the $H \to b\bar{b}$ decay thanks to the VH production mode which offers to date the highest sensitivity to the Higgs-to-bottom Yukawa coupling. In addition to those studies, as highlighted in the Introduction of this thesis, I have been strongly involved in the $VH(b\bar{b}/c\bar{c})$ analysis team from a technical point of view since I have been responsible for the analysis framework since a bit more than one year now (October 2021) and thus was part of the analysis coordination team.

For the b-JES calibration, a feasibility study, fully based on my own work with an analysis technique and a corresponding software framework that I developed *ex nihilo*, has been successfully conducted using $t\bar{t}$ lepton+jets events and led to an inclusive and, for the first time, a differential measurement as a function of the transverse momentum of b-tagged jets. The specific b-tagged jet energy correction in data is found to be consistent with unity, with an uncertainty of 2.5% for the inclusive measurement, and decreasing from 4% to 1.9% for b-tagged jet transverse momenta increasing between 30 and 500 GeV for the differential measurement. The results have been published in a dedicated ATLAS conference note and presented at the Moriond Electroweak 2022 conference during a YSF talk session. This measurement could in the future be improved thanks to the better performances of the b-tagger developed within ATLAS, moreover some theoretical uncertainties such as matrix element and parton showering which are leading systematic uncertainties could be reduced. The measurement could also be performed in 2-dimensions as a function of the transverse momentum and the pseudorapidity of b-jets. It could also be repeated determining simultaneously the top quark mass and the b-JES which would allow not taking into account the top quark mass uncertainty which is also a dominant uncertainty in all p_{T}^{b}-regions. Such measurement paved the way to a future integration of the b-JES calibration to any ATLAS analysis as currently the jet calibration is measured flavour-inclusively in data and ATLAS only prescribes additional b-JES response uncertainties to account for potential discrepancies between light-, c-jets and b-jets energy response. Such dedicated correction of the energy of b-jet would in particular be of interest to all analyses whose final state involves one or several b-jets as for instance the $VH, H \to b\bar{b}$ analysis.

The second part of the thesis is about the study of the $VH, H \to b\bar{b}$ process targeting leptonic decays of the weak vector bosons as it allows for a significant reduction of the QCD multijet background. The corresponding ATLAS analysis is at an advanced and ambitious stage as it aims for a combination of 3 analyses the $VH, H \to b\bar{b}$ analysis (resolved and boosted topologies) with the $VH, H \to c\bar{c}$ analysis. A publication for early 2023 is foreseen, and significant improvement are expected both for the $VH, H \to b\bar{b}$ and $VH, H \to c\bar{c}$ sensitivities as the inclusive signal strength precision for the $VH, H \to b\bar{b}$ process is expected to be improved by 20% (expected result: $\mu_{VH}^{H\to b\bar{b}} = 1.00^{+0.15}_{-0.14}$) while the 95% CL upper limit for the $VH, H \to c\bar{c}$ signal strength should be reduced to 16 times the SM predictions which is about a factor 2 lower than the previous publication and is almost reaching the recent observed limit of 14.4 set by CMS, the expected limit being 7.6. The main reasons for this increase of sensitivity are the use of b- and c-taggers with higher performances and of the extension of the MVA approach to the full analysis as previously only the $VH, H \to b\bar{b}$ resolved analysis was relying on boosted decision trees. My personal work was focused on the $VH, H \to b\bar{b}$ analysis. In particular, I took part in the multivariate analysis studies being in charge of the optimization of the multivariate discriminant (hyperparameters of the training procedure and choice of the input variables) for the $VH, H \to b\bar{b}$ resolved regime in the 0-lepton channel. The MVA training I performed is going to be used for the incoming publication. The main improvements found have been proven to be due to the new DL1r b-tagger algorithm which achieves a better light- and c-jet rejections compared to the MV2c10 algorithm.

For the 1-lepton channel, I performed an optimization study for the combination of the resolved and boosted topologies exploring more complex strategies than previously probed in former publications. This study was of high interest since for the first time the multivariate approach was extended to the boosted topology. Moreover, a previous publication proved that the combination of the two topologies is achieving better constraints for EFT interpretations than the standalone resolved only or boosted only measurements. Based on my work, since the improvement obtained with more sophisticated strategies was not significant, the analysis team decided to keep a simple division of phase space reconstructing events only with the resolved topology for $p_{\mathrm{T}}^{V} < 400\ \mathrm{GeV}$ and only with the boosted topology for $p_{\mathrm{T}}^{V} > 400\ \mathrm{GeV}$. I also studied the impact of the jet mass scale calibration for small-radius jets as it was affecting the mass distribution of the Higgs candidate in the resolved topology. Since the JMS calibration was found to worsen the truth versus reconstructed level jet invariant mass agreement, it was decided not to apply that calibration for the Legacy Run 2 publication. Finally, I was one of the three main editors of an ATLAS internal note which documents all the MVA studies performed for the $VH, H \to b\bar{b}/c\bar{c}$ Legacy Run 2 publication. This internal note is a major update with respect to the previous round since the MVA is now used in all regions of the $VH, H \to b\bar{b}/c\bar{c}$ analysis.

In the future and for different timescales the $VH, H \to b\bar{b}/c\bar{c}$ measurements presented could be further improved. First machine learning algorithms are playing a leading role both at performance level (flavour tagging, object identification, rejection of background, calibration...) and at analysis level for signal to background discrimination for instance. ATLAS is currently developing new algorithms which will allow for better calibration in particular of jets which are essential in the $VH, H \to b\bar{b}/c\bar{c}$ analysis since the invariant mass of the Higgs candidate is found to be the most discriminating variable. Better identification of objects and particles should also be achieved in the future with for example the new incoming b- and c-tagging algorithms which are already outperforming the current default DL1r algorithms. Recent results from CMS concerning the $VH, H \to c\bar{c}$ process have shown that significant improvements could largely surpass sensitivity projections and have opened the door for potentially reaching the level of precision required to measure the Higgs-to-charm coupling which would be an outstanding milestone for ATLAS and CMS as measuring the coupling of the Higgs to the second generation of fermions is very challenging at the LHC. Moreover, work on the theoretical side is needed and is already ongoing as the modelling uncertainties have also a large impact on the precision and could be reduced as mentioned early on. The incoming Run 3 of the LHC and runs of the HL-LHC also should benefit to the $VH, H \to b\bar{b}/c\bar{c}$ analysis as the statistical uncertainty is still a dominant source of uncertainty for those measurements. The increase of the amount of data should also allow for more precise determination of the b-tagging efficiency scale factors, improve the calibration of objects, and hopefully also tune the simulations which should reduce the modelling uncertainties. In few decades the FCC project could allow for much higher precision measurements in particular if the FCC-ee project is approved and built. In that case, an inclusive ZH measurement exploiting the recoil mass of the Z boson candidate would be possible and dedicated $ZH, H \to b\bar{b}/c\bar{c}$ processes measurement would be very precise thanks to the largely reduced QCD background compared to LHC data-taking conditions and the large amount of data collected, leading to a sub-percent level determination of the Higgs boson coupling to bottom quarks.

Bibliography

[1] ATLAS Collaboration. *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*. In: *Phys. Lett. B* 716 (2012), p. 1. DOI: 10.1016/j.physletb.2012.08.020. arXiv: 1207.7214 [hep-ex] (cit. on pp. 15, 19, 43, 45).

[2] CMS Collaboration. *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*. In: *Phys. Lett. B* 716 (2012), p. 30. DOI: 10.1016/j.physletb.2012.08.021. arXiv: 1207.7235 [hep-ex] (cit. on pp. 15, 19, 43).

[3] ATLAS Collaboration. *Observation of $H \to b\bar{b}$ decays and VH production with the ATLAS detector*. In: *Phys. Lett. B* 786 (2018), p. 59. DOI: 10.1016/j.physletb.2018.09.013. arXiv: 1808.08238 [hep-ex] (cit. on pp. 15, 44, 157, 158, 161, 164, 209).

[4] CMS Collaboration. *Observation of Higgs Boson Decay to Bottom Quarks*. In: *Phys. Rev. Lett.* 121 (2018), p. 121801. DOI: 10.1103/PhysRevLett.121.121801. arXiv: 1808.08242 [hep-ex] (cit. on pp. 15, 44, 157, 158, 222).

[5] ATLAS Collaboration. *Combination of measurements of Higgs boson production in association with a W or Z boson in the $b\bar{b}$ decay channel with the ATLAS experiment at $\sqrt{s} = 13$ TeV*. ATLAS-CONF-2021-051. 2021. URL: https://cds.cern.ch/record/2782535 (cit. on pp. 15, 158, 171, 177, 203, 204, 209, 210, 211, 212, 214, 215, 216, 217, 218, 219, 220, 221, 230, 235, 266, 267, 280).

[6] ATLAS Collaboration. *Direct constraint on the Higgs-charm coupling from a search for Higgs boson decays into charm quarks with the ATLAS detector*. In: *Eur. Phys. J. C* 82 (2022), 717. 58 p. DOI: 10.1140/epjc/s10052-022-10588-3. arXiv: 2201.11428. URL: https://cds.cern.ch/record/2800527 (cit. on pp. 15, 158, 169, 170, 180, 203, 209, 224, 225, 226, 227, 242, 249, 250).

[7] ATLAS Collaboration. *Energy scale calibration of b-tagged jets with ATLAS Run 2 data using $t\bar{t}$ lepton+jets events*. Tech. rep. Geneva: CERN, Mar. 2022. URL: https://cds.cern.ch/record/2803523 (cit. on pp. 16, 122).

[8] R. Bouquet and G. Marchiori. *b-jet energy scale calibration using $t\bar{t}$ lepton+jets events with the ATLAS Run 2 data*. In: (Mar. 2022). URL: https://cds.cern.ch/record/2804176 (cit. on pp. 16, 122).

[9] R. Bouquet. *Energy scale calibration of b-tagged jets with ATLAS Run 2 data using $t\bar{t}$ lepton+jets events*. Tech. rep. Geneva: CERN, May 2022. URL: https://cds.cern.ch/record/2808578 (cit. on pp. 16, 122).

[10] *Moriond 2022 Electroweak Interactions & Unified Theories conference*. URL: https://moriond.in2p3.fr/2022/EW/ (cit. on pp. 16, 122).

[11] *Athena MR 37405 (bug in the central code of ATLAS): TEvent::fill() Fix*. URL: https://gitlab.cern.ch/atlas/athena/-/merge_requests/37405 (cit. on p. 17).

[12] *ROOT issue 9939: Hadd super slow since TFileMerger modification*. URL: https://github.com/root-project/root/issues/9939 (cit. on p. 17).

[13] G. Arnison et al. *Experimental observation of isolated large transverse energy electrons with associated missing energy at $\sqrt{s}$ =540 GeV*. In: *Physics Letters B* 122.1 (1983), pp. 103–116. ISSN: 0370-2693. DOI: https://doi.org/10.1016/0370-2693(83)91177-2. URL: https://www.sciencedirect.com/science/article/pii/0370269383911772 (cit. on p. 19).

[14] G. Arnison et al. *Experimental observation of lepton pairs of invariant mass around 95 GeV/c^2 at the CERN SPS collider*. In: *Physics Letters B* 126.5 (1983), pp. 398–410. ISSN: 0370-2693. DOI: https://doi.org/10.1016/0370-2693(83)90188-0. URL: https://www.sciencedirect.com/science/article/pii/0370269383901880 (cit. on p. 19).

[15] P. Bagnaia et al. *Evidence for $Z^0 \to e^+e^-$ at the CERN pp collider*. In: *Physics Letters B* 129.1 (1983), pp. 130–140. ISSN: 0370-2693. DOI: https://doi.org/10.1016/0370-2693(83)90744-X. URL: https://www.sciencedirect.com/science/article/pii/037026938390744X (cit. on p. 19).

[16] CDF Collaboration. *Observation of Top Quark Production in $\overline{p}p$ Collisions with the Collider Detector at Fermilab*. In: *Phys. Rev. Lett.* 74 (14 Apr. 1995), pp. 2626–2631. DOI: 10.1103/PhysRevLett.74.2626. URL: https://link.aps.org/doi/10.1103/PhysRevLett.74.2626 (cit. on p. 19).

[17] DØ Collaboration. *Observation of the Top Quark*. In: *Phys. Rev. Lett.* 74 (14 Apr. 1995), pp. 2632–2637. DOI: 10.1103/PhysRevLett.74.2632. URL: https://link.aps.org/doi/10.1103/PhysRevLett.74.2632 (cit. on p. 19).

[18] M. Tanabashi et al. *Review of Particle Physics*. In: *Phys. Rev. D* 98 (3 Aug. 2018). DOI: 10.1103/PhysRevD.98.030001. URL: https://link.aps.org/doi/10.1103/PhysRevD.98.030001 (cit. on pp. 20, 21, 65, 66, 67, 96, 122).

[19] B. Pontecorvo. *Neutrino Experiments and the Problem of Conservation of Leptonic Charge*. In: *Zh. Eksp. Teor. Fiz.* 53 (1967), pp. 1717–1725. URL: http://www.jetp.ras.ru/cgi-bin/dn/e_026_05_0984.pdf (cit. on p. 20).

[20] M. Aker et al. *Direct neutrino-mass measurement with sub-electronvolt sensitivity*. In: *Nature Phys.* 18.2 (2022), pp. 160–166. DOI: 10.1038/s41567-021-01463-1. arXiv: 2105.08533 [hep-ex] (cit. on p. 20).

[21] Z. Maki, M. Nakagawa, and S. Sakata. *Remarks on the Unified Model of Elementary Particles*. In: *Progress of Theoretical Physics* 28.5 (Nov. 1962), pp. 870–880. ISSN: 0033-068X. DOI: 10.1143/PTP.28.870. eprint: https://academic.oup.com/ptp/article-pdf/28/5/870/5258750/28-5-870.pdf. URL: https://doi.org/10.1143/PTP.28.870 (cit. on p. 20).

[22] M. Kobayashi and T. Maskawa. *CP-Violation in the Renormalizable Theory of Weak Interaction*. In: *Progress of Theoretical Physics* 49.2 (Feb. 1973), pp. 652–657. ISSN: 0033-068X. DOI: 10.1143/PTP.49.652. eprint: https://academic.oup.com/ptp/article-pdf/49/2/652/5257692/49-2-652.pdf. URL: https://doi.org/10.1143/PTP.49.652 (cit. on p. 20).

[23] N. Cabibbo. *Unitary Symmetry and Leptonic Decays*. In: *Phys. Rev. Lett.* 10 (12 June 1963), pp. 531–533. DOI: 10.1103/PhysRevLett.10.531. URL: https://link.aps.org/doi/10.1103/PhysRevLett.10.531 (cit. on p. 20).

[24] Particle Data Group. *CKM Quark-Mixing Matrix*. 2020. URL: https://pdg.lbl.gov/2020/reviews/rpp2020-rev-ckm-matrix.pdf (cit. on p. 21).

[25] Cush. *Standard Model of Elementary Particles*. URL: https://en.wikipedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg (cit. on p. 22).

[26] E. Drexler. *Elementary particle interactions in the Standard Model*. URL: https://en.m.wikipedia.org/wiki/File:Elementary_particle_interactions_in_the_Standard_Model.png (cit. on p. 22).

[27] M. Thomson. *Modern Particle Physics*. Cambridge University Press, 2013. DOI: 10.1017/CBO9781139525367 (cit. on pp. 23, 24, 25).

[28] E. Noether. *Invariante Variationsprobleme*. ger. In: *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse* 1918 (1918), pp. 235–257. URL: http://eudml.org/doc/59024 (cit. on p. 23).

[29] E. Noether. *Invariant variation problems*. In: *Transport Theory and Statistical Physics* 1.3 (Jan. 1971), pp. 186–207. DOI: 10.1080/00411457108231446. URL: https://doi.org/10.1080%2F00411457108231446 (cit. on p. 23).

[30] P. A. M. Dirac. *The Quantum Theory of the Electron*. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 117.778 (1928), pp. 610–624. ISSN: 09501207. URL: http://www.jstor.org/stable/94981 (visited on 10/19/2022) (cit. on p. 24).

[31] J. C. Maxwell. *A Dynamical Theory of the Electromagnetic Field*. In: *Philosophical Transactions of the Royal Society of London* 155 (1865), pp. 459–512. ISSN: 02610523. URL: http://www.jstor.org/stable/108892 (visited on 10/19/2022) (cit. on p. 24).

[32] S. Mandelstam. *Feynman Rules for Electromagnetic and Yang-Mills Fields from the Gauge-Independent Field-Theoretic Formalism*. In: *Phys. Rev.* 175 (5 Nov. 1968), pp. 1580–1603. DOI: 10.1103/PhysRev.175.1580. URL: https://link.aps.org/doi/10.1103/PhysRev.175.1580 (cit. on p. 24).

[33] D. H. Perkins. *Introduction to High Energy Physics*. 4th ed. Cambridge University Press, 2000. DOI: 10.1017/CBO9780511809040 (cit. on pp. 25, 26, 28).

[34] S. L. Glashow. *Partial-symmetries of weak interactions*. In: *Nuclear Physics* 22.4 (1961), pp. 579–588. ISSN: 0029-5582. DOI: https://doi.org/10.1016/0029-5582(61)90469-2. URL: https://www.sciencedirect.com/science/article/pii/0029558261904692 (cit. on p. 26).

[35] S. Weinberg. *A Model of Leptons*. In: *Phys. Rev. Lett.* 19 (21 Nov. 1967), pp. 1264–1266. DOI: 10.1103/PhysRevLett.19.1264. URL: https://link.aps.org/doi/10.1103/PhysRevLett.19.1264 (cit. on p. 26).

[36] A. Salam. *Weak and Electromagnetic Interactions*. In: *Conf. Proc. C* 680519 (1968), pp. 367–377. DOI: 10.1142/9789812795915_0034 (cit. on p. 26).

[37] K. Nishijima. *Charge Independence Theory of V Particles**. In: *Progress of Theoretical Physics* 13.3 (Mar. 1955), pp. 285–304. ISSN: 0033-068X. DOI: 10.1143/PTP.13.285. eprint: https://academic.oup.com/ptp/article-pdf/13/3/285/5425869/13-3-285.pdf. URL: https://doi.org/10.1143/PTP.13.285 (cit. on p. 26).

[38] M. Gell-Mann. *The interpretation of the new particles as displaced charge multiplets*. In: *Nuovo Cim.* 4.S2 (1956), pp. 848–866. DOI: 10.1007/BF02748000 (cit. on p. 26).

[39] M. E. Peskin and D. V. Schroeder. *An introduction to quantum field theory*. Boulder, CO: Westview, 1995. URL: https://cds.cern.ch/record/257493 (cit. on p. 26).

[40] Particle Data Group. *Electroweak Model and Constraints on New Physics*. 2020. URL: https://pdg.lbl.gov/2020/reviews/rpp2020-rev-standard-model.pdf (cit. on p. 28).

[41] F. Englert and R. Brout. *Broken Symmetry and the Mass of Gauge Vector Mesons*. In: *Phys. Rev. Lett.* 13 (9 Aug. 1964), pp. 321–323. DOI: 10.1103/PhysRevLett.13.321. URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.321 (cit. on pp. 28, 43).

[42] P. Higgs. *Broken symmetries, massless particles and gauge fields*. In: *Physics Letters* 12.2 (1964), pp. 132–133. ISSN: 0031-9163. DOI: https://doi.org/10.1016/0031-9163(64)91136-9. URL: https://www.sciencedirect.com/science/article/pii/0031916364911369 (cit. on pp. 28, 43).

[43] P. W. Higgs. *Broken Symmetries and the Masses of Gauge Bosons*. In: *Phys. Rev. Lett.* 13 (16 Oct. 1964), pp. 508–509. DOI: 10.1103/PhysRevLett.13.508. URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.508 (cit. on pp. 28, 43).

[44] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble. *Global Conservation Laws and Massless Particles*. In: *Phys. Rev. Lett.* 13 (20 Nov. 1964), pp. 585–587. DOI: 10.1103/PhysRevLett.13.585. URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.585 (cit. on pp. 28, 43).

[45] I. Neutelings. URL: https://tikz.net/author/izaak/ (cit. on pp. 29, 64, 83, 113).

[46] J. Riebesell. URL: https://tikz.net/author/janosh/ (cit. on p. 29).

[47] Particle Data Group. *Status of Higgs Boson Physics*. 2020. URL: https://pdg.lbl.gov/2020/reviews/rpp2020-rev-higgs-boson.pdf (cit. on pp. 29, 30).

[48] J. C. Romão and J. P. Silva. *A RESOURCE FOR SIGNS AND FEYNMAN DIAGRAMS OF THE STANDARD MODEL*. In: *International Journal of Modern Physics A* 27.26 (Oct. 2012), p. 1230025. DOI: 10.1142/s0217751x12300256. URL: https://doi.org/10.1142%2Fs0217751x12300256 (cit. on p. 30).

[49] V. Blacus. *Running coupling constants*. URL: https://commons.wikimedia.org/wiki/File:Running_coupling_constants.svg (cit. on p. 32).

[50] G. Altarelli and G. Parisi. *Asymptotic freedom in parton language*. In: *Nuclear Physics B* 126.2 (1977), pp. 298–318. ISSN: 0550-3213. DOI: https://doi.org/10.1016/0550-3213(77)90384-4. URL: https://www.sciencedirect.com/science/article/pii/0550321377903844 (cit. on p. 33).

[51] J. J. Ethier and E. R. Nocera. *Parton Distributions in Nucleons and Nuclei*. In: *Annual Review of Nuclear and Particle Science* 70.1 (Oct. 2020), pp. 43–76. DOI: 10.1146/annurev-nucl-011720-042725. URL: https://doi.org/10.1146%2Fannurev-nucl-011720-042725 (cit. on p. 34).

[52] R. D. Ball et al. *Parton distributions from high-precision collider data*. In: *The European Physical Journal C* 77.10 (Oct. 2017). DOI: 10.1140/epjc/s10052-017-5199-5. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-017-5199-5 (cit. on p. 33).

[53] LHC Higgs Cross Section Working Group. *Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector*. In: (2016). DOI: 10.23731/CYRM-2017-002. arXiv: 1610.07922 [hep-ph] (cit. on pp. 35, 37, 46, 158, 224).

[54] N. Berger et al. *Simplified Template Cross Sections - Stage 1.1*. 2019. DOI: 10.48550/ARXIV.1906.02754. URL: https://arxiv.org/abs/1906.02754 (cit. on p. 35).

[55] S. Amoroso et al. *Les Houches 2019: Physics at TeV Colliders: Standard Model Working Group Report*. 2020. DOI: 10.48550/ARXIV.2003.01700. URL: https://arxiv.org/abs/2003.01700 (cit. on p. 35).

[56] LHC Higgs Working Group. *Fiducial And STXS*. URL: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHWGFiducialAndSTXS (cit. on p. 36).

[57] LHC Higgs Cross Section Working Group. *Handbook of LHC Higgs Cross Sections: 3. Higgs Properties*. In: *CERN-2013-004* (CERN, Geneva, 2013). DOI: 10.5170/CERN-2013-004. arXiv: 1307.1347 [hep-ph] (cit. on pp. 37, 46, 206, 224).

[58] ATLAS Collaboration. *A detailed map of Higgs boson interactions by the ATLAS experiment ten years after the discovery*. In: *Nature* 607 (July 2022), 52–59. 12 p. DOI: 10.1038/s41586-022-04893-w. arXiv: 2207.00092. URL: https://cds.cern.ch/record/2814946 (cit. on pp. 37, 43, 47, 48, 49, 50).

[59] ATLAS Collaboration. *Combined measurements of Higgs boson production and decay using up to 80 fb^{-1} of proton–proton collision data at $\sqrt{s} = 13$ TeV collected with the ATLAS experiment*. In: *Phys. Rev. D* 101 (2020), p. 012002. DOI: 10.1103/PhysRevD.101.012002. arXiv: 1909.02845 [hep-ex] (cit. on p. 37).

[60] J. Ellis, V. Sanz, and T. You. *Complete Higgs sector constraints on dimension-6 operators*. In: *JHEP* 07 (2014), p. 036. DOI: 10.1007/JHEP07(2014)036. arXiv: 1404.3667 [hep-ph] (cit. on pp. 37, 219).

[61] R. Contino et al. *Effective Lagrangian for a light Higgs-like scalar*. In: *JHEP* 07 (2013), p. 035. DOI: 10.1007/JHEP07(2013)035. arXiv: 1303.3876 [hep-ph] (cit. on pp. 37, 219).

[62] I. Brivio, Y. Jiang, and M. Trott. *The SMEFTsim package, theory and tools*. In: *JHEP* 12 (2017), p. 070. DOI: 10.1007/JHEP12(2017)070. arXiv: 1709.06492 [hep-ph] (cit. on pp. 37, 219, 220).

[63] B. Grzadkowski et al. *Dimension-six terms in the Standard Model Lagrangian*. In: *Journal of High Energy Physics* 2010.10 (Oct. 2010). DOI: 10.1007/jhep10(2010)085. URL: https://doi.org/10.1007%2Fjhep10%282010%29085 (cit. on p. 38).

[64] ATLAS Collaboration. *Combined effective field theory interpretation of Higgs boson and weak boson production and decay with ATLAS data and electroweak precision observables*. Tech. rep. Geneva: CERN, 2022. URL: http://cds.cern.ch/record/2816369 (cit. on pp. 39, 47, 50).

[65] ATLAS Collaboration. *Measurements of WH and ZH production in the $H \to b\bar{b}$ decay channel in pp collisions at 13 TeV with the ATLAS detector*. In: *Eur. Phys. J. C* 81 (2021), p. 178. DOI: 10.1140/epjc/s10052-020-08677-2. arXiv: 2007.02873 [hep-ex] (cit. on pp. 39, 130, 158, 161, 162, 164, 175, 176, 180, 193, 203, 209, 212, 214, 219, 220, 221, 227, 229, 230, 231, 232, 249, 266).

[66] J. Campbell et al. *Event Generators for High-Energy Physics Experiments*. Tech. rep. Mar. 2022. arXiv: 2203.11110. URL: https://cds.cern.ch/record/2807002 (cit. on p. 40).

[67] S. Höche. *Introduction to parton-shower event generators*. 2014. DOI: 10.48550/ARXIV.1411.4085. URL: https://arxiv.org/abs/1411.4085 (cit. on p. 41).

[68] B. Andersson et al. *Parton fragmentation and string dynamics*. In: *Physics Reports* 97.2 (1983), pp. 31–145. ISSN: 0370-1573. DOI: https://doi.org/10.1016/0370-1573(83)90080-7. URL: https://www.sciencedirect.com/science/article/pii/0370157383900807 (cit. on p. 41).

[69] G. Marchesini and B. Webber. *Simulation of QCD jets including soft gluon interference*. In: *Nuclear Physics B* 238.1 (1984), pp. 1–29. ISSN: 0550-3213. DOI: https://doi.org/10.1016/0550-3213(84)90463-2. URL: https://www.sciencedirect.com/science/article/pii/0550321384904632 (cit. on p. 41).

[70] A. Schälicke et al. *An event generator for particle production in high-energy collisions*. In: *Progress in Particle and Nuclear Physics* 53.1 (July 2004), pp. 329–338. DOI: 10.1016/j.ppnp.2004.02.031. URL: https://doi.org/10.1016%2Fj.ppnp.2004.02.031 (cit. on p. 42).

[71] R. Pasechnik and M. Šumbera. *Different Faces of Confinement*. In: *Universe* 7.9 (Sept. 2021), p. 330. DOI: 10.3390/universe7090330. URL: https://doi.org/10.3390%2Funiverse7090330 (cit. on p. 42).

[72] C. B. Duncan and P. Kirchgaeßer. *Kinematic strangeness production in cluster hadronization*. In: *The European Physical Journal C* 79.1 (Jan. 2019). DOI: 10.1140/epjc/s10052-019-6573-2. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-019-6573-2 (cit. on p. 42).

[73] T. Sjöstrand, S. Mrenna, and P. Skands. *A brief introduction to PYTHIA 8.1*. In: *Comput. Phys. Commun.* 178 (2008), pp. 852–867. DOI: 10.1016/j.cpc.2008.01.036. arXiv: 0710.3820 [hep-ph] (cit. on pp. 41, 128, 162, 163).

[74] T. Sjöstrand et al. *An introduction to PYTHIA 8.2*. In: *Comput. Phys. Commun.* 191 (2015), p. 159. DOI: 10.1016/j.cpc.2015.01.024. arXiv: 1410.3012 [hep-ph] (cit. on pp. 41, 127, 162, 163).

[75] T. Gleisberg et al. *Event generation with SHERPA 1.1*. In: *JHEP* 02 (2009), p. 007. DOI: 10.1088/1126-6708/2009/02/007. arXiv: 0811.4622 [hep-ph] (cit. on pp. 43, 162, 163, 164).

[76] E. Bothmann et al. *Event generation with Sherpa 2.2*. In: *SciPost Phys.* 7.3 (2019), p. 034. DOI: 10.21468/SciPostPhys.7.3.034. arXiv: 1905.09127 [hep-ph] (cit. on pp. 43, 162, 163).

[77] J. Bellm et al. *Herwig 7.0/Herwig++ 3.0 release note*. In: *Eur. Phys. J. C* 76.4 (2016), p. 196. DOI: 10.1140/epjc/s10052-016-4018-8. arXiv: 1512.01178 [hep-ph] (cit. on pp. 43, 128).

[78] J. Bellm et al. *Herwig 7.1 Release Note*. In: (2017). arXiv: 1705.06919 [hep-ph] (cit. on p. 43).

[79] S. Alioli et al. *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*. In: *JHEP* 06 (2010), p. 043. DOI: 10.1007/JHEP06(2010)043. arXiv: 1002.2581 [hep-ph] (cit. on pp. 43, 127, 162, 163).

[80] K. Hamilton et al. *Merging H/W/Z + 0 and 1 jet at NLO with no merging scale: a path to parton shower + NNLO matching*. In: *JHEP* 05 (2013), p. 082. DOI: 10.1007/JHEP05(2013)082. arXiv: 1212.4504 [hep-ph] (cit. on p. 43).

[81] 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*. In: *JHEP* 07 (2014), p. 079. DOI: 10.1007/JHEP07(2014)079. arXiv: 1405.0301 [hep-ph] (cit. on pp. 43, 128).

[82] ATLAS Collaboration. *The ATLAS Simulation Infrastructure*. In: *Eur. Phys. J. C* 70 (2010), p. 823. DOI: 10.1140/epjc/s10052-010-1429-9. arXiv: 1005.4568 [physics.ins-det] (cit. on pp. 43, 127, 162).

[83] GEANT4 Collaboration, S. Agostinelli, et al. *GEANT4 – a simulation toolkit*. In: *Nucl. Instrum. Meth. A* 506 (2003), p. 250. DOI: 10.1016/S0168-9002(03)01368-8 (cit. on pp. 43, 127, 162).

[84] P. W. Higgs. *Spontaneous Symmetry Breakdown without Massless Bosons*. In: *Phys. Rev.* 145 (4 May 1966), pp. 1156–1163. DOI: 10.1103/PhysRev.145.1156 (cit. on p. 43).

[85] T. W. B. Kibble. *Symmetry Breaking in Non-Abelian Gauge Theories*. In: *Phys. Rev.* 155 (5 Mar. 1967), pp. 1554–1561. DOI: 10.1103/PhysRev.155.1554. URL: https://link.aps.org/doi/10.1103/PhysRev.155.1554 (cit. on p. 43).

[86] CMS Collaboration. *A measurement of the Higgs boson mass in the diphoton decay channel*. In: *Phys. Lett. B* 805 (2020), p. 135425. DOI: 10.1016/j.physletb.2020.135425. arXiv: 2002.06398 [hep-ex] (cit. on p. 43).

[87] ATLAS Collaboration. *Measurement of the Higgs boson mass in the $H \to ZZ^* \to 4\ell$ decay channel using 139 fb^{-1} of $\sqrt{s} = 13$ TeV pp collisions recorded by the ATLAS detector at the LHC*. Tech. rep. Geneva: CERN, 2022. arXiv: 2207.00320. URL: https://cds.cern.ch/record/2814431 (cit. on pp. 43, 45).

[88] ATLAS and CMS Collaborations. *Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments*. In: *Phys. Rev. Lett.* 114 (2015), p. 191803. DOI: 10.1103/PhysRevLett.114.191803. arXiv: 1503.07589 [hep-ex] (cit. on pp. 43, 44, 163).

[89] CMS Collaboration. *A portrait of the Higgs boson by the CMS experiment ten years after the discovery*. In: *Nature* (July 2022), 26 p. DOI: 10.1038/s41586-022-04892-x. arXiv: 2207.00043. URL: https://cds.cern.ch/record/2814513 (cit. on pp. 43, 44).

[90] P. Zyla et al. *Review of Particle Physics*. In: *PTEP* 2020.8 (2020), p. 083C01. DOI: 10.1093/ptep/ptaa104 (cit. on pp. 43, 158).

[91] ATLAS Collaboration. *Measurements of the Higgs boson inclusive and differential fiducial cross sections in the 4ℓ decay channel at $\sqrt{s} = 13$ TeV*. In: *Eur. Phys. J. C* 80 (2020), p. 942. DOI: 10.1140/epjc/s10052-020-8223-0. arXiv: 2004.03969 [hep-ex] (cit. on p. 44).

[92] ATLAS Collaboration. *Higgs boson production cross-section measurements and their EFT interpretation in the 4ℓ decay channel at $\sqrt{s} = 13$ TeV with the ATLAS detector*. In: *Eur. Phys. J. C* 80 (2020), p. 957. DOI: 10.1140/epjc/s10052-020-8227-9. arXiv: 2004.03447 [hep-ex] (cit. on p. 44). Erratum: in: *Eur. Phys. J. C* 81 (2021), p. 29. DOI: 10.1140/epjc/s10052-020-08644-x.

[93] ATLAS Collaboration. *Measurement of the properties of Higgs boson production at $\sqrt{s} = 13$ TeV in the $H \to \gamma\gamma$ 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 (cit. on p. 44).

[94] ATLAS Collaboration. *Measurements of gluon-gluon fusion and vector-boson fusion Higgs boson production cross-sections in the $H \to WW^* \to e\nu\mu\nu$ decay channel in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*. In: *Phys. Lett. B* 789 (2019), p. 508. DOI: 10.1016/j.physletb.2018.11.064. arXiv: 1808.09054 [hep-ex] (cit. on p. 44).

[95] ATLAS Collaboration. *Cross-section measurements of the Higgs boson decaying into a pair of τ-leptons in proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*. In: *Phys. Rev. D* 99 (2019), p. 072001. DOI: 10.1103/PhysRevD.99.072001. arXiv: 1811.08856 [hep-ex] (cit. on p. 44).

[96] ATLAS Collaboration. *Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector*. In: *Phys. Lett. B* 784 (2018), p. 173. DOI: 10.1016/j.physletb.2018.07.035. arXiv: 1806.00425 [hep-ex] (cit. on p. 44).

[97] CMS Collaboration. *Measurements of production cross sections of the Higgs boson in the four-lepton final state in proton–proton collisions at $\sqrt{s} = 13$ TeV*. In: *Eur. Phys. J. C* 81 (2021), p. 488. DOI: 10.1140/epjc/s10052-021-09200-x. arXiv: 2103.04956 [hep-ex] (cit. on p. 44).

[98] CMS Collaboration. *Measurements of Higgs boson production cross sections and couplings in the diphoton decay channel at $\sqrt{s} = 13$ TeV*. In: *JHEP* 07 (2021), p. 027. DOI: 10.1007/JHEP07(2021)027. arXiv: 2103.06956 [hep-ex] (cit. on p. 44).

[99] CMS Collaboration. *Measurement of the inclusive and differential Higgs boson production cross sections in the leptonic WW decay mode at $\sqrt{s} = 13$ TeV*. In: *JHEP* 03 (2021), p. 003. DOI: 10.1007/JHEP03(2021)003. arXiv: 2007.01984 [hep-ex] (cit. on p. 44).

[100] CMS Collaboration. *Measurement of the inclusive and differential Higgs boson production cross sections in the decay mode to a pair of τ leptons in pp collisions at $\sqrt{s} = 13$ TeV*. In: (2021). arXiv: 2107.11486 [hep-ex] (cit. on p. 44).

[101] CMS Collaboration. *Observation of $t\bar{t}H$ Production*. In: *Phys. Rev. Lett.* 120 (2018), p. 231801. DOI: 10.1103/PhysRevLett.120.231801. arXiv: 1804.02610 [hep-ex] (cit. on p. 44).

[102] C. N. Yang. *Selection Rules for the Dematerialization of a Particle into Two Photons*. In: *Phys. Rev.* 77 (2 Jan. 1950), pp. 242–245. DOI: 10.1103/PhysRev.77.242. URL: https://link.aps.org/doi/10.1103/PhysRev.77.242 (cit. on p. 44).

[103] L. D. Landau. *On the angular momentum of a system of two photons*. In: *Dokl. Akad. Nauk SSSR* 60.2 (1948), pp. 207–209. DOI: 10.1016/B978-0-08-010586-4.50070-5 (cit. on p. 44).

[104] ATLAS Collaboration. *Study of the spin and parity of the Higgs boson in diboson decays with the ATLAS detector*. In: *Eur. Phys. J. C* 75 (2015), p. 476. DOI: 10.1140/epjc/s10052-015-3685-1. arXiv: 1506.05669 [hep-ex] (cit. on p. 44). Erratum: in: *Eur. Phys. J. C* 76 (2016), p. 152. DOI: 10.1140/epjc/s10052-016-3934-y.

[105] ATLAS Collaboration. *Measuring CP properties of Higgs boson interactions with τ leptons with the ATLAS detector*. Tech. rep. Geneva: CERN, 2022. URL: https://cds.cern.ch/record/2809728 (cit. on p. 44).

[106] ATLAS Collaboration. *Measurements of the Higgs boson inclusive and differential fiducial cross-sections in the diphoton decay channel with pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*. In: *JHEP* 2208 (2022), p. 027. DOI: 10.1007/JHEP08(2022)027. arXiv: 2202.00487. URL: https://cds.cern.ch/record/2800939 (cit. on p. 45).

[107] M. Carena, Z. Liu, and M. Riembau. *Probing the electroweak phase transition via enhanced di-Higgs boson production*. In: *Phys. Rev. D* 97 (9 May 2018), p. 095032. DOI: 10.1103/PhysRevD.97.095032. URL: https://link.aps.org/doi/10.1103/PhysRevD.97.095032 (cit. on p. 51).

[108] ATLAS Collaboration. *Constraining the Higgs boson self-coupling from single- and double-Higgs production with the ATLAS detector using pp collisions at $\sqrt{s} = 13$ TeV*. Tech. rep. Geneva: CERN, 2022. arXiv: 2211.01216. URL: https://cds.cern.ch/record/2839326 (cit. on pp. 51, 52).

[109] ATLAS Collaboration. *Search for invisible Higgs-boson decays in events with vector-boson fusion signatures using 139 fb^{-1} of proton-proton data recorded by the ATLAS experiment*. In: *JHEP* 2208 (2022), p. 104. DOI: 10.1007/JHEP08(2022)104. arXiv: 2202.07953. URL: https://cds.cern.ch/record/2801694 (cit. on p. 53).

[110] F. Zwicky. *On the Masses of Nebulae and of Clusters of Nebulae*. In: *The Astrophysical Journal* 86 (Oct. 1937), p. 217. DOI: 10.1086/143864 (cit. on p. 53).

[111] V. C. Rubin, J. Ford W. K., and N. Thonnard. *Rotational properties of 21 SC galaxies with a large range of luminosities and radii, from NGC 4605 (R=4kpc) to UGC 2885 (R=122kpc)*. In: *The Astrophysical Journal* 238 (June 1980), pp. 471–487. DOI: 10.1086/158003 (cit. on p. 53).

[112] E. Hubble. *A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae*. In: *Proceedings of the National Academy of Science* 15.3 (Mar. 1929), pp. 168–173. DOI: 10.1073/pnas.15.3.168 (cit. on p. 54).

[113] J. D. Wells. *Higgs naturalness and the scalar boson proliferation instability problem*. In: *Synthese* 194.2 (2017), pp. 477–490. DOI: 10.1007/s11229-014-0618-8. arXiv: 1603.06131 [hep-ph] (cit. on p. 55).

[114] L. R. Evans and P. Bryant. *LHC Machine*. In: *JINST* 3 (2008), S08001. 164 p. DOI: 10.1088/1748-0221/3/08/S08001. URL: https://cds.cern.ch/record/1129806 (cit. on p. 57).

[115] A. Breskin and R. Voss. *The CERN Large Hadron Collider: Accelerator and Experiments*. Geneva: CERN, 2009. URL: https://cds.cern.ch/record/1244506 (cit. on p. 57).

[116] A. Lopes and M. L. Perrey. *FAQ-LHC The guide*. Jan. 2022. URL: https://cds.cern.ch/record/2809109 (cit. on pp. 57, 59, 60).

[117] CERN website. URL: https://home.cern/ (cit. on p. 57).

[118] E. Mobs. *The CERN accelerator complex - August 2018. Complexe des accélérateurs du CERN - Août 2018*. In: (Aug. 2018). URL: https://cds.cern.ch/record/2636343 (cit. on p. 58).

[119] O. S. Brüning et al. *LHC Design Report*. CERN Yellow Reports: Monographs. Geneva: CERN, 2004. DOI: 10.5170/CERN-2004-003-V-1. URL: https://cds.cern.ch/record/782076 (cit. on pp. 58, 59).

[120] R. Steerenberg. *Batch Compression Merging and Splitting (BCMS). Batch Compression Merging and Splitting (BCMS)*. In: (Apr. 2017). URL: https://cds.cern.ch/record/2259071 (cit. on p. 59).

[121] R. Steerenberg. *LHC report: full house for the LHC*. In: (July 2017). URL: https://cds.cern.ch/record/2272573 (cit. on p. 59).

[122] A. A. Gorzawski. *Luminosity control and beam orbit stability with beta star leveling at LHC and HL-LHC*. 2016. DOI: 10.5075/epfl-thesis-7338. URL: https://cds.cern.ch/record/2238309 (cit. on p. 59).

[123] ATLAS Collaboration. *ATLAS data quality operations and performance for 2015–2018 data-taking*. In: *JINST* 15 (2020), P04003. DOI: 10.1088/1748-0221/15/04/P04003. arXiv: 1911.04632 [physics.ins-det] (cit. on pp. 59, 60, 127, 161).

[124] ATLAS Collaboration. *Measurement of the Inelastic Proton–Proton Cross Section at $\sqrt{s} = 13$ TeV with the ATLAS Detector at the LHC*. In: *Phys. Rev. Lett.* 117 (2016), p. 182002. DOI: 10.1103/PhysRevLett.117.182002. arXiv: 1606.02625 [hep-ex] (cit. on p. 59).

[125] S. Karppinen et al. *Design of Beampipes for LHC Experiments*. In: *Vacuum* 64 (July 2003), 476–473. 12 p. URL: https://cds.cern.ch/record/631307 (cit. on p. 60).

[126] H. Wiedemann. *Particle Accelerator Physics II, Nonlinear and Higher-Order Beam Dynamics*. Springer Berlin, Heidelberg. DOI: doi.org/10.1007/978-3-642-97550-9 (cit. on p. 61).

[127] K. Wille. *Synchrotron Radiation (lecture)*. URL: https://indico.cern.ch/event/218284/contributions/1520454/ (cit. on p. 61).

[128] ATLAS Collaboration. *The ATLAS Experiment at the CERN Large Hadron Collider*. In: *JINST* 3 (2008), S08003. DOI: 10.1088/1748-0221/3/08/S08003 (cit. on pp. 61, 62, 67, 68, 69, 70, 71, 72, 73, 75, 76).

[129] ATLAS Collaboration. *ATLAS: letter of intent for a general-purpose pp experiment at the large hadron collider at CERN*. In: (1992). URL: http://cds.cern.ch/record/291061 (cit. on pp. 61, 62).

[130] The ATLAS experiment website. URL: https://atlas.cern/ (cit. on pp. 61, 62).

[131] CMS Collaboration. *The CMS Experiment at the CERN LHC*. In: *JINST* 3 (2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004 (cit. on p. 62).

[132] CMS Collaboration. *CMS: letter of intent by the CMS Collaboration for a general purpose detector at LHC*. Tech. rep. Geneva: CERN, 1992. URL: https://cds.cern.ch/record/290808 (cit. on p. 62).

[133] The CMS experiment website. URL: https://cms.cern/ (cit. on p. 62).

[134] LHCb Collaboration. *The LHCb Detector at the LHC*. In: *JINST* 3 (2008), S08005. DOI: 10.1088/1748-0221/3/08/S08005. URL: https://cds.cern.ch/record/1129809 (cit. on p. 62).

[135] H. Dijkstra et al. *LHCb Letter of Intent, LHCb Collaboration*. Tech. rep. Geneva: CERN, Jan. 1995. URL: http://cds.cern.ch/record/691698 (cit. on p. 62).

[136] The LHCb experiment website. URL: https://lhcb-outreach.web.cern.ch/ (cit. on p. 62).

[137] Alice Collaboration. *The ALICE experiment at the CERN LHC. A Large Ion Collider Experiment*. In: *JINST* 3 (2008), S08002. 259 p. DOI: 10.1088/1748-0221/3/08/S08002. URL: https://cds.cern.ch/record/1129812 (cit. on p. 62).

[138] Alice Collaboration. *Letter of Intent for A Large Ion Collider Experiment [ALICE]*. Tech. rep. Geneva: CERN, 1993. URL: http://cds.cern.ch/record/290825 (cit. on p. 62).

[139] The ALICE experiment website. URL: https://alice-collaboration.web.cern.ch/ (cit. on p. 62).

[140] J. Pequenao. *Computer generated image of the whole ATLAS detector*. Mar. 2008. URL: https://cds.cern.ch/record/1095924 (cit. on p. 63).

[141] J. Pequenao and P. Schaffner. *How ATLAS detects particles: diagram of particle paths in the detector*. Jan. 2013. URL: https://cds.cern.ch/record/1505342 (cit. on p. 63).

[142] M. Gupta. *Calculation of radiation length in materials*. Tech. rep. Geneva: CERN, July 2010. URL: https://cds.cern.ch/record/1279627 (cit. on p. 65).

[143] M. Delmastro. *Particle interactions in particle detectors*. URL: https://indico.cern.ch/event/294651/contributions/671929/ (cit. on p. 67).

[144] E. Longo and I. Sestili. *Monte Carlo calculation of photon-initiated electromagnetic showers in lead glass*. In: *Nuclear Instruments and Methods* 128.2 (1975), pp. 283–307. ISSN: 0029-554X. DOI: https://doi.org/10.1016/0029-554X(75)90679-5. URL: https://www.sciencedirect.com/science/article/pii/0029554X75906795 (cit. on p. 67).

[145] Quantum Diaries. URL: https://www.quantumdiaries.org/wp-content/uploads/2011/05/exp-magnets.png (cit. on p. 68).

[146] A. M. Rodriguez Vera and J. Antunes Pequenao. *ATLAS Detector Magnet System*. May 2021. URL: https://cds.cern.ch/record/2770604 (cit. on p. 68).

[147] ATLAS Collaboration. *ATLAS inner detector: Technical Design Report, 1*. Technical design report. ATLAS. Geneva: CERN, 1997. URL: https://cds.cern.ch/record/331063 (cit. on p. 69).

[148] N. C. Benekos et al. *ATLAS Inner Detector Performance*. Tech. rep. Geneva: CERN, Dec. 2003. URL: https://cds.cern.ch/record/688762 (cit. on p. 69).

[149] J. Pequenao. *Computer generated image of the ATLAS inner detector*. Mar. 2008. URL: https://cds.cern.ch/record/1095926 (cit. on p. 70).

[150] ATLAS Collaboration. *Experiment Briefing: Keeping the ATLAS Inner Detector in perfect alignment*. July 2020. URL: https://cds.cern.ch/record/2723878 (cit. on pp. 70, 82).

[151] M. Capeans et al. *ATLAS Insertable B-Layer Technical Design Report*. Tech. rep. Sept. 2010. URL: https://cds.cern.ch/record/1291633 (cit. on pp. 70, 71).

[152] ATLAS Collaboration. *Modelling radiation damage to pixel sensors in the ATLAS detector*. In: *JINST* 14 (2019), P06012. DOI: 10.1088/1748-0221/14/06/P06012. arXiv: 1905.03739 [hep-ex] (cit. on pp. 71, 72).

[153] Y. Takubo. *ATLAS IBL operational experience*. Tech. rep. Geneva: CERN, Nov. 2016. DOI: 10.22323/1.287.0004. URL: https://cds.cern.ch/record/2235541 (cit. on p. 71).

[154] ATLAS Collaboration. *Efficiency and Hit Spatial Resolution of ATLAS IBL Sensors in LHC Run 2 Collision Events*. ATL-INDET-PUB-2016-001. 2016. URL: https://cds.cern.ch/record/2203893 (cit. on p. 71).

[155] C. Goessling et al. *Planar n+-in-n silicon pixel sensors for the ATLAS IBL upgrade*. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 650.1 (2011), pp. 198–201. ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2010.11.186. URL: https://www.sciencedirect.com/science/article/pii/S016890021002735X (cit. on p. 71).

[156] C. Da Via et al. *3D silicon sensors: Design, large area production and quality assurance for the ATLAS IBL pixel detector upgrade*. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 694 (2012), pp. 321–330. ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2012.07.058. URL: https://www.sciencedirect.com/science/article/pii/S0168900212008509 (cit. on p. 71).

[157] ATLAS Collaboration. *Operation and performance of the ATLAS semiconductor tracker in LHC Run 2*. In: (2021). arXiv: 2109.02591 [physics.ins-det] (cit. on p. 72).

[158] ATLAS Collaboration. *Performance of the ATLAS Transition Radiation Tracker in Run 1 of the LHC: tracker properties*. In: *JINST* 12 (2017), P05002. DOI: 10.1088/1748-0221/12/05/P05002. arXiv: 1702.06473 [hep-ex] (cit. on p. 72).

[159] A. Vogel. *ATLAS Transition Radiation Tracker (TRT): Straw Tube Gaseous Detectors at High Rates*. Tech. rep. Geneva: CERN, Apr. 2013. URL: https://cds.cern.ch/record/1537991 (cit. on p. 72).

[160] J. Pequenao. *Computer Generated image of the ATLAS calorimeter*. Mar. 2008. URL: https://cds.cern.ch/record/1095927 (cit. on p. 73).

[161] ATLAS Collaboration. *ATLAS liquid-argon calorimeter: Technical Design Report*. Technical design report. ATLAS. Geneva: CERN, 1996. DOI: 10.17181/CERN.FWRW.FOOQ. URL: https://cds.cern.ch/record/331061 (cit. on p. 73).

[162] ATLAS Collaboration. *Electron and photon energy calibration with the ATLAS detector using LHC Run 1 data*. In: *Eur. Phys. J. C* 74 (2014), p. 3071. DOI: 10.1140/epjc/s10052-014-3071-4. arXiv: 1407.5063 [hep-ex] (cit. on p. 74).

[163] J. Pequenao. *Computer generated image of the ATLAS Muons subsystem*. Mar. 2008. URL: https://cds.cern.ch/record/1095929 (cit. on p. 77).

[164] A. M. Rodriguez Vera et al. *Trigger and Data Acquisition - ATLAS Fact Sheet*. In: (July 2021). URL: https://cds.cern.ch/record/2775198 (cit. on p. 77).

[165] ATLAS Collaboration. *ATLAS High-Level Trigger, Data Acquisition and Controls: Technical Design Report*. ATLAS-TDR-16; CERN-LHCC-2003-022. 2003. URL: https://cds.cern.ch/record/616089 (cit. on p. 77).

[166] ATLAS Collaboration. *ATLAS Approved Plots DAQ*. URL: https://twiki.cern.ch/twiki/bin/view/AtlasPublic/ApprovedPlotsDAQ (cit. on pp. 77, 79).

[167] 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*. In: *Eur. Phys. J. C* 79 (2019), p. 639. DOI: 10.1140/epjc/s10052-019-7140-6. arXiv: 1902.04655 [hep-ex] (cit. on p. 82).

[168] G. Borissov et al. *ATLAS strategy for primary vertex reconstruction during Run-2 of the LHC*. Tech. rep. Geneva: CERN, May 2015. DOI: 10.1088/1742-6596/664/7/072041. URL: https://cds.cern.ch/record/2015220 (cit. on p. 82).

[169] W. Waltenberger, R. Frühwirth, and P. Vanlaer. *Adaptive vertex fitting*. In: *Journal of Physics G: Nuclear and Particle Physics* 34.12 (Nov. 2007), N343–N356. DOI: 10.1088/0954-3899/34/12/n01. URL: https://doi.org/10.1088/0954-3899/34/12/n01 (cit. on p. 82).

[170] ATLAS Collaboration. *Reconstruction of primary vertices at the ATLAS experiment in Run 1 proton–proton collisions at the LHC*. In: *Eur. Phys. J. C* 77 (2017), p. 332. DOI: 10.1140/epjc/s10052-017-4887-5. arXiv: 1611.10235 [hep-ex] (cit. on pp. 82, 129).

[171] ATLAS Collaboration. *Performance of the ATLAS track reconstruction algorithms in dense environments in LHC Run 2*. In: *Eur. Phys. J. C* 77 (2017), p. 673. DOI: 10.1140/epjc/s10052-017-5225-7. arXiv: 1704.07983 [hep-ex] (cit. on pp. 82, 85).

[172] ATLAS Collaboration. *ATLAS Tracking Software Tutorial*. Aug. 2022. URL: https://atlassoftwaredocs.web.cern.ch/trackingTutorial/index.html (cit. on pp. 82, 83).

[173] Y. Amhis et al. *Averages of b-hadron, c-hadron, and* τ*-lepton properties as of 2021*. Tech. rep. June 2022. arXiv: 2206.07501. URL: https://cds.cern.ch/record/2815545 (cit. on p. 82).

[174] T. G. Cornelissen et al. *Updates of the ATLAS Tracking Event Data Model (Release 13)*. Tech. rep. Geneva: CERN, June 2007. URL: https://cds.cern.ch/record/1038095 (cit. on p. 83).

[175] R. Frühwirth. *Application of Kalman filtering to track and vertex fitting*. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 262.2 (1987), pp. 444–450. ISSN: 0168-9002. DOI: https://doi.org/10.1016/0168-9002(87)90887-4. URL: https://www.sciencedirect.com/science/article/pii/0168900287908874 (cit. on p. 84).

[176] D. Stampfer, M. Regler, and R. Frühwirth. *Track fitting with energy loss*. In: *Computer Physics Communications* 79.2 (1994), pp. 157–164. ISSN: 0010-4655. DOI: https://doi.org/10.1016/0010-4655(94)90064-7. URL: https://www.sciencedirect.com/science/article/pii/0010465594900647 (cit. on p. 84).

[177] ATLAS Collaboration. *Impact Parameter Resolution*. URL: https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/IDTR-2015-007/ (cit. on p. 85).

[178] ATLAS Collaboration. *Electron and photon performance measurements with the ATLAS detector using the 2015–2017 LHC proton–proton collision data*. In: *JINST* 14 (2019), P12006. DOI: 10.1088/1748-0221/14/12/P12006. arXiv: 1908.00005 [hep-ex] (cit. on pp. 86, 88, 89, 90, 91, 92, 97).

[179] ATLAS Collaboration. *Electron and photon reconstruction and performance in ATLAS using a dynamical, topological cell clustering-based approach*. ATL-PHYS-PUB-2017-022. 2017. URL: https://cds.cern.ch/record/2298955 (cit. on pp. 86, 89, 91).

[180] ATLAS Collaboration. *Electron and photon energy calibration with the ATLAS detector using 2015–2016 LHC proton–proton collision data*. In: *JINST* 14 (2019), P03017. DOI: 10.1088/1748-0221/14/03/P03017. arXiv: 1812.03848 [hep-ex] (cit. on pp. 86, 89, 91).

[181] W. Lampl et al. *Calorimeter Clustering Algorithms: Description and Performance*. Tech. rep. Geneva: CERN, Apr. 2008. URL: https://cds.cern.ch/record/1099735 (cit. on p. 86).

[182] ATLAS Collaboration. *Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1*. In: *Eur. Phys. J. C* 77 (2017), p. 490. DOI: 10.1140/epjc/s10052-017-5004-5. arXiv: 1603.02934 [hep-ex] (cit. on pp. 86, 87, 106).

[183] ATLAS Collaboration. *Muon reconstruction and identification efficiency in ATLAS using the full Run 2 pp collision data set at* $\sqrt{s} = 13$ *TeV*. In: *Eur. Phys. J. C* 81 (2021), p. 578. DOI: 10.1140/epjc/s10052-021-09233-2. arXiv: 2012.00578 [hep-ex] (cit. on pp. 93, 94, 95, 96).

[184] ATLAS Collaboration. *Muon reconstruction performance of the ATLAS detector in proton–proton collision data at* $\sqrt{s} = 13$ *TeV*. In: *Eur. Phys. J. C* 76 (2016), p. 292. DOI: 10.1140/epjc/s10052-016-4120-y. arXiv: 1603.05598 [hep-ex] (cit. on pp. 93, 94).

[185] J. Illingworth and J. Kittler. *A survey of the hough transform*. In: *Computer Vision, Graphics, and Image Processing* 44.1 (1988), pp. 87–116. ISSN: 0734-189X. DOI: https://doi.org/10.1016/S0734-189X(88)80033-1. URL: https://www.sciencedirect.com/science/article/pii/S0734189X88800331 (cit. on p. 93).

[186] ATLAS Collaboration. *Measurement of the tau lepton reconstruction and identification performance in the ATLAS experiment using pp collisions at* $\sqrt{s} = 13$ *TeV*. ATLAS-CONF-2017-029. 2017. URL: https://cds.cern.ch/record/2261772 (cit. on p. 96).

[187] ATLAS Collaboration. *Reconstruction, Energy Calibration, and Identification of Hadronically Decaying Tau Leptons in the ATLAS Experiment for Run-2 of the LHC*. ATL-PHYS-PUB-2015-045. 2015. URL: https://cds.cern.ch/record/2064383 (cit. on p. 97).

[188] 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 (cit. on p. 97).

[189] M. Cacciari, G. P. Salam, and G. Soyez. *The anti-*k_t *jet clustering algorithm*. In: *JHEP* 04 (2008), p. 063. DOI: 10.1088/1126-6708/2008/04/063. arXiv: 0802.1189 [hep-ph] (cit. on pp. 98, 99, 130).

[190] G. P. Salam. *Towards jetography*. In: *The European Physical Journal C* 67.3-4 (May 2010), pp. 637–686. DOI: 10.1140/epjc/s10052-010-1314-6. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-010-1314-6 (cit. on pp. 98, 158).

[191] J. M. Butterworth et al. *Jet Substructure as a New Higgs-Search Channel at the Large Hadron Collider*. In: *Phys. Rev. Lett.* 100 (24 June 2008), p. 242001. DOI: 10.1103/PhysRevLett.100.242001. URL: https://link.aps.org/doi/10.1103/PhysRevLett.100.242001 (cit. on pp. 98, 158).

[192] ATLAS Collaboration. *Jet reconstruction and performance using particle flow with the ATLAS Detector*. In: *Eur. Phys. J. C* 77 (2017), p. 466. DOI: 10.1140/epjc/s10052-017-5031-2. arXiv: 1703.10485 [hep-ex] (cit. on p. 100).

[193] ATLAS Collaboration. *Jet energy scale and resolution measured in proton–proton collisions at* $\sqrt{s} =$ 13 *TeV with the ATLAS detector*. In: *Eur. Phys. J. C* 81 (2020), p. 689. DOI: 10.1140/epjc/s10052-021-09402-3. arXiv: 2007.02645 [hep-ex] (cit. on pp. 100, 101, 102, 103, 104, 105, 106, 139).

[194] 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 (cit. on pp. 100, 117, 119).

[195] W. Lampl et al. *Calorimeter Clustering Algorithms: Description and Performance*. ATL-LARG-PUB-2008-002. 2008. URL: https://cds.cern.ch/record/1099735 (cit. on p. 100).

[196] ATLAS Collaboration. *Jet energy scale measurements and their systematic uncertainties in proton–proton collisions at* $\sqrt{s} = 13$ *TeV with the ATLAS detector*. In: *Phys. Rev. D* 96 (2017), p. 072002. DOI: 10.1103/PhysRevD.96.072002. arXiv: 1703.09665 [hep-ex] (cit. on pp. 100, 123, 139).

[197] ATLAS Collaboration. *In situ calibration of large-radius jet energy and mass in* 13 *TeV proton–proton collisions with the ATLAS detector*. In: *Eur. Phys. J. C* 79 (2019), p. 135. DOI: 10.1140/epjc/s10052-019-6632-8. arXiv: 1807.09477 [hep-ex] (cit. on pp. 106, 108, 109).

[198] ATLAS Collaboration. *Optimisation of large-radius jet reconstruction for the ATLAS detector in* 13 *TeV proton–proton collisions*. In: *Eur. Phys. J. C* 81 (2020), p. 334. DOI: 10.1140/epjc/s10052-021-09054-3. arXiv: 2009.04986 [hep-ex] (cit. on p. 106).

[199] ATLAS Collaboration. *Performance of jet substructure techniques for large-R jets in proton–proton collisions at* $\sqrt{s} = 7$ *TeV using the ATLAS detector*. In: *JHEP* 09 (2013), p. 076. DOI: 10.1007/JHEP09(2013)076. arXiv: 1306.4945 [hep-ex] (cit. on p. 107).

[200] D. Krohn, J. Thaler, and L.-T. Wang. *Jet trimming*. In: *Journal of High Energy Physics* 2010.2 (Feb. 2010). DOI: 10.1007/jhep02(2010)084. URL: https://doi.org/10.1007%2Fjhep02%282010%29084 (cit. on p. 107).

[201] S. D. Ellis and D. E. Soper. *Successive combination jet algorithm for hadron collisions*. In: *Physical Review D* 48.7 (Oct. 1993), pp. 3160–3166. DOI: 10.1103/physrevd.48.3160. URL: https://doi.org/10.1103%2Fphysrevd.48.3160 (cit. on p. 107).

[202] ATLAS Collaboration. *Variable Radius, Exclusive-k_T, and Center-of-Mass Subjet Reconstruction for Higgs($\to b\bar{b}$) Tagging in ATLAS*. ATL-PHYS-PUB-2017-010. 2017. URL: https://cds.cern.ch/record/2268678 (cit. on pp. 110, 111).

[203] D. Krohn, J. Thaler, and L.-T. Wang. *Jets with variable R*. In: *Journal of High Energy Physics* 2009.06 (June 2009), pp. 059–059. DOI: 10.1088/1126-6708/2009/06/059. URL: https://doi.org/10.1088/1126-6708/2009/06/059 (cit. on p. 110).

[204] M. Cacciari, G. P. Salam, and G. Soyez. *The Catchment Area of Jets*. In: *JHEP* 04 (2008), p. 005. DOI: 10.1088/1126-6708/2008/04/005. arXiv: 0802.1188 [hep-ph] (cit. on p. 110).

[205] A. Hoecker et al. *TMVA - Toolkit for Multivariate Data Analysis*. 2007. arXiv: physics/0703039 [physics.data-an] (cit. on pp. 112, 190, 191).

[206] 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 (cit. on p. 112).

[207] ATLAS Collaboration. *ATLAS b-jet identification performance and efficiency measurement with $t\bar{t}$ events in pp collisions at $\sqrt{s} = 13$ TeV*. In: *Eur. Phys. J. C* 79 (2019), p. 970. DOI: 10.1140/epjc/s10052-019-7450-8. arXiv: 1907.05120 [hep-ex] (cit. on pp. 112, 115, 130).

[208] P. J. Windischhofer. *Heavy-flavour jet tagging in ATLAS*. In: (Jan. 2020). URL: https://cds.cern.ch/record/2706702 (cit. on pp. 112, 114).

[209] ATLAS Collaboration. *Expected performance of the 2019 ATLAS b-taggers*. URL: http://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/FTAG-2019-005 (cit. on pp. 112, 114, 175).

[210] ATLAS Collaboration. *Performance of 2019 recommendations of ATLAS Flavor Tagging algorithms with Variable Radius track jets*. URL: http://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/FTAG-2019-006 (cit. on pp. 112, 114, 177).

[211] ATLAS Collaboration. *Measurement of the c-jet mistagging efficiency in $t\bar{t}$ events using pp collision data at $\sqrt{s} = 13$ TeV collected with the ATLAS detector*. In: (2021). arXiv: 2109.10627 [hep-ex] (cit. on p. 112).

[212] ATLAS Collaboration. *Identification of Boosted Higgs Bosons Decaying Into $b\bar{b}$ With Neural Networks and Variable Radius Subjets in ATLAS*. ATL-PHYS-PUB-2020-019. 2020. URL: https://cds.cern.ch/record/2724739 (cit. on pp. 112, 246, 248).

[213] 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 (cit. on p. 112).

[214] 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 (cit. on p. 112).

[215] 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 (cit. on p. 113).

[216] *Illustrating the random forest algorithm in TikZ*. URL: https://tex.stackexchange.com/questions/503883/illustrating-the-random-forest-algorithm-in-tikz (cit. on p. 113).

[217] ATLAS Collaboration. *Constituent-level pile-up mitigation techniques in ATLAS*. ATLAS-CONF-2017-065. 2017. URL: https://cds.cern.ch/record/2281055 (cit. on p. 116).

[218] ATLAS Collaboration. *Tagging and suppression of pileup jets with the ATLAS detector*. ATLAS-CONF-2014-018. 2014. URL: https://cds.cern.ch/record/1700870 (cit. on p. 116).

[219] ATLAS Collaboration. *Performance of pile-up mitigation techniques for jets in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector*. In: *Eur. Phys. J. C* 76 (2016), p. 581. DOI: 10.1140/epjc/s10052-016-4395-z. arXiv: 1510.03823 [hep-ex] (cit. on pp. 116, 130).

[220] ATLAS Collaboration. *Forward jet vertex tagging using the particle flow algorithm*. ATL-PHYS-PUB-2019-026. 2019. URL: https://cds.cern.ch/record/2683100 (cit. on p. 116).

[221] D. Adams et al. *Recommendations of the Physics Objects and Analysis Harmonisation Study Groups 2014*. Tech. rep. Geneva: CERN, May 2014. URL: https://cds.cern.ch/record/1700874 (cit. on p. 116).

[222] ATLAS Collaboration. *Performance of missing transverse momentum reconstruction with the ATLAS detector using proton–proton collisions at $\sqrt{s} = 13$ TeV*. In: *Eur. Phys. J. C* 78 (2018), p. 903. DOI: 10.1140/epjc/s10052-018-6288-9. arXiv: 1802.08168 [hep-ex] (cit. on p. 117).

[223] 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 (cit. on p. 117).

[224] N. Besson and M. Boonekamp. *Determination of the Absolute Lepton Scale Using Z Boson Decays: Application to the Measurement of MW*. Tech. rep. ATL-PHYS-PUB-2006-007. ATL-COM-PHYS-2005-072. Geneva: CERN, Nov. 2005. URL: http://cds.cern.ch/record/910107 (cit. on pp. 122, 138).

[225] ATLAS Collaboration. *Jet energy measurement and its systematic uncertainty in proton–proton collisions at* $\sqrt{s} = 7$ *TeV with the ATLAS detector*. In: *Eur. Phys. J. C* 75 (2015), p. 17. DOI: 10.1140/epjc/s10052-014-3190-y. arXiv: 1406.0076 [hep-ex] (cit. on pp. 122, 139).

[226] ATLAS Collaboration. *Measurement of the top quark mass in the* $t\bar{t} \to$ *lepton+jets and* $t\bar{t} \to$ *dilepton channels using* $\sqrt{s} = 7$ *TeV ATLAS data*. In: *Eur. Phys. J. C* 75 (2015), p. 330. DOI: 10.1140/epjc/s10052-015-3544-0. arXiv: 1503.05427 [hep-ex] (cit. on pp. 122, 153).

[227] ATLAS Collaboration. *Measurement of the top quark mass in the* $t\bar{t} \to$ *lepton+jets channel from* $\sqrt{s} = 8$ *TeV ATLAS data and combination with previous results*. In: *Eur. Phys. J. C* 79 (2019), p. 290. DOI: 10.1140/epjc/s10052-019-6757-9. arXiv: 1810.01772 [hep-ex] (cit. on pp. 122, 126, 128, 153).

[228] CMS Collaboration. *Jet energy scale and resolution in the CMS experiment in pp collisions at* 8 *TeV*. In: *JINST* 12 (2017), P02014. DOI: 10.1088/1748-0221/12/02/P02014. arXiv: 1607.03663 [hep-ex] (cit. on p. 123).

[229] Particle Data Group. *Top quark review*. 2020. URL: https://pdg.lbl.gov/2020/reviews/rpp2020-rev-top-quark.pdf (cit. on pp. 123, 124).

[230] Physics Modelling Group. *Top processes Recommendation*. URL: https://twiki.cern.ch/twiki/bin/viewauth/AtlasProtected/PmgTopProcesses (cit. on pp. 123, 127, 141, 150).

[231] ATLAS-CMS recommended predictions for top-quark-pair cross sections using the Top++v2.0 program (M. Czakon, A. Mitov, 2013). URL: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/TtbarNNLO (cit. on p. 123).

[232] ATLAS Collaboration. *Standard Model Summary Plots June 2021*. Tech. rep. Geneva: CERN, July 2021. URL: http://cds.cern.ch/record/2777014 (cit. on p. 124).

[233] CMS Collaboration. *Measurement of the top quark mass using proton–proton data at* $\sqrt{s} = 7$ *and* 8 *TeV*. In: *Phys. Rev. D* 93 (2016), p. 072004. DOI: 10.1103/PhysRevD.93.072004. arXiv: 1509.04044 [hep-ex] (cit. on pp. 126, 128).

[234] CMS Collaboration. *A profile likelihood approach to measure the top quark mass in the lepton+jets channel at* $\sqrt{s} = 13$ TeV. Tech. rep. Geneva: CERN, 2022. URL: https://cds.cern.ch/record/2806509 (cit. on p. 126).

[235] LHC Top Working Group. *History of LHCTopWG Top Mass Summary Plots*. URL: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/TopMassHistory (cit. on p. 126).

[236] ATLAS Collaboration. *Performance of electron and photon triggers in ATLAS during LHC Run 2*. In: *Eur. Phys. J. C* 80 (2020), p. 47. DOI: 10.1140/epjc/s10052-019-7500-2. arXiv: 1909.00761 [hep-ex] (cit. on pp. 127, 179).

[237] ATLAS Collaboration. *Performance of the ATLAS muon triggers in Run 2*. In: *JINST* 15 (2020), P09015. DOI: 10.1088/1748-0221/15/09/p09015. arXiv: 2004.13447 [hep-ex] (cit. on pp. 127, 179).

[238] ATLAS Collaboration. *Luminosity determination in pp collisions at* $\sqrt{s} = 8$ *TeV using the ATLAS detector at the LHC*. In: *Eur. Phys. J. C* 76 (2016), p. 653. DOI: 10.1140/epjc/s10052-016-4466-1. arXiv: 1608.03953 [hep-ex] (cit. on p. 127).

[239] ATLAS Production Group. *MC production campaigns*. URL: https://twiki.cern.ch/twiki/bin/view/AtlasProtected/AtlasProductionGroup (cit. on p. 127).

[240] ATLAS Production Group. *MC16 production campaign*. URL: https://twiki.cern.ch/twiki/bin/view/AtlasProtected/AtlasProductionGroupMC16 (cit. on p. 127).

[241] 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 (cit. on p. 127).

[242] ATLAS Collaboration. *Performance of the Fast ATLAS Tracking Simulation (FATRAS) and the ATLAS Fast Calorimeter Simulation (FastCaloSim) with single particles*. ATL-SOFT-PUB-2014-001. 2014. URL: https://cds.cern.ch/record/1669341 (cit. on p. 127).

[243] ATLAS Collaboration et al. *The simulation principle and performance of the ATLAS fast calorimeter simulation FastCaloSim*. Tech. rep. ATL-PHYS-PUB-2010-013. Geneva: CERN, Oct. 2010. URL: https://cds.cern.ch/record/1300517 (cit. on p. 127).

[244] K. Edmonds et al. *The Fast ATLAS Track Simulation (FATRAS)*. Tech. rep. ATL-SOFT-PUB-2008-001. ATL-COM-SOFT-2008-002. Geneva: CERN, Mar. 2008. URL: https://cds.cern.ch/record/1091969 (cit. on p. 127).

[245] Top working group. *Systematic affecting Top MC for* 13 TeV *analysis in release 21*. URL: https://twiki.cern.ch/twiki/bin/view/AtlasProtected/TopMCSystematicsR21 (cit. on pp. 127, 141).

[246] S. Frixione, G. Ridolfi, and P. Nason. *A positive-weight next-to-leading-order Monte Carlo for heavy flavour hadroproduction*. In: *JHEP* 09 (2007), p. 126. DOI: 10.1088/1126-6708/2007/09/126. arXiv: 0707.3088 [hep-ph] (cit. on pp. 127, 162).

[247] S. Alioli et al. *NLO single-top production matched with shower in POWHEG: s- and t-channel contributions*. In: *JHEP* 09 (2009), p. 111. DOI: 10.1088/1126-6708/2009/09/111. arXiv: 0907.4076 [hep-ph] (cit. on pp. 127, 162). Erratum: in: *JHEP* 02 (2010), p. 011. DOI: 10.1007/JHEP02(2010)011.

[248] E. Re. *Single-top Wt-channel production matched with parton showers using the POWHEG method*. In: *Eur. Phys. J. C* 71 (2011), p. 1547. DOI: 10.1140/epjc/s10052-011-1547-z. arXiv: 1009.2450 [hep-ph] (cit. on pp. 127, 162).

[249] S. Frixione et al. *Single-top hadroproduction in association with a W boson*. In: *JHEP* 07 (2008), p. 029. DOI: 10.1088/1126-6708/2008/07/029. arXiv: 0805.3067 [hep-ph] (cit. on pp. 127, 163).

[250] ATLAS Collaboration. *Studies on top-quark Monte Carlo modelling for Top2016*. ATL-PHYS-PUB-2016-020. 2016. URL: https://cds.cern.ch/record/2216168 (cit. on pp. 127, 141).

[251] *Script for generation of $t\bar{t}$ non-fully-hadronic sample DSID = 410470*. URL: https://gitlab.cern.ch/atlas-physics/pmg/infrastructure/mc15joboptions/blob/master/share/DSID410xxx/MC15.410470.PhPy8EG_A14_ttbar_hdamp258p75_nonallhad.py (cit. on p. 127).

[252] R. D. Ball et al. *Parton distributions for the LHC run II*. In: *JHEP* 04 (2015), p. 040. DOI: 10.1007/JHEP04(2015)040. arXiv: 1410.8849 [hep-ph] (cit. on pp. 127, 162, 163, 164).

[253] M. Czakon, P. Fiedler, and A. Mitov. *Total Top-Quark Pair-Production Cross Section at Hadron Colliders Through $O(\alpha_S^4)$*. In: *Phys. Rev. Lett.* 110 (2013), p. 252004. DOI: 10.1103/PhysRevLett.110.252004. arXiv: 1303.6254 [hep-ph] (cit. on p. 127).

[254] M. Aliev et al. *HATHOR – HAdronic Top and Heavy quarks crOss section calculatoR*. In: *Comput. Phys. Commun.* 182 (2011), pp. 1034–1046. DOI: 10.1016/j.cpc.2010.12.040. arXiv: 1007.1327 [hep-ph] (cit. on p. 127).

[255] P. Kant et al. *HATHOR for single top-quark production: Updated predictions and uncertainty estimates for single top-quark production in hadronic collisions*. In: *Computer Physics Communications* 191 (2015), pp. 74–89. ISSN: 0010-4655. DOI: https://doi.org/10.1016/j.cpc.2015.02.001. URL: https://www.sciencedirect.com/science/article/pii/S0010465515000454 (cit. on p. 127).

[256] 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 (cit. on pp. 128, 162).

[257] R. D. Ball et al. *Parton distributions with LHC data*. In: *Nucl. Phys. B* 867 (2013), p. 244. DOI: 10.1016/j.nuclphysb.2012.10.003. arXiv: 1207.1303 [hep-ph] (cit. on pp. 128, 162).

[258] H. Arnold, B. Moser, and T. A. Du Pree. *Calibration of the b-jet tagging efficiency using a tag-and-probe approach in single-lepton $t\bar{t}$ events at the ATLAS experiment*. Tech. rep. ATL-COM-PHYS-2019-116. Geneva: CERN, Feb. 2019. URL: https://cds.cern.ch/record/2659619 (cit. on p. 129).

[259] ATLAS Collaboration. *Selection of jets produced in* 13 *TeV proton–proton collisions with the ATLAS detector*. ATLAS-CONF-2015-029. 2015. URL: https://cds.cern.ch/record/2037702 (cit. on p. 130).

[260] H. C. Rytgaard. *Models for Robust Spline Smoothing – function space restrictions, L1 splines and generalizations*. PhD thesis. Aug. 2016. URL: http://curis.ku.dk/ws/files/164301283/thesis_helene_rytgaard.pdf (cit. on pp. 134, 309).

[261] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference and prediction*. 2nd ed. Springer, 2009. URL: http://www-stat.stanford.edu/~tibs/ElemStatLearn/ (cit. on pp. 134, 135, 136, 309).

[262] B. Grimstad et al. *SPLINTER: a library for multivariate function approximation with splines*. http://github.com/bgrimstad/splinter. 2015 (cit. on pp. 134, 309).

[263] F. James and M. Winkler. *Minuit 2 documentation*. URL: https://root.cern.ch/root/htmldoc/guides/minuit2/Minuit2.html (cit. on p. 136).

[264] *ROOT Minuit2 documentation*. URL: https://root.cern/doc/master/Minuit2Page.html (cit. on p. 136).

[265] Robert Reid. *Chi-squared distribution table with sigma values*. 2012. URL: http://www.reid.ai/2012/09/chi-squared-distribution-table-with.html (cit. on p. 138).

[266] Particle Data Group. *Statistics review*. 2020. URL: https://pdg.lbl.gov/2020/reviews/rpp2020-rev-statistics.pdf (cit. on p. 138).

[267] ATLAS Collaboration. *Measurements of b-jet tagging efficiency with the ATLAS detector using $t\bar{t}$ events at $\sqrt{s} = 13$ TeV*. In: *JHEP* 08 (2018), p. 089. DOI: 10.1007/JHEP08(2018)089. arXiv: 1805.01845 [hep-ex] (cit. on p. 140).

[268] I. Luise. *Observation of the Higgs boson coupling to b-quarks with the ATLAS detector*. 2019. URL: https://cds.cern.ch/record/2698480 (cit. on pp. 140, 186).

[269] ATLAS Collaboration. *ATLAS Pythia 8 tunes to 7 TeV data*. ATL-PHYS-PUB-2014-021. 2014. URL: https://cds.cern.ch/record/1966419 (cit. on pp. 141, 162, 163).

[270] LHAPDF Documentation. URL: https://lhapdf.hepforge.org/ (cit. on p. 141).

[271] J. Butterworth et al. *PDF4LHC recommendations for LHC Run II*. In: *J. Phys. G* 43 (2016), p. 023001. DOI: 10.1088/0954-3899/43/2/023001. arXiv: 1510.03865 [hep-ph] (cit. on pp. 141, 162, 163).

[272] ATLAS Collaboration. *Measurement of the $t\bar{t}$ production cross-section in the lepton+jets channel at $\sqrt{s} = 13$ TeV with the ATLAS experiment*. In: *Phys. Lett. B* 810 (2020), p. 135797. DOI: 10.1016/j.physletb.2020.135797. arXiv: 2006.13076 [hep-ex] (cit. on p. 142).

[273] F. E. James. *Statistical Methods in Experimental Physics; 2nd ed.* Singapore: World Scientific, 2006. URL: https://cds.cern.ch/record/1019859 (cit. on p. 144).

[274] S. Glashow, D. V. Nanopoulos, and A. Yildiz. *Associated production of Higgs bosons and Z particles*. In: *Phys. Rev. D* 18 (1978), pp. 1724–1727. DOI: 10.1103/PhysRevD.18.1724 (cit. on p. 157).

[275] ATLAS Collaboration. *Measurement of the associated production of a Higgs boson decaying into b-quarks with a vector boson at high transverse momentum in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*. In: *Phys. Lett. B* 816 (2021), p. 136204. DOI: 10.1016/j.physletb.2021.136204. arXiv: 2008.02508 [hep-ex] (cit. on pp. 158, 160, 175, 177, 180, 203, 209, 212, 214, 219, 220, 221, 230, 233, 234, 266).

[276] ATLAS Collaboration. *Measurement of the fiducial and differential cross-section of WH/ZH production for the b-jets + E_T^{miss} final state in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*. Tech. rep. Geneva: CERN, Apr. 2022. URL: https://cds.cern.ch/record/2805712 (cit. on pp. 158, 209).

[277] ATLAS Collaboration. *ATLAS Event Displays: Higgs boson decaying to two b-quarks*. Aug. 2018. URL: https://cds.cern.ch/record/2636049 (cit. on p. 160).

[278] 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: https://cds.cern.ch/record/2677054 (cit. on pp. 161, 208).

[279] G. Avoni et al. *The new LUCID-2 detector for luminosity measurement and monitoring in ATLAS*. In: *JINST* 13.07 (2018), P07017. DOI: 10.1088/1748-0221/13/07/P07017 (cit. on p. 161).

[280] ATLAS Collaboration. *Public ATLAS Luminosity Results for Run-2 of the LHC*. URL: https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2 (cit. on p. 162).

[281] ATLAS Collaboration. *Good Run Lists For Analysis Run 2*. URL: https://twiki.cern.ch/twiki/bin/viewauth/AtlasProtected/GoodRunListsForAnalysisRun2 (cit. on p. 162).

[282] ATLAS Collaboration. *Measurement of the Z/γ^* boson transverse momentum distribution in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector*. In: *JHEP* 09 (2014), p. 145. DOI: 10.1007/JHEP09(2014)145. arXiv: 1406.3660 [hep-ex] (cit. on pp. 162, 163).

[283] G. Cullen et al. *Automated one-loop calculations with GoSam*. In: *Eur. Phys. J. C* 72 (2012), p. 1889. DOI: 10.1140/epjc/s10052-012-1889-1. arXiv: 1111.2034 [hep-ph] (cit. on pp. 162, 163).

[284] K. Hamilton, P. Nason, and G. Zanderighi. *MINLO: multi-scale improved NLO*. In: *JHEP* 10 (2012), p. 155. DOI: 10.1007/JHEP10(2012)155. arXiv: 1206.3572 [hep-ph] (cit. on pp. 162, 163).

[285] G. Luisoni et al. *$HW^{\pm}$/HZ + 0 and 1 jet at NLO with the POWHEG BOX interfaced to GoSam and their merging within MiNLO*. In: *JHEP* 10 (2013), p. 083. DOI: 10.1007/JHEP10(2013)083. arXiv: 1306.2542 [hep-ph] (cit. on pp. 162, 163).

[286] M. L. Ciccolini, S. Dittmaier, and M. Krämer. *Electroweak radiative corrections to associated WH and ZH production at hadron colliders*. In: *Phys. Rev. D* 68 (2003), p. 073003. DOI: 10.1103/PhysRevD.68.073003. arXiv: hep-ph/0306234 [hep-ph] (cit. on pp. 162, 163).

[287] O. Brein, A. Djouadi, and R. Harlander. *NNLO QCD corrections to the Higgs-strahlung processes at hadron colliders*. In: *Phys. Lett. B* 579 (2004), pp. 149–156. DOI: 10.1016/j.physletb.2003.10.112. arXiv: hep-ph/0307206 (cit. on pp. 162, 163).

[288] G. Ferrera, M. Grazzini, and F. Tramontano. *Associated Higgs-W-Boson Production at Hadron Colliders: a Fully Exclusive QCD Calculation at NNLO*. In: *Phys. Rev. Lett.* 107 (2011), p. 152003. DOI: 10.1103/PhysRevLett.107.152003. arXiv: 1107.1164 [hep-ph] (cit. on pp. 162, 163).

[289] O. Brein et al. *Top-quark mediated effects in hadronic Higgs-Strahlung*. In: *Eur. Phys. J. C* 72 (2012), p. 1868. DOI: 10.1140/epjc/s10052-012-1868-6. arXiv: 1111.0761 [hep-ph] (cit. on pp. 162, 163).

[290] G. Ferrera, M. Grazzini, and F. Tramontano. *Higher-order QCD effects for associated WH production and decay at the LHC*. In: *JHEP* 04 (2014), p. 039. DOI: 10.1007/JHEP04(2014)039. arXiv: 1312.1669 [hep-ph] (cit. on pp. 162, 163).

[291] G. Ferrera, M. Grazzini, and F. Tramontano. *Associated ZH production at hadron colliders: The fully differential NNLO QCD calculation*. In: *Phys. Lett. B* 740 (2015), pp. 51–55. DOI: 10.1016/j.physletb.2014.11.040. arXiv: 1407.4747 [hep-ph] (cit. on pp. 162, 163).

[292] J. M. Campbell, R. K. Ellis, and C. Williams. *Associated production of a Higgs boson at NNLO*. In: *JHEP* 06 (2016), p. 179. DOI: 10.1007/JHEP06(2016)179. arXiv: 1601.00658 [hep-ph] (cit. on pp. 162, 163).

[293] L. Altenkamp et al. *Gluon-induced Higgs-strahlung at next-to-leading order QCD*. In: *JHEP* 02 (2013), p. 078. DOI: 10.1007/JHEP02(2013)078. arXiv: 1211.5015 [hep-ph] (cit. on pp. 162, 163).

[294] B. Hespel, F. Maltoni, and E. Vryonidou. *Higgs and Z boson associated production via gluon fusion in the SM and the 2HDM*. In: *JHEP* 06 (2015), p. 065. DOI: 10.1007/JHEP06(2015)065. arXiv: 1503.01656 [hep-ph] (cit. on pp. 162, 163).

[295] R. V. Harlander et al. *Soft gluon resummation for gluon-induced Higgs Strahlung*. In: *JHEP* 11 (2014), p. 082. DOI: 10.1007/JHEP11(2014)082. arXiv: 1410.0217 [hep-ph] (cit. on pp. 162, 163).

[296] R. V. Harlander, S. Liebler, and T. Zirke. *Higgs Strahlung at the Large Hadron Collider in the 2-Higgs-doublet model*. In: *JHEP* 02 (2014), p. 023. DOI: 10.1007/JHEP02(2014)023. arXiv: 1307.8122 [hep-ph] (cit. on pp. 162, 163).

[297] O. Brein, R. V. Harlander, and T. J. E. Zirke. *vh@nnlo – Higgs Strahlung at hadron colliders*. In: *Comput. Phys. Commun.* 184 (2013), pp. 998–1003. DOI: 10.1016/j.cpc.2012.11.002. arXiv: 1210.5347 [hep-ph] (cit. on pp. 162, 163).

[298] M. Czakon and A. Mitov. *Top++: A program for the calculation of the top-pair cross-section at hadron colliders*. In: *Comput. Phys. Commun.* 185 (2014), p. 2930. DOI: 10.1016/j.cpc.2014.06.021. arXiv: 1112.5675 [hep-ph] (cit. on pp. 162, 163).

[299] N. Kidonakis. *Next-to-next-to-leading logarithm resummation for s-channel single top quark production*. In: *Phys. Rev. D* 81 (2010), p. 054028. DOI: 10.1103/PhysRevD.81.054028. arXiv: 1001.5034 [hep-ph] (cit. on pp. 162, 163).

[300] N. Kidonakis. *Next-to-next-to-leading-order collinear and soft gluon corrections for t-channel single top quark production*. In: *Phys. Rev. D* 83 (2011), p. 091503. DOI: 10.1103/PhysRevD.83.091503. arXiv: 1103.2792 [hep-ph] (cit. on pp. 162, 163).

[301] N. Kidonakis. *Two-loop soft anomalous dimensions for single top quark associated production with a W^- or H^-*. In: *Phys. Rev. D* 82 (2010), p. 054018. DOI: 10.1103/PhysRevD.82.054018. arXiv: 1005.4451 [hep-ph] (cit. on pp. 162, 163).

[302] F. Cascioli, P. Maierhöfer, and S. Pozzorini. *Scattering Amplitudes with Open Loops*. In: *Phys. Rev. Lett.* 108 (2012), p. 111601. DOI: 10.1103/PhysRevLett.108.111601. arXiv: 1111.5206 [hep-ph] (cit. on pp. 162, 163, 164).

[303] T. Gleisberg and S. Höche. *Comix, a new matrix element generator*. In: *JHEP* 12 (2008), p. 039. DOI: 10.1088/1126-6708/2008/12/039. arXiv: 0808.3674 [hep-ph] (cit. on pp. 162, 163, 164).

[304] S. Schumann and F. Krauss. *A parton shower algorithm based on Catani–Seymour dipole factorisation*. In: *JHEP* 03 (2008), p. 038. DOI: 10.1088/1126-6708/2008/03/038. arXiv: 0709.1027 [hep-ph] (cit. on p. 162).

[305] S. Höche et al. *QCD matrix elements + parton showers. The NLO case*. In: *JHEP* 04 (2013), p. 027. DOI: 10.1007/JHEP04(2013)027. arXiv: 1207.5030 [hep-ph] (cit. on p. 162).

[306] S. Catani et al. *Vector Boson Production at Hadron Colliders: a Fully Exclusive QCD Calculation at NNLO*. In: *Phys. Rev. Lett.* 103 (2009), p. 082001. DOI: 10.1103/PhysRevLett.103.082001. arXiv: 0903.2120 [hep-ph] (cit. on pp. 162, 164).

[307] A. Denner et al. *Electroweak corrections to Higgs-strahlung off W/Z bosons at the Tevatron and the LHC with Hawk*. In: *JHEP* 03 (2012), p. 075. DOI: 10.1007/JHEP03(2012)075. arXiv: 1112.5142 [hep-ph] (cit. on pp. 162, 163).

[308] A. Denner et al. *HAWK 2.0: A Monte Carlo program for Higgs production in vector-boson fusion and Higgs strahlung at hadron colliders*. In: *Comput. Phys. Commun.* 195 (2015), pp. 161–171. DOI: 10.1016/j.cpc.2015.04.021. arXiv: 1412.5390 [hep-ph] (cit. on pp. 162, 163).

[309] D. J. Lange. *The EvtGen particle decay simulation package*. In: *Nucl. Instrum. Meth. A* 462 (2001), p. 152. DOI: 10.1016/S0168-9002(01)00089-4 (cit. on p. 162).

[310] ATLAS Collaboration. *Evidence for the $H \to b\bar{b}$ decay with the ATLAS detector*. In: *JHEP* 12 (2017), p. 024. DOI: 10.1007/JHEP12(2017)024. arXiv: 1708.03299 [hep-ex] (cit. on pp. 165, 175, 179).

[311] ATLAS Collaboration. *Improving jet substructure performance in ATLAS using Track-CaloClusters*. ATL-PHYS-PUB-2017-015. 2017. URL: https://cds.cern.ch/record/2275636 (cit. on p. 166).

[312] ATLAS Collaboration. *Optimisation of large-radius jet reconstruction for the ATLAS detector in 13 TeV proton–proton collisions*. ATLAS-CONF-2020-021. 2020. URL: https://cds.cern.ch/record/2723736 (cit. on p. 166).

[313] ATLAS Collaboration. *Recommended isolation working points (Rel. 21)*. URL: https://twiki.cern.ch/twiki/bin/view/AtlasProtected/RecommendedIsolationWPs?rev=29 (cit. on p. 173).

[314] ATLAS Collaboration. *Performance of the missing transverse momentum triggers for the ATLAS detector during Run-2 data taking*. In: *JHEP* 08 (2020), p. 080. DOI: 10.1007/JHEP08(2020)080. arXiv: 2005.09554 [hep-ex] (cit. on pp. 178, 179).

[315] M. Proissl. *Dijet Invariant Mass Studies in the Higgs boson $H \to b\bar{b}$ resonance search in association with a W/Z boson using the ATLAS detector*. 2014. URL: https://cds.cern.ch/record/1997808 (cit. on p. 180).

[316] C. Kato. *Evidence for the Higgs boson decaying to a pair of b quarks*. 2017. URL: https://cds.cern.ch/record/2644553 (cit. on p. 180).

[317] A. D. Bukin. *Fitting function for asymmetric peaks*. 2007. DOI: 10.48550/ARXIV.0711.4449. URL: https://arxiv.org/abs/0711.4449 (cit. on p. 180).

[318] *Flavor Tagging Efficiency Parametrisations with Graph Neural Networks*. Tech. rep. Geneva: CERN, 2022. URL: https://cds.cern.ch/record/2825433 (cit. on pp. 186, 187, 188).

[319] D. Gonçalves and J. Nakamura. *Role of the Z polarization in the $H \to b\bar{b}$ measurement*. In: *Phys. Rev. D* 98.9 (2018), p. 093005. DOI: 10.1103/PhysRevD.98.093005. arXiv: 1805.06385 [hep-ph] (cit. on p. 193).

[320] G. Cowan et al. *Asymptotic formulae for likelihood-based tests of new physics*. In: *Eur. Phys. J. C* 71 (2011), p. 1554. DOI: 10.1140/epjc/s10052-011-1554-0. arXiv: 1007.1727 [physics.data-an] (cit. on pp. 195, 198, 199, 201). Erratum: in: *Eur. Phys. J. C* 73 (2013), p. 2501. DOI: 10.1140/epjc/s10052-013-2501-z.

[321] R. Barlow and C. Beeston. *Fitting using finite Monte Carlo samples*. In: *Computer Physics Communications* 77.2 (1993), pp. 219–228. ISSN: 0010-4655. DOI: https://doi.org/10.1016/0010-4655(93)90005-W. URL: https://www.sciencedirect.com/science/article/pii/001046559390005W (cit. on p. 197).

[322] N. Metropolis and S. Ulam. *The Monte Carlo Method*. In: *Journal of the American Statistical Association* 44.247 (1949), pp. 335–341. DOI: 10.1080/01621459.1949.10483310. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1949.10483310. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1949.10483310 (cit. on p. 198).

[323] S. S. Wilks. *The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses*. In: *The Annals of Mathematical Statistics* 9.1 (1938), pp. 60–62. DOI: 10.1214/aoms/1177732360. URL: https://doi.org/10.1214/aoms/1177732360 (cit. on p. 198).

[324] D. Pati. *Likelihood Ratio tests*. 2022. URL: https://web.stat.tamu.edu/~debdeep/LRTests.pdf (cit. on p. 198).

[325] A. Wald. *Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large*. In: *Transactions of the American Mathematical Society* 54.3 (1943), pp. 426–482. ISSN: 00029947. URL: http://www.jstor.org/stable/1990256 (visited on 10/01/2022) (cit. on p. 199).

[326] E. Gross. *Practical Statistics for High Energy Physics*. In: *CERN Yellow Rep. School Proc.* 3 (2018). Ed. by M. Mulders and G. Zanderighi, pp. 199–221. DOI: 10.23730/CYRSP-2018-003.199 (cit. on pp. 200, 201).

[327] K. Cranmer, J. Pavez, and G. Louppe. *Approximating Likelihood Ratios with Calibrated Discriminative Classifiers*. 2015. DOI: 10.48550/ARXIV.1506.02169. URL: https://arxiv.org/abs/1506.02169 (cit. on p. 205).

[328] ATLAS Collaboration. *Measurement of VH, $H \to b\bar{b}$ production as a function of the vector-boson transverse momentum in 13 TeV pp collisions with the ATLAS detector*. In: *JHEP* 05 (2019), p. 141. DOI: 10.1007/JHEP05(2019)141. arXiv: 1903.04618 [hep-ex] (cit. on p. 209).

[329] ATLAS Collaboration. *Simplified template cross section measurements of Higgs boson produced in association with vector bosons in the $H \to b\bar{b}$ decay channel in proton-proton collisions at $\sqrt{s} = 13$ TeV*. Tech. rep. Geneva: CERN, 2022. URL: https://cds.cern.ch/record/2827421 (cit. on pp. 222, 223).

[330] A. L. Read. *Presentation of search results: the CL_s technique*. In: *Journal of Physics G: Nuclear and Particle Physics* 28.10 (Sept. 2002), pp. 2693–2704. DOI: 10.1088/0954-3899/28/10/313. URL: https://doi.org/10.1088/0954-3899/28/10/313 (cit. on p. 224).

[331] C. Delaunay et al. *Enhanced Higgs boson coupling to charm pairs*. In: *Phys. Rev. D* 89 (2014), p. 033014. DOI: 10.1103/PhysRevD.89.033014 (cit. on p. 227).

[332] G. Perez et al. *Constraining the charm Yukawa and Higgs-quark coupling universality*. In: *Phys. Rev. D* 92 (2015), p. 033016. DOI: 10.1103/PhysRevD.92.033016 (cit. on p. 227).

[333] A. Bazavov et al. *Up-, down-, strange-, charm-, and bottom-quark masses from four-flavor lattice QCD*. In: *Phys. Rev. D* 98 (Sept. 2018), p. 054517. DOI: 10.1103/PhysRevD.98.054517. URL: https://link.aps.org/doi/10.1103/PhysRevD.98.054517 (cit. on p. 227).

[334] CMS Collaboration. *Search for Higgs boson decay to a charm quark-antiquark pair in proton-proton collisions at $\sqrt{s} = 13$ TeV*. Tech. rep. Geneva: CERN, 2022. arXiv: 2205.05550. URL: https://cds.cern.ch/record/2809290 (cit. on pp. 228, 229, 242, 246).

[335] H. Qu and L. Gouskos. *Jet tagging via particle clouds*. In: *Phys. Rev. D* 101 (5 Mar. 2020), p. 056019. DOI: 10.1103/PhysRevD.101.056019. URL: https://link.aps.org/doi/10.1103/PhysRevD.101.056019 (cit. on p. 228).

[336] CMS Collaboration. *A search for the standard model Higgs boson decaying to charm quarks*. In: *JHEP* 03 (2020), p. 131. DOI: 10.1007/JHEP03(2020)131. arXiv: 1912.01662 [hep-ex] (cit. on pp. 228, 229).

[337] CMS Collaboration. *Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques*. In: *JINST* 15 (2020), P06005. DOI: 10.1088/1748-0221/15/06/P06005. arXiv: 2004.08262 [hep-ex] (cit. on p. 228).

[338] ATLAS Collaboration. *Evaluation of theoretical uncertainties for simplified template cross section measurements of V-associated production of the Higgs boson*. ATL-PHYS-PUB-2018-035. 2018. URL: https://cds.cern.ch/record/2649241 (cit. on p. 235).

[339] ATLAS Collaboration. *Extrapolation of ATLAS sensitivity to $H \to b\bar{b}$ and $H \to c\bar{c}$ decays in VH production at the HL-LHC*. ATL-PHYS-PUB-2021-039. 2021. URL: https://cds.cern.ch/record/2788490 (cit. on pp. 246, 249, 250, 251).

[340] ATLAS Collaboration. *Graph Neural Network Jet Flavour Tagging with the ATLAS Detector*. Tech. rep. Geneva: CERN, June 2022. URL: https://cds.cern.ch/record/2811135 (cit. on pp. 246, 247).

[341] ATLAS Collaboration. *Dependence of the Jet Energy Scale on the Particle Content of Hadronic Jets in the ATLAS Detector Simulation*. Tech. rep. Geneva: CERN, 2022. URL: https://cds.cern.ch/record/2808016 (cit. on p. 248).

[342] P. Azzurri et al. *A special Higgs challenge: Measuring the mass and production cross section with ultimate precision at FCC-ee*. In: *Eur. Phys. J. Plus* 137.1 (2021), p. 23. DOI: 10.1140/epjp/s13360-021-02202-4. arXiv: 2106.15438. URL: https://cds.cern.ch/record/2775926 (cit. on pp. 251, 252).

[343] ATLAS Collaboration. *Search for resonances in diphoton events at $\sqrt{s} = 13$ TeV with the ATLAS detector*. In: *JHEP* 09 (2016), p. 001. DOI: 10.1007/JHEP09(2016)001. arXiv: 1606.03833 [hep-ex] (cit. on p. 252).

[344] G. Frattari. *Investigating the nature of dark matter and of the Higgs boson with jets and missing transverse momentum at the LHC*. Oct. 2021. URL: https://cds.cern.ch/record/2802389 (cit. on p. 287).

[345] *Python Scipy Interpolate module documentation*. URL: https://docs.scipy.org/doc/scipy/reference/interpolate.html (cit. on p. 309).

[346] *Python Statmodels NonParametric module*. URL: https://www.statsmodels.org/dev/nonparametric.html (cit. on p. 309).

Appendices

A Penalized spline: some technical details 309
A.1 Cross-validated residual sum of squares tests 309

A Penalized spline: some technical details

As mentionned already in Section 4.4, the open-source C++ Splinter (SPLine INTERpolation) library [262] has been used in order to compute penalized spline curves.

Other C++ libraries exist such as ALGLIB, but Splinter was chosen for its simplicity to use and to integrate in a C++ code. An alternative implementation of smoothing splines is available in Python in the Scipy module [345] with the functions `scipy.interpolate.UnivariateSpline` and `scipy.interpolate.BivariateSpline` functions.

A.1 Cross-validated residual sum of squares tests

The minimization of cross-validated residual sum of squares (CVRSS), also called leave-one out cross-validation, is a method to find the optimal smoothing parameter $\lambda_{\text{opt.}}$ for a penalized spline as explained in appendix B.3 (titled "Choosing the Smoothing Parameter λ") of Ref. [260] or in section 7.10 ("Cross validation") of Ref. [261]. The CVRSS is computed as follows:

$$\text{CVRSS}(\lambda) = \sum_{i=1}^{n} (y_i - f_\lambda^{(-i)}(x_i))^2 \tag{A.1}$$

where $f_\lambda^{(-i)}$ is the penalized spline obtained when removing the i^{th} point (x_i, y_i) of the dataset. Thus, $f_\lambda^{(-i)}(x_i)$ is the value of $f_\lambda^{(-i)}$ precisely at the i^{th} point. Basically the function CVRSS(λ) assesses how stable is the penalized spline when removing points. The CVRSS is called a loss function, *i.e.* it is the function to minimize in order to find the optimal parameter $\lambda_{\text{opt.}}$.

This method was tested unsuccessfully in the case of the χ^2 curves: the penalized curves obtained are overfitting the original χ^2 curves and are sensitive to fluctuations because $\lambda_{\text{opt.}}$ found is too small. This can be explained by the fact that for b-JES scans, a χ^2 curve contains 200 points and the step between dataset points 0.1% is rather small so when removing one point the value $f_\lambda^{(-i)}(x_i)$ is still very close to y_i for small values of λ.

Similarly, the same technique was used removing several points instead of one ("leave k-out cross validation" or "k-fold cross-validation" method) as explained in Ref. [261], but similar conclusions were obtained.

To test that obtaining a too small $\lambda_{\text{opt.}}$ was not coming from particularities of the Splinter library, another smoothing technique called LOWESS (for Locally Weighted Scatterplot Smoothing) was tested using the function `statsmodels.nonparametric.smoothers_lowess.lowess` of the Statmodels Python module [346]. That function uses a smoothing parameter $s \in [0, 1]$ and the smoothing is performed using neighbouring points whose weight decreases with their distance to that position. s determines the percentage of points to use in the dataset to realize the smoothing. The same conclusions were obtained: the optimal parameter $s_{\text{opt.}}$ was too small and the smoothed curve was overfitting the original χ^2 curve and its fluctuations.

As a conclusion, the CVRSS loss function is not relevant for the χ^2 curve smoothing so one has to find a new loss function denoted CV(λ) that would be more adequate.

No solutions to this problem were found both in literature and on StackExchange even after asking questions on different forums and on the Splinter library repository:

➢ StackExchange Statistics section:
https://stats.stackexchange.com/questions/485832/hyper-parameter-optimization-for-regression-to-avoid-overfiting-underfiting

➢ StackExchange Mathematic section:
https://math.stackexchange.com/questions/3813122/hyper-parameter-optimization-for-regression-to-avoid-overfiting-underfiting

➢ Splinter library: https://github.com/bgrimstad/splinter/issues/125

Therefore an empirical search of a more general loss function has been performed: see Equation (4.4.4).