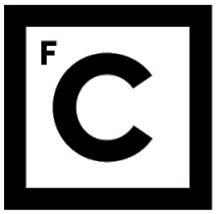
UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS



Ciências
ULisboa

Search for the Higgs boson at ATLAS/LHC in WH associated production and decay to b-quark pairs

Doutoramento em Física

Rute Costa Batalha Pedro

Tese orientada por:
Prof. Doutor José Carvalho Maneira
Prof. Doutora Patricia Conde Muíño

Documento especialmente elaborado para a obtenção do grau de doutor

2017

UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

Search for the Higgs boson at ATLAS/LHC in WH associated production and decay to b-quark pairs

Doutoramento em Física

Rute Costa Batalha Pedro

Tese orientada por:
Prof. Doutor José Carvalho Maneira
Prof. Doutora Patricia Conde Muíño

Júri:
Presidente:

- Doutora Margarida Maria Telo da Gama, Professora Catedrática
  Faculdade de Ciências da Universidade de Lisboa

Vogais:

- Doutor Andrew Mehta, Reader
  Departamento de Física da Universidade de Liverpool, Reino Unido;
- Doutor João Carlos Lopes de Carvalho, Professor Associado com Agregação
  Faculdade de Ciências e Tecnologia da Universidade de Coimbra;
- Doutor João Manuel Coelho dos Santos Varela, Professor Associado
  Instituto Superior Técnico da Universidade de Lisboa;
- Doutora Amélia Arminda Teixeira Maio, Professora Associada Jubilada
  Faculdade de Ciências da Universidade de Lisboa;
- Doutor António Joaquim Rosa Amorim Barbosa, Professor Catedrático
  Faculdade de Ciências da Universidade de Lisboa;
- Doutor José Carvalho Maneira, Professor Auxiliar Convidado
  Faculdade de Ciências da Universidade de Lisboa (Orientador).

Documento especialmente elaborado para a obtenção do grau de doutor

Fundação para a Ciência e Tecnologia - PhD grant SFRH/BD/81204/2011

2017

Resumo

O mecanismo de Higgs foi introduzido no Modelo Padrão das partículas elementares e suas interacções na década de 1960, para resolver o conflito existente entre partículas massivas e leis de conservação da física de partículas. Uma das consequências deste mecanismo é a previsão de uma nova partícula fundamental, o bosão de Higgs, observado pela primeira vez em 2012 pelas experiências ATLAS e CMS do LHC/CERN.

Esta tese descreve a pesquisa pelo bosão de Higgs através do seu decaimento em pares de quarks b com o detector ATLAS, usando acontecimentos de colisões pp com uma energia de centro-de-massa de 8 TeV. Este modo de decaimento ainda não foi observado, embora a razão de bifurcação (BR) seja dominante relativamente aos processos alternativos: para um Higgs de massa $m_H = 125$ GeV, $BR(H \to b\bar{b})$ =57.7%. A sua procura é desafiante e difícil devido à quantidade de acontecimentos de fundo com jactos de partículas. Para reduzir esse fundo, escolhe-se a produção do bosão de Higgs associada a um bosão W/Z, uma vez que os leptões resultantes do decaimento do W/Z constituem uma forma efectiva de identificar o sinal.

Considera-se a produção associada a um W e seleccionam-se os acontecimentos de acordo com a topologia do sinal $WH \to \ell\nu b\bar{b}$: um electrão ou um muão, energia em falta associada ao neutrino e dois jactos resultantes da fragmentação dos quarks b. Acontecimentos que resultam em jactos e leptões carregados, como a produção de quarks top e de W+jactos, constituem os fundos principais da análise. Como a secção eficaz de produção destes processos é muito superior à do sinal, a proporção de acontecimentos de sinal (S) e fundo (B) $S/\sqrt{B}$ é de apenas 0.3.

A análise usa uma técnica Multivariacional, que explora correlações entre diferentes observáveis através do método Boosted Decision Tree (BDT), para aumentar a sensibilidade aos acontecimentos de sinal. Realizou-se um estudo que permitiu melhorar o desempenho da BDT até 12%.

São também analisadas amostras de simulação de sinal e fundos nas mesmas condições. Dada a pequena significância do sinal relativamente ao fundo, foi indispensável verificar que a simulação modela correctamente os fundos e avaliar as incertezas sistemáticas na sua previsão. Neste contexto, foi efectuado um estudo que levou à determinação das incertezas sistemáticas associadas à modelação da produção do quark top.

A razão entre a taxa de acontecimentos de sinal observada e a prevista pelo Modelo Padrão foi $1.65^{+0.58}_{-0.56}(\text{stat})^{+0.58}_{-0.48}(\text{syst}) = 1.65^{+0.82}_{-0.74}$, medida compatível com a previsão tendo em conta as incertezas obtidas. A significância do sinal medido, que representa a probabilidade dos dados observados serem compatíveis com a hipótese de ausência de sinal, corresponde à probabilidade gaussiana de observar um valor superior a 2.02 desvios padrão e é insuficiente para se declarar a observação do processo $WH \to \ell\nu b\bar{b}$.

Palvaras-Chave: Higgs Modelo Padrão, Produção associada, Decaimento $b\bar{b}$, ATLAS/LHC, Boosted Decision Tree

Sumário

O bosão de Higgs foi teorizado pelo mecanismo de quebra espontânea da simetria electrofraca proposto em 1964 por R. Brout e F. Englert, P. W. Higgs, e G. Guralnik, C. R. Hagen, and T. Kibble para resolver a problemática da massa das partículas fundamentais no Modelo Padrão (MP) e as divergências previstas na dispersão WW. Este mecanismo prevê todas as propriedades da partícula: carga eléctrica nula, spin 0 e paridade positiva, excepto a sua massa. Desde a previsão foram realizadas várias pesquisas para encontrar o bosão de Higgs em experiências envolvendo colisão de partículas a altas energias, como o LEP e o Tevatrão e mais recentemente o LHC. A 4 de Julho de 2012, as experiências ATLAS e CMS do LHC no CERN anunciaram a observação independente de um novo bosão, com massa aproximada de 125 GeV, compatível com o bosão de Higgs do MP.

O acelerador de hadrões instalado no complexo de aceleradores do CERN, LHC, foi construído para acelerar e colidir protões e núcleos de chumbo. Este acelerador foi projectado para acelerar protões até 7 TeV, o que resulta numa energia de centro-de-massa das colisões com limite superior de 14 TeV. Em 2011 e 2012, período de operações habitualmente designado por Run I, o LHC colidiu protões a 7 e 8 TeV dando origem a amostras de dados com um tamanho correspondente à luminosodade integrada de 5.6 fb^{-1}e 20.3 fb^{-1}, respectivamente. Os acontecimentos de colisão de partículas são detectados por quatro detectores instalados em torno dos quatro pontos nominais de colisão do LHC: ATLAS, CMS, ALICE e LHCb.

A experiência ATLAS dedica-se à investigação de um espectro largo de tópicos de Física de colisões a altas energias, que se estende desde a procura pelo bosão de Higgs às medidas de precisão do MP e ao teste de uma vasta quantidade de modelos para além do MP, cujo objectivo é responder às questões em aberto no ramo da Física de Partículas. O detector utiliza diferentes tecnologias para medir e distinguir a variedade de produtos das colisões do LHC e é composto por diferentes camadas funcionais: um detector de traços de partículas carregadas electricamente, um calorímetro electromagnético e hadrónico e um espectrómetro de muões. A experiência dispõe também de um sistema de selecção de acontecimentos em tempo real, desenhado para seleccionar os acontecimentos interessantes para investigação, e que mantém o fluxo de dados compatível com o sistema de aquisição e armazenamento.

Os pricipais mecanismos de produção do bosão de Higgs a partir de colisão de protões são fusão de gluões, fusão de bosões vectoriais, produção associada a um bosão W/Z e a um par de quarks top. Para uma dada energia de centro-de-massa das colisões de protões, a secção eficaz destes processos depende apenas da massa do bosão de Higgs, parâmetro livre da teoria. A intensidade do acoplamento entre o bosão de Higgs e as diferentes partículas fundamentais do MP é proporcional à massa das partículas no caso dos fermiões e ao quadrado da massa no caso dos bosões. Por este motivo, o bosão de Higgs tende a decair mais frequentemente em partículas massivas. No entanto, factores cinemáticos do decaimento fazem com que a razão de bifurcação, definida como a razão entre a largura parcial de um dado modo de decaimento

e a largura total da partícula, dependa também da massa do próprio bosão de Higgs. Para uma massa de 125 GeV, este decai maioritariamente em pares de quarks b (57.7%) e de bosões W (21%).

A descoberta do bosão compatível com o Higgs baseou-se essencialmente em modos de decaimento bosónicos: $H \to WW \to \ell\nu\ell\nu$, $H \to ZZ \to \ell\ell\ell\ell$ e $H \to \gamma\gamma$, com $\ell = e, \mu$. Estes processos possuem uma assinatura experimental limpa, e por isso constituíram as primeiras apostas das experiências ATLAS e CMS no âmbito da procura do bosão de Higgs no primeiro conjunto de dados. A descoberta consistiu na observação de um excesso de acontecimentos estatisticamente significante nos espectros de massa invariante ou massa transversa dos produtos finais da cadeia de decaimento, relativamente ao fundo total previsto pelo MP.

Os resultados da análise da amostra completa de dados da Run I do LHC constituem a esta data o estado-da-arte no que respeita ao campo de investigação experimental sobre o modelo de Higgs. Incluem a medida mais precisa da massa, $m_H = 125.09 \pm 0.24$ GeV, resultante da combinação das medidas individuais de ATLAS e CMS. Da combinação obteve-se também a medida da secção eficaz dos diversos modos de produção do bosão de Higgs e das razões de bifurcação dos modos de decaimento mais significativos. A intensidade dos acoplamentos entre o Higgs e diferentes partículas elementares foi obtida de forma semelhante. ATLAS e CMS conduziram de forma individual testes ao spin e à paridade. Todos os resultados sugerem que a partícula observada é compatível com a previsão do MP.

No entanto, o modo de decaimento principal do bosão de Higgs, $H \to b\bar{b}$, não foi ainda observado. Este canal tem dois jactos como assinatura experimental e é por isso particularmente difícil de detectar devido à abundante produção de jactos de fundo em colisões de hadrões, com uma secção eficaz $\sim 10^{17}$ vezes maior do que a secção eficaz prevista para a produção do Higgs. A procura pelo sinal $H \to b\bar{b}$ não beneficia portanto de uma busca independente do mecanismo de produção da partícula. Por sua vez, a produção associada a um bosão W/Z, ou acompanhada de um par de quarks top é relevante para a identificação do sinal. Estes modos de produção têm no entanto secções eficazes reduzidas e não foram observados até ao momento, ao contrário dos restantes mecanismos de produção principais.

Dada a grandeza da razão de bifurcação de $H \to b\bar{b}$, os resultados da sua procura têm um papel fundamental na medida das propriedades do bosão de Higgs, desde a largura total, à identificação da sua natureza.

O trabalho descrito nesta tese corresponde à procura pelo bosão de Higgs com decaimento em dois quarks b e com produção associada a um bosão W. Considera-se o decaimento leptónico do W originando um electrão ou um muão porque estes leptões têm uma assinatura experimental limpa e permitem que o acontecimento de sinal seja seleccionado em tempo real. São analisados os dados de colisões de protões do LHC a uma energia de centro-de-massa de $\sqrt{s}$ =8 TeV correspondendo a uma luminosidade integrada de 20.3 fb^{-1}adquiridos pelo detector ATLAS em 2012.

Os acontecimentos são seleccionados de acordo com a topologia do sinal $WH \to \ell\nu b\bar{b}$: um electrão ou um muão isolados e de alto momento transverso p_{T}, elevada energia transversa em

falta associada ao neutrino que atravessa o detector sem depositar energia, e dois jactos isolados de alto p_{T} resultantes da fragmentação de quarks b. Estes objectos são reconstruídos a partir das medidas das diferentes sub-camadas do detector. Os electrões são identificados por um aglomerado de energia depositada no calorímetro electromagnético espacialmente combinado a um traço no detector de traços. Os muões são identificados no espectrómetro de muões e associados a um traço no detector de traços. Os jactos resultam da combinação de aglomerados de energia nos calorímetros electromagnético e hadrónico com o algoritmo anti-kt. Para a determinação do sabor do partão que originou o jacto utiliza-se essencialmente as medidas do detector de traços. O hadrão b resultante da fragmentação de um quark b desloca-se do ponto de colisão antes de decair. A medição desse deslocamento e a reconstrução do vértice do seu decaimento constituem a base da identificação de jactos b.

O desenho da selecção de acontecimentos foi estabelecido pelo grupo de análise de ATLAS e teve como objectivo a maximização da eficiência na selecção do processo de sinal e da rejeição dos fundo, tendo como base informação de simulação Monte-Carlo (MC). Os fundos principais da análise são pares de quarks top, $t\bar{t}$; W+jactos; quark top; dibosões (WW, WZ e ZZ) e multijactos. A previsão da composição da amostra de dados é retirada da simulação MC para os diferentes processos de fundo e sinal, excepto para o fundo de multijactos, que é derivada a partir de uma amostra de dados enriquecida neste tipo de acontecimentos. A razão sinal/fundo $S/\sqrt{B}$ obtida é de cerca de 0.3, e aqui reside a maior dificuldade da busca por $H \to b\bar{b}$.

A análise emprega uma abordagem Multivariacional para extrair do conjunto de dados seleccionados a máxima sensibilidade ao sinal. A técnica utilizada, Boosted Decision Tree (BDT), explora correlações entre diferentes observáveis para aceder a espaços de fase da amostra onde a proporção sinal/fundo é maior, compondo um classificador discriminante no final. A BDT permitiu alargar até 30% o ganho na sensibilidade ao sinal relativamente a técnicas tradicionais de selecção de acontecimentos. É gerada a partir de simulação e aplicada em dados reais, pelo que o seu desempenho depende de forma crítica de quão precisos são os modelos de simulação na previsão dos acontecimentos reais.

A procura pelo sinal $WH \to \ell\nu b\bar{b}$ culmina numa análise estatística em que os dados são comparados à previsão MC e é medida a compatibilidade entre os dois. Tecnicamente, maximiza-se uma função de verosimilhança que incorpora um termo de Poisson para descrever a probabilidade dos dados corresponderem à previsão MC e vários termos que modelam o efeito das incertezas sistemáticas, experimentais e teóricas, nessa previsão.

O trabalho descrito nesta tese foi realizado no contexto de um sub-grupo de trabalho da experiência ATLAS, HSG5, cujo objectivo é a procura pelo decaimento do bosão de Higgs num par de quarks bottom, produzido em associação a um bosão W ou Z. Foi feita uma análise completa do canal $WH \to \ell\nu b\bar{b}$ com um código de selecção de acontecimentos desenvolvido independentemente. A implementação do código permitiu inter-validar as ferramentas de análise utilizadas na colaboração HSG5 e estabeleceu o ponto de partida das contribuições pessoais para a análise oficial de ATLAS.

Uma das contribuições pessoais para a análise consistiu na determinação das incertezas

sistemáticas na previsão do fundo top, a que correspondem três canais de produção - s, t e Wt, a partir da comparação de diferentes modelos de simulação. Os resultados obtidos reflectem o impacto na análise das incertezas nos processos de hadronização, de radiação partónica e da composição dos protões, e traduzem ainda efeitos relacionados com a ordem na teoria de perturbações a que os acontecimentos são gerados. Obteve-se uma incerteza máxima na previsão do número de acontecimentos top de 30%, 52% e 15% para o canal s, t e Wt, respectivamente.

Adicionalmente, realizei um estudo de optimização do desempenho da BDT, através da inclusão de novos observáveis com potencial poder de discriminação de sinal. Foram testados cerca de 20 novos observáveis com natureza cinemática, angular e de forma do acontecimento. Dois deles permitiram aumentar o poder de separação de sinal e fundo da BDT em 12%. Para que pudessem ser utilizados, foi necessário verificar que a simulação dos diversos fundos descrevia correctamente os dois observáveis, através de comparações de dados e MC em regiões enriquecidas em cada um dos fundos principais. No caso do sinal, a validação foi realizada por via da comparação de diferentes modelos de simulação. As discrepâncias observadas estavam dentro da incerteza tida em conta na análise para a previsão do sinal. As novas variáveis foram adoptadas na análise oficial de ATLAS e foram utilizadas no primeiro conjunto de dados da Run II do LHC.

A análise estatística dos dados resultou na medida da taxa de acontecimentos de sinal relativa à prevista pelo MP de $1.65\,^{+0.58}_{-0.56}(\text{stat})^{+0.58}_{-0.48}(\text{syst}) = 1.65^{+0.82}_{-0.74}$. A medida é compatível com a previsão do MP tendo em conta as incertezas da medição, igualmente repartidas entre a natureza sistemática e estatística. A significância do sinal medido, que representa a probabilidade dos dados observados serem compatíveis com a hipótese de ausência de sinal, corresponde à probabilidade gaussiana de observar um valor superior a 2.02 desvios padrão e é insuficiente para se declarar a observação do processo $WH \to \ell\nu b\bar{b}$.

Abstract

The Higgs mechanism was incorporated in the Standard Model of elementary particles and interactions in the 1960's to solve the existent conflict between massive particles and conservation laws of particle physics. A consequence of this mechanism is the prediction of a new fundamental particle, the Higgs boson, observed for the first time in 2012 by the ATLAS and CMS experiments at the Large Hadron Collider.

This thesis describes the search for the Higgs decay into a pair of $b-$quarks with the ATLAS experiment, using pp collision events with an 8 TeV center-of-mass energy provided by the LHC in 2012. Although the branching fraction of the $H \to b\bar{b}$ decay is dominant ($BR(H \to b\bar{b})$ =57.7% for $m_H = 125\ GeV$), this decay mode was not yet observed. The search is particularly challenging given the huge amount of background events containing jets. To reduce this background, the Higgs production associated with a W/Z boson is usually explored, as the leptons resulting from the W/Z decay can effectively trigger the signal.

The W associated production with the W boson decaying leptonically is considered. The data analysis searches for events compatible with the $WH \to \ell\nu b\bar{b}$ signal topology: one electron or muon, missing transverse energy associated with the undetected neutrino and two jets resulting from $b-$ quark fragmentation. Events containing jets and charged leptons, as top-quark production and $W+$jets, are the main backgrounds of the analysis. Since their production cross-section is much larger than the signal cross-section, the resulting signal-to-background proportion, $S/\sqrt{B}$ is only 0.3.

The analysis comprehends a Multivariate technique, Boosted Decision Tree (BDT), to exploit correlations in the event observables aiming at increasing the sensitivity to the signal. A study that resulted in a 12% gain in the BDT performance was carried on.

Samples of signal and background simulated in the same conditions as data are also analysed. Given the small $S/\sqrt{B}$, it was indispensable to verify that the simulation models correctly the background processes, and to evaluate the systematical uncertainties associated with their prediction. In this context, a study to determine the systematic uncertainties of the single top background modelling was conducted.

The ratio between the observed signal event rate and the Standard Model prediction was $1.65^{+0.58}_{-0.56}(\text{stat})^{+0.58}_{-0.48}(\text{syst}) = 1.65^{+0.82}_{-0.74}$, and therefore the measurement is compatible with the SM prediction within uncertainties. The signal significance, representing the compatibility between the data observation and the background-only hypothesis, corresponds to the gaussian probability of observing a value larger than 2.02 standard deviations and is not sufficient to state the observation of the $WH \to \ell\nu b\bar{b}$ process.

Keywords: Standard Model Higgs, Associated Production, $b\bar{b}$ decay, ATLAS/LHC, Boosted Decision Tree

Acknowledgements

I would like to express my sincere gratitude to my PhD supervisors, José Maneira and Patricia Conde, for their thoughtful guidance and support, and stimulating advice on this work. I appreciated that they always found the time to help me, and is mostly to them that I need to thank for concluding this work, and therefore having had very joyful PhD years.

I acknowledge the funding of this research by Fundação para a Ciência e Tecnologia - FCT, through the PhD grant SFRH/BD/81204/2011.

I am thankful to the colleagues from the ATLAS group at LIP with whom I worked directly in the VH search, Alberto Palma and Mário Sousa, particularly for their help while introducing me to the work. I thank Ricardo Gonçalo and João Gentil who also collaborated in this research topic and provided most helpful contributions and were great motivators.

I extend my gratitude to all the members of the ATLAS Portuguese group, for welcoming me and being available to discuss and share ideas and follow my progress. To the colleagues in Lisboa, I also thank the companionship and the stimulating work atmosphere.

A special word of gratitude is owed to Professor Amélia Maio, for she gave me the opportunity to join this journey. It is a privilege to have her support and learn from her enthusiasm and dedication.

I address to Professor Mário Pimenta, on behalf of the LIP laboratory, my great appreciation for this institute, and I thank its members for kindly hosting me. I would also like to acknowledge the valuable technical support from the LIP IT department and secretariat.

I wish to thank the VH analysis group at CERN, and their conveners along the years, for always giving me the opportunity to present my work at the meetings and for the constructive criticism. For their help during the cut flow challenge, I thank Kenji Kiuchi, Inês Ochoa and Gabriel Facini. Special thanks should be given to Heather Gray and Giacinto Piacquadio, for their availability while helping me through the single top modelling study. For providing me the pre-selected 13 TeV data, I thank Fumiaki Ito.

Concerning my contribution to the TileCal calibration and data quality control, my thanks to Djamel Boumediene and Emmanuelle Dubreuil, for their indispensable help.

And I have to thank broadly the ATLAS Collaboration because this thesis would not have been possible without the work of all its members.

Mes remerciements à Paola, Sandra, Alex et Marc-Andrè qui m'ont fait sentir chez moi pendant mes voyages au CERN.

Obrigada aos meus amigos pelo interesse e curiosidade que sempre mostraram pelo meu trabalho: todas as Anas e Marias, Xana, Louis, Sarah, Koen, Anna, José, Célia e Jorge Grill.

Agradeço por fim a toda a minha família, especialmente a Maria de Lourdes, Maria, António, Sara, Maria Helena, Luís e Ana, pelo apoio e carinho.

Contents

Chapter 1

Introduction

Particle physics studies the elementary composition of matter and the dynamics of its constituents, governed, as known today, by four fundamental forces: Electromagnetic, Weak, Strong and Gravitational. Throughout the 1960's, the Standard Model (SM) of particles and interactions was successfully developed as a common theory integrating the Electromagnetism, and the Weak and Strong interactions, predicting the experimental measurements with an extreme accuracy from fundamental symmetry principles. However, the model was not able to conciliate the existence of massive particles with the symmetries associated with their interactions, that are both observed in nature. Besides, it predicted non-physical values for the cross-section of the WW scattering. The Electroweak Spontaneous Symmetry Breaking mechanism addresses these issues by spontaneously generating the mass of the particles without violating the conservation laws underlying the broken symmetries. By predicting a new particle and new interactions, it cures the WW scattering problem. The new particle is named after the first theorist who propose it, Peter Higgs, as the Higgs boson. Both the mechanism and the Higgs boson were included in the Standard Model of particle physics.

The discovery of a Standard Model-like Higgs boson, with an approximate mass of 125 GeV, by the ATLAS and CMS experiments at the LHC/CERN in 2012 was therefore a remarkable event in the field of experimental particle physics, as this was the only particle predicted by the SM that had not been observed. Since then, these experiments have been committed to the characterisation of this particle, leading to an experimental picture compatible with the SM. Of major relevance to this effort, is the search for the Higgs decay into a pair of $b-$quarks. Despite being the dominant decay mode, happening 57.7% of the times for $m_H = 125$ GeV, it was not observed yet and the results of such a search can still change the current understanding of the Higgs. In particular, it constitutes the best way of probing the Higgs interaction with down-type quarks at the LHC, yet to be observed, providing a test to models alternative to the SM. Additionally, given the magnitude of its branching fraction, measurements with the $H \to b\bar{b}$ channel are powerful to constrain the Higgs boson width.

The work presented on this thesis focuses on the search for the Standard Model (SM) Higgs boson decaying to $b-$quark pairs and produced in association with a W boson at the

LHC proton-proton collisions, with the ATLAS detector. The WH production is chosen for providing an effective way of triggering the $H \to b\bar{b}$ signal if the leptonic decay of the W boson is considered.

Chapter 2 presents an overview of the theoretical foundations of the Standard Model of elementary particle physics and introduces the Electroweak Spontaneous Symmetry Breaking mechanism proposed in 1964 to describe the origin of the mass of the SM particles, and that led to the prediction of the Higgs boson. The phenomenological aspects of the pp collisions taking place at the LHC and the most precise Higgs-related measurements are also discussed.

The LHC accelerates and collides protons or lead nuclei with unprecedented conditions of energy and instantaneous luminosity. For pp collisions the centre-of-mass energy reached in 2012 was 8 TeV. Four nominal collision points are equipped with large-scale detectors able to detect, identify and measure the products of the collisions. ATLAS is one of them. It comprises inner trackers, electromagnetic and hadronic calorimeters and a muon spectrometer. Combined with magnetic field systems, these sub-detectors track charged particles, and measure the energy and momentum of electrons, photons, hadrons and muons. The LHC and the ATLAS detector are presented in more detail in Chapter 3.

Chapter 4 describes the techniques employed by the ATLAS collaboration to reconstruct and calibrate final state physics objects, such as electrons, photons, muons, or jets of particles from quark or gluon hadronisation. These are the key ingredients on which any physics analysis relies on. Equally relevant to the $H \to b\bar{b}$ search is the identification of the flavour of the quark that originated the jet. $b-$tagging algorithms are used for this purpose and this topic is addressed as well.

A continuous assessment of the detector performance and operation is essential to evaluate the quality of the data taken during the LHC collisions. ATLAS integrates several independent calibration systems to monitor the full detector and ensure an early detection of any malfunctioning. Chapter 5 presents a discussion on these topics from the perspective of the TileCal hadronic calorimeter of ATLAS, and presents an algorithm that I developed to automatically identify channels with some kind of malfunctioning using data from the laser calibration system. This was my very first contribution to the experiment and consisted on the detector task to be qualified as an author of the ATLAS publications.

Chapter 6 describes the search for the $WH \to \ell\nu b\bar{b}$ signal using 20.3 fb^{-1}of $\sqrt{s} = 8$ TeV pp collision data recorded by ATLAS. The work was integrated into a broader analysis group within the ATLAS collaboration, HSG5, that additionally searches for the $H \to b\bar{b}$ decay in the ZH production mode. The design of the event selection was established by the working group with the objective of maximising the signal sensitivity and was based on simulated samples of signal and backgrounds. For the WH channel case, these are one isolated and high p_{T} electron or muon, large missing transverse energy associated with the undetected neutrino and two jets originating from $b-$quarks. I contributed to the development and validation of an independent code to perform the full WH analysis within the LIP group, implementing all the event selection conditions, corrections to simulation and data calibrations. This allowed to inter-validate the

analysis tools of the different groups participating in the WH/ZH search and provide backup inputs to the HSG5 global fit. It also established the starting point of my own contributions to the HSG5 effort and allowed me to perform the full analysis of data myself.

One of the greatest challenges of the WH search is the amount of competing background processes. Top-quark pairs and single top production, W+jets, dibosons and multijets have a final state signature very similar to the signal, and happen at a rate that is several orders of magnitude larger than the SM Higgs is expected to be produced. The signal significance $S/\sqrt{B}$ is predicted to be only 0.3 after event selection. This motivated the usage of a Multivariate approach to discriminate signal events. A Boosted Decision Tree (BDT) is used for its performance, offering also simplicity and straightforward interpretation as advantages when compared to other methods. One of my main contributions to the HSG5 analysis was a study that aimed to improve the BDT performance in the signal discrimination. It consisted of the test of new observables to be used as inputs to the BDT and concluded with the identification of two variables that improved the expected significance of the WH search by 12%. The study was an input to the ATLAS Run II analysis and is also reported in Chapter 6.

Chapter 7 describes the statistical analysis of data and presents the results. A maximum likelihood binned fit is used to measure the WH signal and simultaneously constrain the normalisation of the main backgrounds with observed data, taking into account the uncertainties affecting the search. I used the fit to validate the improvement observed in the BDT discrimination with new variables considering all the uncertainties of the analysis. The latter are also discussed. Particular emphasis is put on the uncertainties related to the modelling of the single top background by simulation, as these result from a study carried out by myself as a contribution to the HSG5 analysis.

Finally, the conclusions and a last discussion of this PhD work are drawn in Chapter 8.

Chapter 2

Theoretical and Experimental Overview of the Higgs Mechanism

This Chapter presents a summarised overview of the theoretical foundations of the Standard Model of elementary particle physics in Section 2.1, where the spontaneous symmetry breaking mechanism that led to the prediction of the Higgs boson is described.

The key phenomenological aspects of the proton-proton collisions taking place at the LHC are reviewed in Section 2.2. Moreover, this Section discusses the main mechanisms of Higgs production at the LHC. In Section 2.3, the most precise measurements of the Higgs boson properties by the ATLAS and CMS experiments are presented.

2.1 The Standard Model of Particle Physics

Fundamental particles are the basic units that constitute matter. These, along with their interactions, are described by the Standard Model (SM) of particle physics. An overview of this theoretical framework is given on this Section, starting with the presentation of its elementary particles. Then, the electromagnetic, weak and strong interactions that this model describes are briefly introduced in terms of their formal quantum field theory descriptions. Finally, the mechanism behind the generation of the mass of the SM elementary particles is explained.

This Section was written based on the textbooks referenced by [1, 2, 3], where further details not covered here can be found.

2.1.1 Particles and Interactions

The particle content of the Standard Model is listed in Table 2.1. It can be naturally divided in fermions (with spin-1/2) and bosons (with integer spin), where the spin-1 bosons enter in the SM as mediators of the interactions.

The fermionic sector is organised in quarks and leptons, each composed of three families or generations corresponding to the three columns of the Table. The quantum numbers of charge and spin are the same across generations and particles of different generation differ only on their

			Electric Charge	Spin
		Leptons		
Electron e	Muon μ	Tau τ	-1	1/2
e-neutrino ν_e	μ-neutrino ν_μ	τ-neutrino ν_τ	0	1/2
		Quarks		
Up u	Charm c	Top t	2/3	1/2
Down d	Strange s	Bottom b	-1/3	1/2
		Bosons		
Photon γ			0	1
Z, W^+, W^-			0, +1, -1	1
Gluons g			0	1
Higgs H			0	0

Table 2.1: Particle content of the Standard Model of fundamental particle physics.

mass, that increases with the family order. Electrically charged leptons have the charge of an electron but neutrinos are electrically neutral. Quarks are the only known particles that have fractional electric charge and also possess colour charge.

The identity of quarks and leptons is commonly referred to as flavour and the set of quarks composed of the u , c and t flavours are usually designated by $up-$type quarks. Conversely, the $d-$, $s-$ and $b-$flavoured quarks belong to the $down-$type ensemble. Usually, the lighter mass quarks u, d and s are designated as light quarks.

To every spin-1/2 particle there is a correspondent anti-particle with the same mass and spin, but opposite charge and quantum numbers.

Spin-1 bosons have the fundamental role of mediating the interactions between quarks and leptons and between each other. Each fundamental force described by the SM - electromagnetic, weak and strong - have then at least one boson associated to it. Electromagnetism is mediated by photons and acts upon electrically charged particles. All the quarks and leptons participate in weak interactions established by the massive weak bosons Z and $W^\pm$. Gluons are the mediators of the strong force acting only on the coloured charged quarks, and gluons themselves.

All the particles and their anti-particles described were experimentally observed. The Higgs boson was postulated in 1964 and only recently, in 2012, a new particle compatible with it was observed. This spin-0 boson is theorised in the SM as a consequence of the mechanism that generates the mass of the fundamental particles.

2.1.2 Electromagnetic Interactions

The electromagnetic interaction is responsible for the atomic structure, molecular arrangement of atoms and constitutes the basis of all optical and chemical phenomena. The relativistic and quantum theory of electromagnetism is the Quantum Electrodynamics (QED) field theory. As a Quantum Field Theory (QFT), particles are quanta of their correspondent field representations.

QED is, in the context of particle physics, more suitable described in terms of the

Lagrangian formalism, where the equations of motion of particles and the system dynamics are inscribed. The lagrangian density, from now on just referred to as lagrangian for short, of a free electron spinor field ψ is

$$\mathscr{L} = i\bar{\psi}\gamma^{\mu}\partial_{\mu}\psi - m\bar{\psi}\psi \tag{2.1}$$

where γ^{μ} are the Dirac matrices and m the electron mass. The first term is the kinetic term associated with the electron propagation and the second term, quadratic on the field, is the mass term. Using the Euler-Lagrange equations, it can be shown that this lagrangian is equivalent to the Dirac equation describing spin-1/2 particles, which has positive energy solutions describing the electron and solutions of negative energy interpreted as the electron anti-particle. Electrons and positrons have helicity eigenstates of +/-1 (right/left) according to their spin projection on the momentum vector.

The concept of symmetry plays a central role in QFT. The rules followed by interactions can all be derived from a few symmetry principles. For instance, the physical laws are intrinsically symmetric with respect to the transformations of the Poincaré group, defined as the set of space-time translations and Lorentz transformations. Poincaré invariance implies conservation of energy and momentum and invariance under rotations is connected with angular momentum conservation. The equivalence between symmetry and conservation laws is established by Noether's theorem, sometimes referred to as the "spinal cord" of the Standard Model of particle physics.

Eq. 2.1 exhibits also a gauge symmetry. The gauge terminology derives from phase and it can be easily shown that modifying the phase of the electron field, $\psi \to \psi' = e^{ie\alpha}\psi$[1], has no consequences on the initial lagrangian. However, the consequences for physics are important, since by Noether's theorem it implies electric charge conservation. The class of such gauge transformations is called the Abelian U(1) group.

If charge conservation is required not only globally but also at each space-time point, it must be assured that the lagrangian is invariant under local gauge transformations of the type $\psi \to \psi' = e^{ie\alpha(x)}\psi$, where now $\alpha \equiv \alpha(x)$. Replacing into Eq. 2.1

$$\mathscr{L} \to \mathscr{L}' = \mathscr{L} + ie\bar{\psi}\gamma^{\mu}e^{i\alpha(x)}\psi\partial_{\mu}\alpha(x) \tag{2.2}$$

one can see that the lagrangian is not invariant. This fact can be restored by making use of the covariant derivative D_{μ} instead of ∂_{μ}, defined explicitly to cancel the term breaking the local gauge symmetry as

$$D_{\mu} \equiv \partial_{\mu} - ieA_{\mu} \tag{2.3}$$

where A_{μ} is the electromagnetic vector field, identified as the field of the physical spin-1 photon. Making use of the covariant derivative D_{μ}, thus introduces the photon and its interaction with

[1] $\bar{\psi}$ transforms as $\bar{\psi} \to \bar{\psi}' = e^{-ie\alpha}\bar{\psi}$

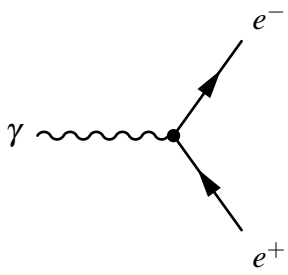


Figure 2.1: The basic QED interaction vertex.

the electron field. But for this to be coherently done, a kinetic term for the photon itself must be introduced as well. The final QED lagrangian reads

$$\mathscr{L} = i\bar{\psi}\gamma^{\mu}\partial_{\mu}\psi - m\bar{\psi}\psi + e\bar{\psi}\gamma^{\mu}A_{\mu}\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} \tag{2.4}$$

where the term involving the electromagnetic field strength tensor $F_{\mu\nu} = \partial_{\mu}A_{\nu} - \partial_{\nu}A_{\mu}$ corresponds to the photon kinetic term and gives rise to the Maxwell equations using the Euler-Lagrange principle. The third term describes the interaction between photons and electrons and relates to the basic QED vertex depicted in Figure 2.1. If the Eq. 2.4 is generalised to all electrically charged leptons and quarks, by adding their corresponding fields and interaction terms, the equivalent lagrangian contains the description of all electromagnetic interactions.

The electromagnetic interaction mediated by photons is a consequence of the the U(1) local gauge symmetry and for this reason, the photon is also called the U(1) gauge field. It is important to notice that this description of the gauge field matches the experimental evidence of massless photons. In fact, a mass term in the lagrangian would break the local gauge symmetry.

Once the QED lagrangian is set, the cross-sections for the electromagnetic processes and decay rates can be calculated, predicted and experimentally tested. The probability amplitude $\mathscr{A}$ for a transition from the initial state i to the final state f is

$$\mathscr{A}(i \to f) = (2\pi)^4\delta^{(4)}(\sum p_i - \sum p_f) \times i\mathscr{M}(p_i \to p_f) \tag{2.5}$$

where p_i and p_f are the total four-momentum associated to the initial and final state particles. The first factor encloses the conservation of four-momentum through the 4-dimentional δ function. Secondly, the amplitude function $\mathscr{M}$ holds both the dynamic information about the evolution of the system and the kinematic information related with the phase space available for it to take place. The latter corresponds to an integral in the four-momentum space over all the possible final states. On its turn, dynamics is intimately connected with the interactions described in the theory lagrangian and possible diagrams of the process. The solutions of $\mathscr{M}$ are a perturbative series in $\alpha = e^2/4\pi\hbar c = 1/137$, the QED coupling, also called fine structure constant. The leading order (LO) QED calculation, quadratic in α, is the one associated with the minimal number of QED vertices needed to describe a process, while a next-to-leading order (NLO), $\propto \alpha^4$, admits loop effects.

Figure 2.2 shows an example of a LO and a NLO diagram of the Bhabha scattering.

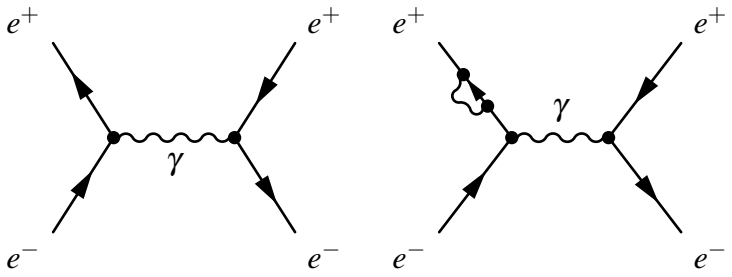


Figure 2.2: Example diagram of the Bhabha scattering at (left) LO and (right) NLO in α.

Loop effects result in divergent contributions to the cross-section amplitude expression leading to unphysical predictions. To solve this issue, the theory is renormalised. In fact, even the definition of the charge and mass of the electron is affected by loop effects, and it turns out that these theory parameters can be redefined to hide and eliminate the divergent loop terms. The same happens to the QED coupling α, that is in turn redefined as a function of the momentum Q transferred in the interaction. The $\alpha \equiv \alpha(Q^2)$ is known as the running coupling constant.

2.1.3 Strong Interactions

The strong force is what holds together neutrons and positively charged protons within atomic nuclei. This is a residual effect of what happens deeper inside the nucleons made of strongly interacting quarks. As for the electromagnetic interaction, the gauge QFT known as Quantum Chromodynamics (QCD) was developed to describe the dynamics of strong interactions between quarks. Quarks are spin-1/2 particles and the observation of spin-3/2 hadrons composed of three quarks with the same flavour, as the $\Omega^- = (sss)$ baryon, was contrary to the Pauli exclusion principle. To work around this puzzle, three additional quantum numbers were needed, and this led to the prediction of three colour charges: red, blue and green. Thus, QCD can be regarded as an extension of the U(1) group that suits the conservation of three colours, resulting in the SU(3) gauge group theory. The QCD gauge invariant lagrangian is

$$\mathscr{L}_{QCD} = \bar{q}_{j,\alpha}(i\gamma^\mu \partial_\mu - m)q_{j,\alpha} - g(\bar{q}_{j,\alpha}\gamma^\mu T_a q_{j,\alpha})G^a_\mu - \frac{1}{4}G^a_{\mu\nu}G^{\mu\nu}_a \tag{2.6}$$

where $q_{j,\alpha}$ represents the quark field with colour j and flavour α, T_a with $a = 1, ..., 8$ are the 3×3 Gell-Man matrices generators of the SU(3) group. G^a_μ are the eight gauge fields required by demanding invariance under local phase transformations of the quark fields. These are then obtained using the same principles that in QED led to the photon field, and have strength tensors $G^a_{\mu\nu}$ defined as follows

$$G^a_{\mu\nu} = \partial_\mu G^a_\nu - \partial_\nu G^a_\mu - g f_{abc} G^b_\mu G^c_\nu \tag{2.7}$$

where f_{abc} are real constants. The vector gauge fields represent the spin-1 massless gluons that mediate the strong interaction and conserve colour. The basic QCD vertex, coupling two quarks

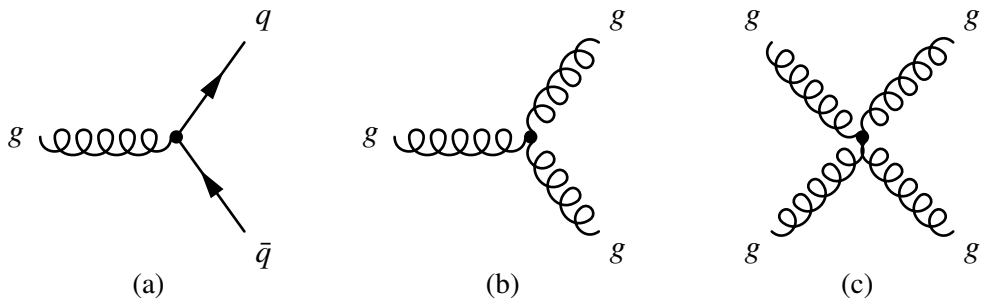


Figure 2.3: The basic QCD interaction vertices. (a) $gq\bar{q}$ vertex, (b) gluon triple coupling and (c) gluon quartic coupling.

with a gluon, shown in Figure 2.3(a), is inscribed on the second term of Eq. 2.6, where g is set to define the coupling strength.

An important distinction with respect to QED is that the last term on the QCD lagrangian encloses triple and quartic couplings between the gluons, as shown in Figures 2.3(b) and 2.3(c) respectively, thus gluons interact with other gluons. This is inevitable because in order to conserve three colour charges and to establish the interaction between different coloured quarks, the gluons are bicoloured particles, unlike photons that do not possess electric charge.

The most relevant difference between QCD and QED is, however, the running coupling behaviour. Whilst the electromagnetic coupling dependence on the momentum transfer Q is very small, the strong coupling $\alpha_s(Q^2)$ strongly depends on Q. For large Q, or short distance interactions, the strength of the colour interaction is very small and quarks and gluons are asymptotically free particles. In this regime, the perturbative series of the QCD lagrangian in α_s converges and perturbation theory can be used to calculate observables. For small Q, this is no longer valid. In this regime, α_s becomes so large that higher-order effects dominate the expansion and the theory becomes non-perturbative. Non-perturbative effects are very hard to calculate precisely with QCD and therefore models must enter into play on its turn.

It is due to the asymptotic freedom that quarks exist only confined within colourless hadrons. Whenever a quark is ripped off of a hadron, α_s increases and the strong interaction will give rise to more quarks and gluons and bind them into hadrons again. This process is known as hadronisation or fragmentation and is essentially non-perturbative. Jets are the observed manifestation of hadronisation and the experimental signature of quarks. A quark resultant from a particle decay, for instance, hadronises into collimated hadrons, which many are unstable and decay, giving rise to a jet of particles.

2.1.4 Electroweak Interaction

The successful confirmation of the QED theory through experiment came along with the ambition to unify the electromagnetic and weak interactions in a same, broader theoretical framework, formalised by Weinberg, Glashow and Salam in the 1960s. According to it, both

interactions are manifestations of the electroweak force. However, the electromagnetic coupling strength is much larger than the weak, indicating that massive bosons should intermediate the weak interaction for the electroweak unification to be achievable.

Weak charged currents are axial-vector (A-V), i.e. only couple left fermions, while weak neutral currents, as happens for QED, couple both helicity states. Both the weak and electromagnetic interactions couple leptons of the same family. This suggested that, in the context of the electroweak unification, fermions were better represented as chiral states of isospin doublets and singlets. The left-handed isospin doublets χ_L are defined for the first generation of leptons and quarks as

$$\chi_L : \begin{pmatrix} \nu_e \\ e \end{pmatrix}_L, \begin{pmatrix} u \\ d' \end{pmatrix}_L \tag{2.8}$$

and the same for the second and third generations. The down-type quarks d' are mixed representations of the physical quarks, with the mixture established through the Cabibo-Kobayashi-Maskawa (CKM) matrix. This allows incorporating flavour changing in the quark sector through charged currents. Right-handed chiral states form isospin singlets χ_R since there are no right-handed neutrinos

$$\chi_R : e_R, u_R, d'_R \tag{2.9}$$

and will only interact through neutral electroweak currents. The $U(1)_Y \times SU(2)$ is the electroweak gauge theory symmetry group. The weak hypercharge Y is conserved by $U(1)$ invariance and the weak isospin through $SU(2)$ invariance. The quantum electroweak gauge invariant lagrangian is

$$\begin{aligned} \mathscr{L}_{EW} = \bar{\chi}_{L,\beta}\gamma^\mu \left(i\partial_\mu - \frac{g}{2}\tau \cdot \mathbf{W}_\mu - g'\frac{Y}{2}B_\mu \right) \chi_{L,\beta} + \bar{\chi}_{R,\beta}\gamma^\mu \left(i\partial_\mu - g'\frac{Y}{2}B_\mu \right) \chi_{R,\beta} \\ - \frac{1}{4}\mathbf{W}_{\mu\nu} \cdot \mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} \end{aligned} \tag{2.10}$$

where β runs over the correspondent chiral states of the three families of quarks and leptons. τ are the Pauli matrices generators of the SU(2) symmetry group, and g and g' tune the coupling strengths. $\mathbf{W}_\mu = (W^1_\mu, W^2_\mu, W^3_\mu)$ and B_μ are the four massless gauge fields with corresponding strength tensors $\mathbf{W}_{\mu\nu}$ and $B_{\mu\nu}$. The latter compose the third and fourth terms of the lagrangian to define the free propagation and self-interaction vertices of the gauge bosons.

The physical photon and $W^\pm$ and Z bosons fields are defined as a mixture of the electroweak fields

$$\begin{aligned} W^\pm_\mu &= (W^1_\mu \mp W^2_\mu)/\sqrt{2} \\ A_\mu &= \cos\theta_W B_\mu + \sin\theta_W W^3_\mu \\ Z_\mu &= -\sin\theta_W B_\mu + \cos\theta_W W^3_\mu \end{aligned} \tag{2.11}$$

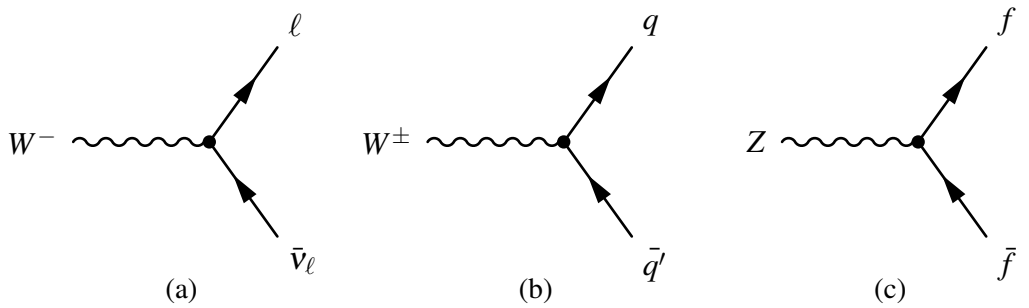


Figure 2.4: Weak fermion couplings: (a) $W\ell\bar{\nu}_\ell$, (b) $Wq\bar{q'}$ and (c) $Zf\bar{f}$ vertices.

where θ_W is the Weinberg angle. The electroweak lagrangian of Eq. 2.10 can then be written explicitly in terms of the physical fields. The weak charged currents sector reads

$$\mathscr{L}_{EW,CC} = -\frac{g}{\sqrt{2}} W_\mu^- \left(\bar{\nu}_{\ell,f} \gamma^\mu (1-\gamma^5) \ell_f + \bar{u}_f \gamma^\mu (1-\gamma^5) d_f \right) + \text{h.c.} \tag{2.12}$$

where ν_ℓ and ℓ are the spinor fields of neutrinos and charged leptons and u and d represent respectively the up- and down-type quarks spinor fields. The f index runs over each generation of fermions and h.c. refers to the hermitian conjugate expression involving the W_μ^+ field. This lagrangian describes the original A-V nature of the charged weak interaction, with the W field coupling exclusively left-handed chiral states, through the $\gamma^\mu(1-\gamma^5)$ structure of the couplings. The corresponding vertices are shown in Figures 2.4(a) and Figures 2.4(b).

In a similar manner, one obtains the lagrangian containing the electroweak neutral interactions in terms of the physical A_μ and Z_μ fields

$$\mathscr{L}_{EW,NC} = -\bar{\psi}_\beta \gamma^\mu \left[A_\mu (g \frac{\sigma^3}{2} \sin\theta_\text{W} + \text{g}' \frac{\text{Y}}{2} \cos\theta_\text{W}) + \text{Z}_\mu (\text{g} \frac{\sigma^3}{2} \cos\theta_\text{W} - \text{g}' \frac{\text{Y}}{2} \cos\theta_\text{W}) \right] \psi_\beta \tag{2.13}$$

here ψ are the fermion spinors with f indexing all fermions. While the nature of charged currents is A-V, neutral weak currents and QED present vector-like symmetries in the electroweak description. The weak neutral coupling vertices are shown in Figure 2.4(c). Although not represented in the Eq. 2.12 and 2.13, the electroweak model describes further interaction terms involving the photon and the weak bosons as predicted by the properties of these particles, namely the couplings between weak vector bosons - $Z_\mu W_\mu^- W_\mu^+$, $Z_\mu Z_\mu W_\mu^- W_\mu^+$ and $W_\mu^- W_\mu^+ W_\mu^- W_\mu^+$ - and the couplings mixing the photon and weak fields - $A_\mu W_\mu^- W_\mu^+$, $A_\mu A_\mu W_\mu^- W_\mu^+$ and $A_\mu Z_\mu W_\mu^- W_\mu^+$.

Although on one hand the electroweak unification was a successful effort, it came at the cost of not being able to incorporate the particles masses. Since left-handed fermions are isospin doublets and right-handed fermions are singlets, and these representations have different gauge transformations, a mass term for fermions is no longer gauge invariant. In addition, mass terms for the gauge fields would also break the $U(1)_Y \times SU(2)$ symmetry.

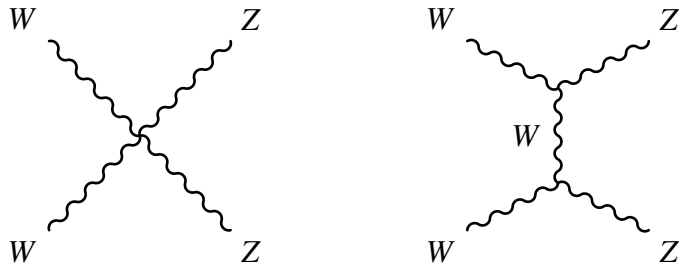


Figure 2.5: Diagrams contributing to the diboson scattering process at LO.

This not only contradicts experimental observation but also the a priori notion that the $W^{\pm}$ and Z bosons need to be massive mediators to conciliate the weak interaction with the much stronger electromagnetic coupling constant. And indeed, this fact leaves a door open for a spontaneous break of the $U(1)_Y \times SU(2)$ symmetry, where the electroweak force decomposes into separate electromagnetic and weak interactions by generating the weak bosons masses, while maintaining hidden the original symmetry of the lagrangian in what concerns the associated conservation laws.

But other facts hinted to the incompleteness of the electroweak model. According to it, the cross-section of some processes violated the unitarity principle as is the case of the diboson scattering process exhibited in Figure 2.5. The calculations for the amplitude of this process result in $\sigma_{WW\to ZZ} \propto Q^2$ and led to the suspicion that other diagrams involving other particles should exist for cancellations to be possible and the theory to be renormalisable.

2.1.5 Electroweak Spontaneous Symmetry Breaking

A spontaneous symmetry breaking mechanism was proposed in 1964 in three independent papers by R. Brout and F. Englert [4], P. W. Higgs [5] and G. Guralnik, C. R. Hagen, and T. Kibble [6] and incorporated in the electroweak theory by Glashow and Salam in 1967.

The $U(1)_Y \times SU(2)$ symmetry breaking mechanism introduces a complex scalar field isospin doublet ϕ with four degrees of freedom $(\phi_1, \phi_2, \phi_3, \phi_4)$

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix} \tag{2.14}$$

and potential $V(\phi)$ given by

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

into the gauge sector of the electroweak lagrangian as follows

$$\mathscr{L}_{EW\phi} = (D^\mu\phi)^\dagger(D_\mu\phi) - V(\phi) \tag{2.16}$$

In Eq. 2.15, λ and μ are free parameters of the model and μ^2, affecting the term quadratic

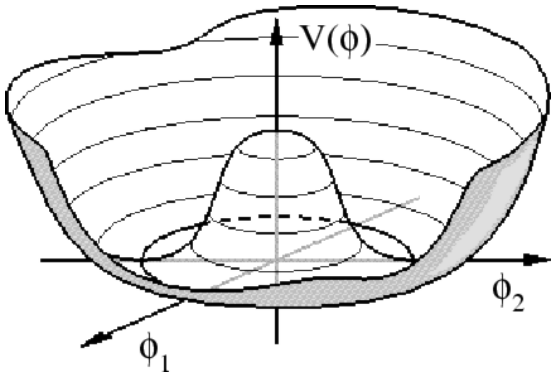


Figure 2.6: Representation of the scalar field potential $V(\phi)$ in 2 dimensions (ϕ_1, ϕ_2).

in the field ϕ, is related with the scalar mass. Since the scalar field is an isospin doublet, the lagrangian of Eq. 2.16 is invariant under $U(1)_Y \times SU(2)$ transformations.

For $\mu^2 < 0$, the field potential has a degenerate minimum, as can be seen from Figure 2.6.

The minimum can be fixed at, for instance, $\phi_0 : (0,0,\phi_3 = v,0)$, where v is the vacuum expectation value (vev), the value of the field yielding the minimum potential $V(\phi_0) = 0$. With this choice, only the neutral component of the doublet ϕ_0 is non-vanishing leading to an electrically neutral vacuum, without loss of generality since the system can always be rotated in the $SU(2)$ space. By allowing small perturbations h around this minimum state, the scalar doublet reads

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v+h \end{pmatrix} \tag{2.17}$$

The system apparently loses three degrees of freedom and the remaining one is associated with a physical boson, the Higgs particle h. The particle is named after P. W. Higgs, the only theorist that predicted a new particle following the spontaneous breaking of the $U(1)_Y \times SU(2)$ symmetry. The $V(\phi_0) = 0$ condition also results in the following relation between the model parameters

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

Mass of the gauge bosons

Now the striking point is how the electroweak gauge lagrangian evolves at this potential minimum with small perturbations h. By explicitly writing the electroweak covariant derivative in terms of the electroweak fields, $D_\mu = \partial_\mu + ig\frac{g}{2}\tau \cdot \mathbf{W}_\mu + ig'\frac{Y}{2}B_\mu$, and recalling their relation

with the physical bosons given by Eq. 2.11, the Eq. 2.16 results in[2]

$$\mathscr{L}_{EW\phi} = \frac{1}{2}(\partial^{\mu}\phi)^{\dagger}(\partial_{\mu}\phi) + \frac{1}{8}(v+h)^2[g^2(W_{\mu}^{+})^2 + g^2(W_{\mu}^{-})^2 + (g^2+g'^2)Z_{\mu}] - V(\phi) \quad (2.19)$$

from where can be seen that terms proportional to $W_{\mu}^{\pm 2}$ and Z_{μ}^2 arose. These are identified with mass terms for the $W^{\pm}$ and Z bosons, that break the $U(1)_Y \times SU(2)$ invariance. Their masses are given by

$$\begin{aligned} M_{W^-} &= M_{W^+} = \frac{1}{2}vg \\ M_Z &= \frac{1}{2}v\sqrt{g^2+g'^2} \end{aligned} \quad (2.20)$$

The photon field remains massless in the theory since there is no mass term proportional to A_{μ}^2. So, by introducing the isospin doublet ϕ, with degenerate minima, the $U(1)_Y \times SU(2)$ is broken spontaneously when ϕ adopts the configuration of minimum potential. The field itself loses three degrees of freedom, that are transformed into the longitudinal polarisation, or mass, of the weak mediators.

The mechanism also predicts that the vacuum is permeated with a scalar field and vev of v=246 GeV. This value is obtained from the muon decay width, that allows to determine the strength of the weak interaction, and making use of the M_W expression in Eq. 2.20. Furthermore, the mechanism also predicts the interaction between the Higgs field h and the massive bosons, no direct coupling with the photon and Higgs self-couplings.

Fermion masses

In the context of the electroweak unified theory, the fermion mass terms are not gauge-invariant as discussed in Section 2.1.4. But contrary to the gauge bosons, their masses result from the introduction of the following Yukawa coupling to the Higgs field

$$\mathscr{L}_{Yukawa} = -\lambda_f(\bar{\chi}_{L,f}\phi\chi_{R,f} + \bar{\chi}_{R,f}\phi\chi_{L,f}) \quad (2.21)$$

with strength λ_f, to be determined for each fermion flavour f. The Yukawa lagrangian for leptons results in

$$\begin{aligned} \mathscr{L}_{Yukawa-leptons} &= -\frac{1}{\sqrt{2}}\lambda_{\ell}\left[(\bar{\nu},\bar{\ell})_L\begin{pmatrix}0\\ v+h\end{pmatrix}\ell_R + \bar{\ell_R}(0,v+h)\begin{pmatrix}\nu\\ \ell\end{pmatrix}_L\right] \\ &= -\frac{\lambda_{\ell}(v+h)}{\sqrt{2}}[\bar{\ell}_L\ell_R + \bar{\ell}_R\ell_L] = -\frac{\lambda_l}{\sqrt{2}}v\bar{\ell}\ell - \frac{\lambda_l}{\sqrt{2}}h\bar{\ell}\ell \end{aligned} \quad (2.22)$$

where it can be identified a mass term for leptons plus an interaction term with the Higgs field. Through the inclusion of the Yukawa coupling, the resultant mass term is, however,

[2]Spurious terms proportional to $\partial_{\mu}\phi Z^{\mu}$, for instance, were discarded for not having physical meaning.

gauge invariant, since the chiral states are recombined into the original QED spinor fields, here represented by ℓ. The mass of the leptons is given by $m_\ell = \lambda_\ell v/\sqrt{2}$ and since λ_ℓ is a free parameter, the model does not predict the lepton masses and presents no attempt to justify their values.

In the quark sector is as not straightforward to obtain the mass terms due to quark mixing, but in the end the result is similar to the one presented for leptons. According to the Higgs mechanism, the mass of the fermions is given by

$$m_f = \lambda_f v/\sqrt{2} \tag{2.23}$$

2.1.6 Final Standard Model Lagrangian

The Standard Model of particle physics thus includes all the fundamental particles observed up to now and describes their interactions based on the $U(1)_Y \times SU(2) \times SU(3)$ symmetries of the electroweak and strong sector. The masses of the particles are obtained from spontaneous breaking the $U(1)_Y \times SU(2)$ symmetry from what results the Higgs boson and its interactions with massive particles. The final SM lagrangian comprehends the pieces shown so far in Eq. 2.10, 2.16, 2.21 and on the gauge sector of 2.6

$$\mathscr{L}_{SM} = \mathscr{L}_{QCD,gauge} + \mathscr{L}_{EW} + \mathscr{L}_{EW\phi} + \mathscr{L}_{Yukawa} \tag{2.24}$$

where $\mathscr{L}_{QCD,gauge} = -g(\bar{q}_{j,\alpha}\gamma^\mu T_a q_{j,\alpha})G^a_\mu - \frac{1}{4}G^a_{\mu\nu}G^{\mu\nu}_a$, with the quark propagation and mass terms already contained in $\mathscr{L}_{EW}$ and $\mathscr{L}_{Yukawa}$, respectively.

Although the Standard Model demonstrates great agreement with the experimental observation up to now, it is far beyond giving an explanation to the major questions of particle physics at the moment, such is the nature of Dark Matter and Dark Energy and the origin of matter/anti-mater asymmetry in the Universe. And in particular, in what concerns the Higgs mechanism, it is important to notice that it does not accommodate the mass of neutrinos, an evidence provided by the observation of neutrino oscillations.

2.1.7 Couplings and Properties of the Higgs boson

The discovery of a new boson compatible with the SM Higgs boson in July 2012 by the ATLAS and CMS experiments at LHC [7, 8] makes the characterisation of the particle found the major priority in the Higgs research field. The strength of the couplings of the Higgs boson to other particles, a growing precision on the mass measurement, and spin and parity measurements are the main experimental results needed to probe the Higgs mechanism.

Couplings

The interaction term of the Higgs field with the gauge bosons is given in Eq. 2.16. The correspondent vertices and coupling strengths are shown in Figure 2.7. The Higgs can have a

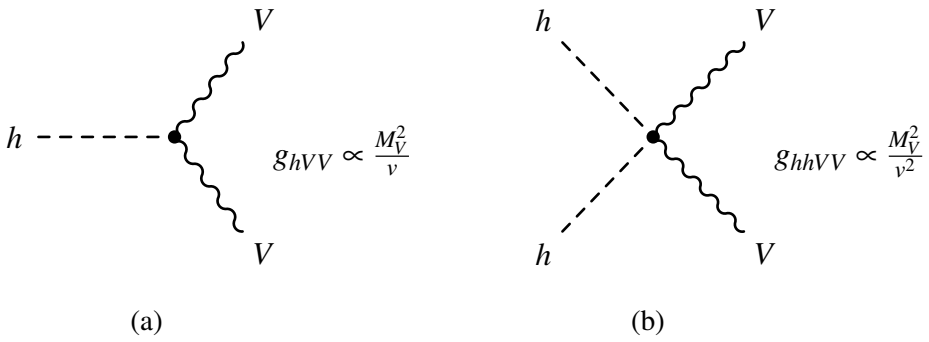


Figure 2.7: Higgs-vector bosons (a) triple and (b) quartic interaction vertices and coupling strengths. V represents either a W or a Z.

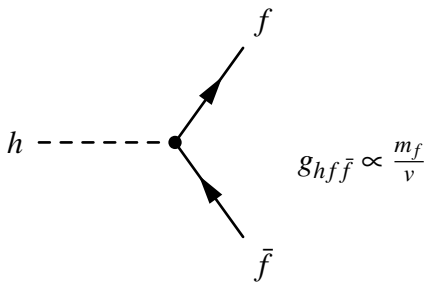


Figure 2.8: Higgs-fermion interaction vertex and coupling strength. f represents either a quark or an electrically charged lepton.

triple or quartic coupling to massive vector bosons with coupling strength proportional to M_V^2.

The vertices and coupling strengths between the Higgs and the fermions, exhibited in Figure 2.8, are obtained in the same way by examining the Yukawa lagrangian of Eq. 2.21. Only triple couplings are predicted by the theory, with coupling strength proportional to the fermion mass m_f.

Concerning the Higgs boson itself and its properties, the particle spectra of the field ϕ is analysed by expanding its potential and recalling the relation of Eq. 2.18

$$\begin{aligned} \mathscr{L}_\phi &= \frac{1}{2}(\partial^\mu \phi)^\dagger (\partial_\mu \phi) - V(\phi) \\ &= \frac{1}{2}(\partial^\mu \phi)^\dagger (\partial_\mu \phi) - \frac{1}{4}v^2 - \lambda v^2 h^2 + \frac{3}{2}\lambda v h^3 - \frac{3}{2}\lambda h^4 \end{aligned} \tag{2.25}$$

This is the lagrangian of a spin-0 particle of positive parity, with the first term describing its free propagation and a mass term, proportional to h^2, yielding $m_h = \sqrt{2\lambda v^2}$. The terms proportional to h^3 and h^4 describe the Higgs triple and fourth self-couplings. These are presented in Figure 2.9 and, as happens for the gauge bosons, the coupling strength is proportional to the squared boson mass.

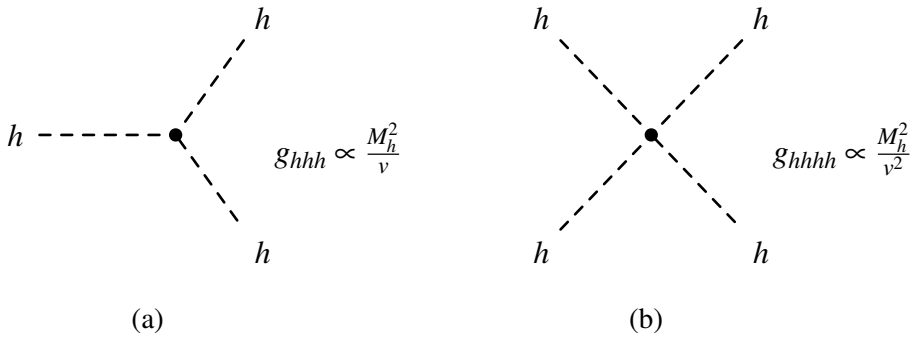


Figure 2.9: Higgs bosons (a) triple and (b) quartic self-interaction vertices and coupling strengths.

Mass

Since λ is a free parameter of the model, the Higgs boson mass, given by $m_h = \sqrt{-2\lambda v^2}$, is also not predicted. However, several theoretical arguments set constraints on this quantity. First off, one of the theoretical successes of the Higgs mechanism is that the Electroweak theory becomes renormalisable. The subject was briefly discussed in Section 2.1.4. With the Higgs mechanism, the diagrams involving hVV vertices contribute to the diboson scattering, depicted in Figure 2.5, curing the amplitude divergence if and only $m_h < 700$ GeV. This reasoning is called unitarity. Above this value, the Higgs can still exist but it comes at the price of not fixing this issue.

Another argument is related to the running of the Higgs λ coupling, that increases with the cut-off scale Λ of the SM validity. For large values of Λ, λ has a singularity known as the Landau Pole. Avoiding that singularity results in a upper limit on m_h. On the opposite sense, requiring the coupling to be positive prevents the Higgs potential to have an unstable minimum and leads to a lower bond on m_h. Figure 2.10 shows the allowed Higgs mass range as a function of the cut-off scale Λ. It also justifies the importance of research in the Higgs sector, as the measurement of the Higgs mass gives insight to the SM validity and to the scale where new physics is expected.

Furthermore, the experiments carried on at the e^+e^- collider LEP and at the $p\bar{p}$ collider Tevatron were able to more strictly unveil the Higgs mass. Precision measurements of the electroweak parameters are sensitive to the Higgs mass given the fact that m_h enters in the radiative corrections to the top and W mass. By additionally including the constraints to the Higgs mass from the direct searches at LEP and Tevatron, a combined statistical analysis of the data resulted in the prediction of $m_h = 116.4^{+18.3}_{-1.3}$ GeV [10].

Higgs decay

The Higgs boson decays directly to fundamental massive particles and, through massive particle loops, to gluons and photons. Figure 2.11 shows the most important diagrams of the

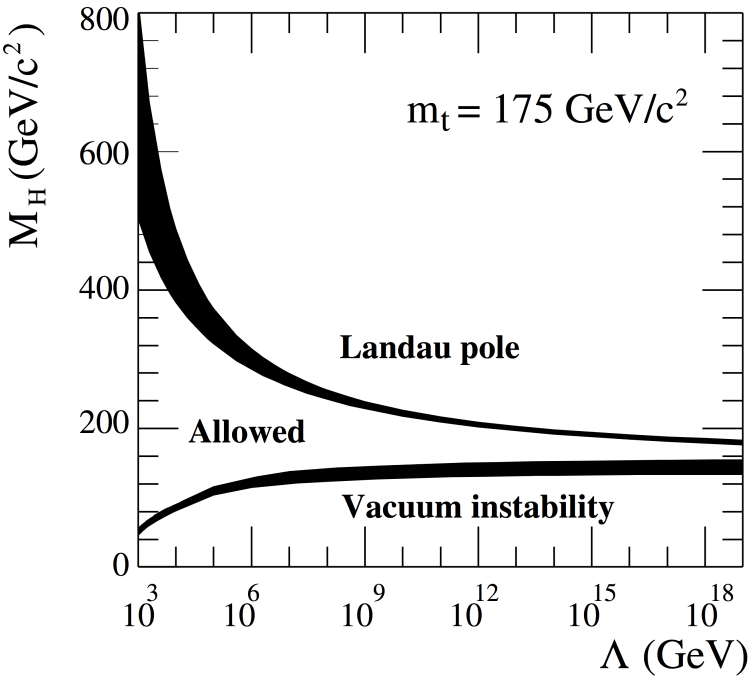


Figure 2.10: SM Higgs mass bounds as a function of the Λ cut-off scale. The upper limit is set by the Landau pole and the lower limit by vacuum stability. Taken from [9].

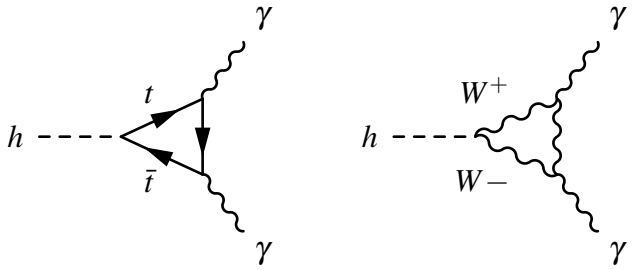


Figure 2.11: Diagram of the $h \to \gamma\gamma$ decay at LO.

Higgs decay to photons. The decay to gluons is similar but does not involve the W loop since gluons do not couple to W bosons.
The partial width Γ of the Higgs decay to massive particles is given by

$$\begin{aligned} \Gamma(h \to f\bar{f}) &\propto m_f^2 m_h \sqrt{1-x} && \text{, with } x = 4m_f^2/m_h^2 \\ \Gamma(h \to VV) &\propto m_h^3 (1 - x + \frac{3}{4}x^2) \sqrt{1-x} && \text{, with } x = 4m_V^2/m_h^2 \end{aligned} \tag{2.26}$$

Γ increases with the decay daughters mass as a result of the Higgs coupling strength proportional to m_f and m_V^2. On another hand, it depends on the available phase space, inscribed in the $\sqrt{1-x}$ factor above, that benefits decays to lighter particles. The Higgs branching ratio defined as the ratio of the partial to the total width is shown in Figure 2.12 as a function of the scalar mass. The interplay between the coupling strength and available phase space is clearly visible. If there is sufficient energy available, i.e. for large Higgs masses, the

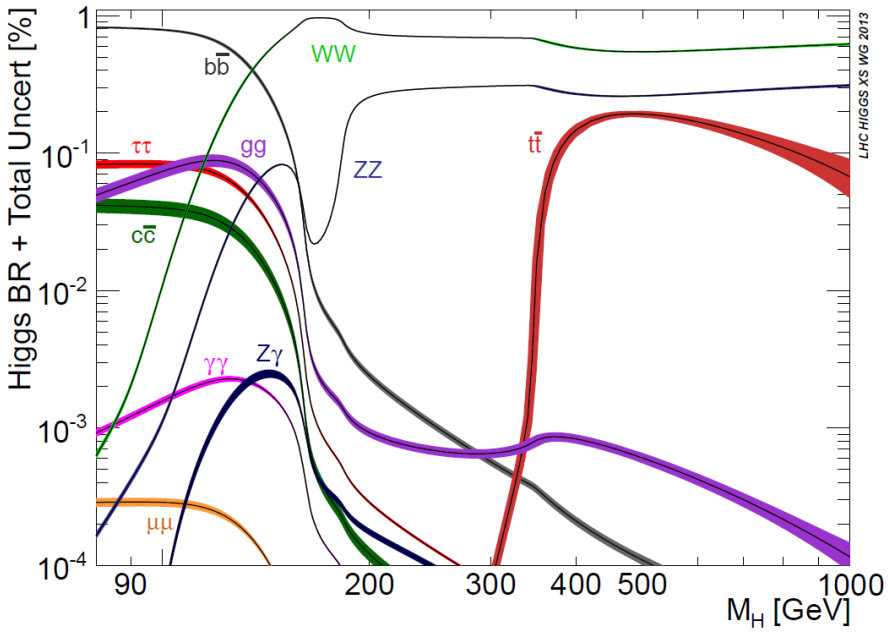


Figure 2.12: SM Higgs branching ratio as a function of the Higgs mass. Taken from [11].

Higgs decay preference is to W, Z and top-quark pairs. Then, the $t\bar{t}$ mode is suppressed at $m_h < 2m_t = 350$ GeV and the same happens to the dibosonic modes at $m_h < 2m_V$. The decays can result in off-shell particles, and that is why below these values the partial widths are small but do not reach the absolute zero. A lighter Higgs decays predominantly to bottom quark pairs, the heaviest elementary particle excluding the ones already mentioned and the Higgs boson itself.

The uncertainties on the branching ratios, shown in Figure 2.12 come from missing higher-order corrections and uncertainties on the theory parameters, such as the masses of the decay products and coupling constants.

2.2 Higgs Phenomenology at the LHC

This section introduces the main phenomena occurring in proton-proton collisions and the main mechanisms of Higgs production at the LHC.

2.2.1 Proton-proton collisions

Figure 2.13 depicts the essential phenomena underlying a pp collision event at the LHC. Protons are compound particles made of the uud valence quarks and a sea of quarks and gluons resultant from strong interactions between partons. When two protons collide, the most energetic inelastic reaction, also called hard scatter, usually involves only a parton from each proton. The remnant of the proton participates in softer interactions denominated underlying event. Moreover, initial and final state partons frequently undergo soft gluon emission, a phenomenon known as initial and final state radiation (ISR and FSR), respectively.

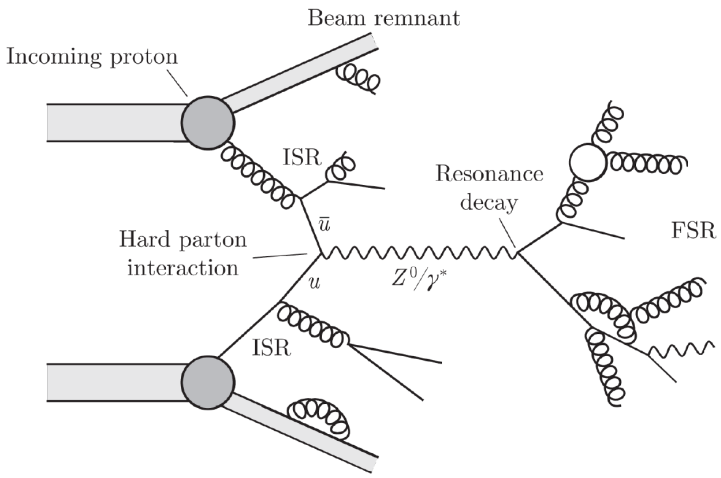


Figure 2.13: Schematic view of a proton-proton collision.

What the detector sees are the final state particles of these reactions or stable decay daughters of the short-lived resonances produced. Quarks and gluons are a special kind of final states. Due to the asymptotic freedom of QCD, final state quarks and gluons, either resultant from radiation or resonance decays, go through the non-perturbative hadronisation or fragmentation process leading to a single hadron or a spray of particles, called jet. Jets are the typical signature of free quarks and gluons in the detector.

The pp collisions at the LHC are provided by the crossing of proton bunches. More than one collision can occur in a single bunch crossing. This multiple interaction effect is known as pile-up and, as will be discussed in Section 3.1, has severe implications on the detection and analysis of events.

Hard Scatter

The hard scatter is the process of largest energy transfer, or larger Q^2, and therefore where the interesting physics events originate. The number of events N of an arbitrary process $pp \to X$ yielding the X final state is the product of the cross-section $\sigma_{pp\to X}$ by the integrated luminosity L, with the latter given by the time integral of the instantaneous luminosity $\mathscr{L}$, characteristic of the accelerator and defined later on Section 3.1

$$N = \sigma_{pp\to X} \times L = \sigma_{pp\to X} \times \int \mathscr{L} \mathrm{dt} \tag{2.27}$$

However, since the initial state particles of hadron colliders are the partons compounding the hadrons, the calculation of $\sigma_{pp\to X}$ must take into account the probability of finding the initial state partons a and b carrying a fraction x_a and x_b of the momenta of the colliding hadrons (designated parton distribution function PDF), and the partonic cross-section $\hat{\sigma}_{ab\to X}$.

While the hard scatter reaction $ab \to X$ can be described by perturbative QCD, given that α_s is small for large Q^2, soft interactions happening inside protons prevent perturbative QCD to be used to determine precisely the PDFs. Combining both aspects to describe $\sigma_{pp\to X}$

is therefore not straightforward. The relation is established by the factorisation theorem [12] stating that perturbative and non-perturbative effects can be factorised at a fixed scale, assuming a non-interference approximation. According to it, the cross-section is given by

$$\sigma_{pp\to X} = \sum_{a,b} \int dx_a dx_b f(x_a, \mu_F^2) f(x_b, \mu_F^2) \hat{\sigma}_{ab\to X}(x_a p_a, x_b p_b, \mu_F^2, \mu_R^2) \tag{2.28}$$

The sum runs over every parton able to initiate the process and the $f(x, \mu_F^2)$ functions represent the PDFs. The $\hat{\sigma}_{ab\to X}$ dependence on the parton momentum is given by the xp factor, where p is the proton momentum. Both the PDFs and the cross-section have an explicit dependence on the factorisation scale μ_F separating the perturbative and non-perturbative regimes. The renormalisation scale μ_R is the scale until QCD is renormalisable and usually defines μ_F itself. The definition of these scales is a source of theoretical uncertainty on the prediction of the expected number of events for a given process.

The partonic cross-section is calculated as a perturbation expansion on the electroweak and strong coupling constants. LO calculations exist for the great majority of processes happening at the LHC, but beyond LO calculations are available only for a limited number of processes. Typically, higher-order corrections to total cross-sections are calculated, but they neglect the dependence of the correction on the kinematics of the final state particles, i.e., they only scale the lower-order differential cross-section. Since α_s is much greater than α_{EW}, the most relevant higher-order corrections to consider in LHC processes are usually related with QCD.

The PDFs are determined in deep inelastic scattering and hadron-hadron collision experiments. The cross-sections calculated for the LHC use the PDFs based mainly on the Tevatron and HERA data.

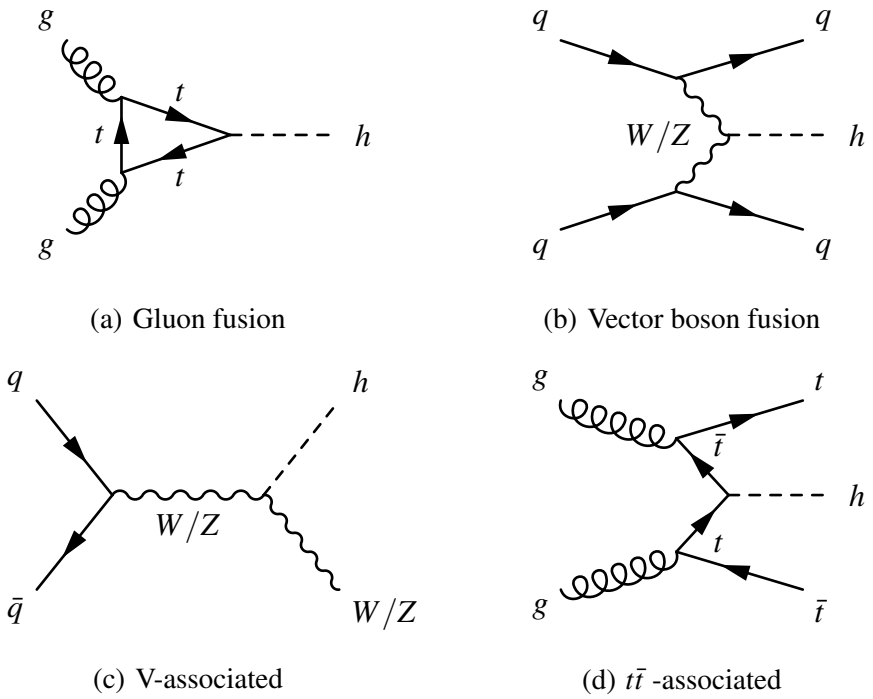
Hadronisation

Hadronisation is essentially a non-perturbative effect, so phenomenological models have to be used instead. The most used models are the string model and the cluster model, and usually, the uncertainties on hadronisation are determined by physics analysis by comparing the predictions between the two.

In the string model, a $q\bar{q}$ pair is regarded as connected by a string [13]. When the pair moves apart, the potential energy in the string increases and eventually breaks originating a new $q'\bar{q}'$ pair. The system splits into the $q\bar{q}'$ and $q'\bar{q}$ systems composed of a mixture of the original quarks with the new pair of quarks. Each of them can either form an on-shell mass hadron or split again giving rise to an extra $q\bar{q}$ pair. The process continues until only on-shell hadrons are present.

The basic idea behind the cluster model [14] is the formation of parton clusters defining colour singlets, mostly through gluon splitting, that then decay into the observed hadrons.

Both models were derived from experimental data and their parameters can be tuned to provide better predictions. The tuning aspects include the dynamics of outgoing hadrons, their



(a) Gluon fusion

(b) Vector boson fusion

(c) V-associated

(d) $t\bar{t}$ -associated

Figure 2.14: Main LO diagrams of Higgs production at the LHC.

multiplicity and flavour fraction, and the resulting jet shape.

2.2.2 Higgs Production

The main mechanisms of Higgs production at the LHC are shown in Figure 2.14, and the corresponding cross-section as a function of the Higgs mass is exhibited in Figure 2.15. These are the gluon fusion ggF, vector boson fusion VBF, associated production with a W or Z vector boson WH or ZH, respectively, or with a top and anti-top pair ttH. Since the Higgs does not couple directly to gluons, gluon fusion involves a loop of virtual elementary particles, that should mainly be top quarks for having the larger coupling strength with the Higgs. The cross-section of these reactions depends on the PDF of the colliding protons and on the coupling strengths of the involved vertices. So, for instance, even if the $u\bar{u} \to h$ process is extremely favoured by the proton valence composition, the fact that the $u-$quark is extremely light prevents it from having a large coupling with the Higgs and the cross-section for this process is negligible.

The calculation of the cross-sections for Higgs production at the LHC starts with a fixed-order calculation in QCD. Then, corrections up to NLO in Electroweak and NNLO QCD are added to the total cross-section for ggF, VBF, WH and ZH. The ttH cross-section is corrected at NLO in QCD. The uncertainties include uncalculated higher-order corrections, uncertainties on the theory parameters and on the PDFs.

Figure 2.16 shows the magnitude of the NLO and NNLO QCD corrections to the WH production cross-section. The NLO increases the total WH LO cross-section by $\sim$20%, but

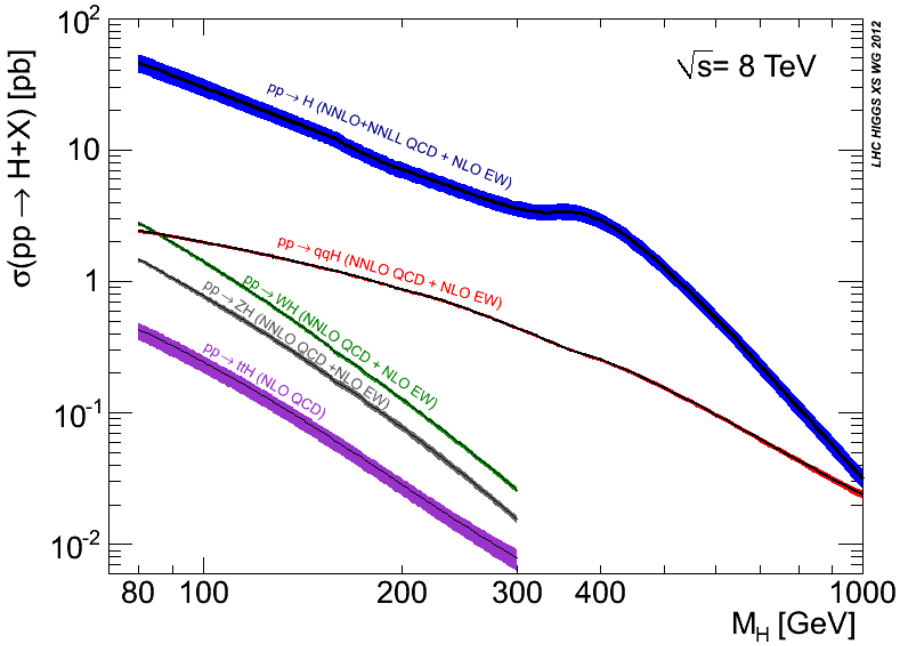
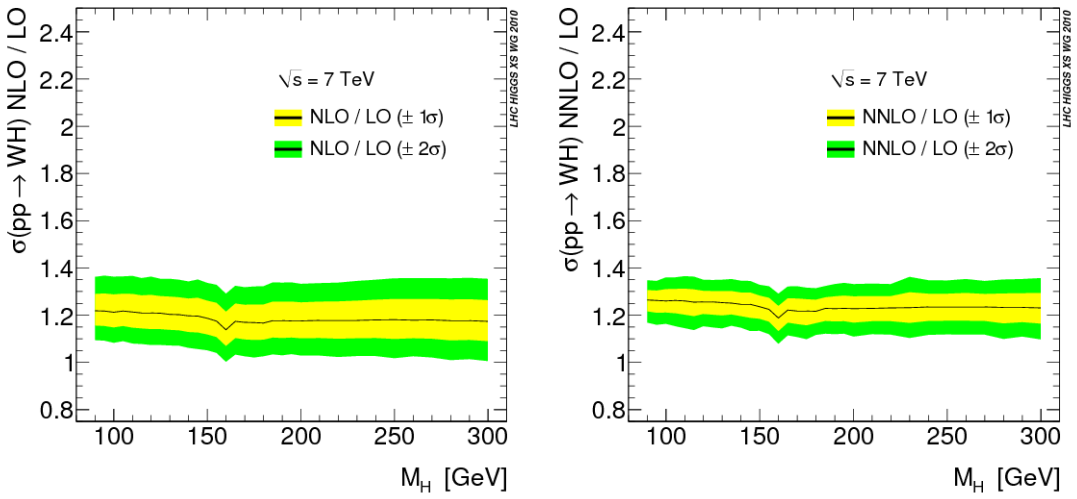


Figure 2.15: SM Higgs production cross-section as a function of the Higgs mass. Taken from [11].

considering NNLO only adds 2% more, indicating that higher-orders corrections converge fast. For processes dominated by QCD dynamics, higher-order corrections are far more relevant. In ggF for instance, NLO corrections increase the LO fixed calculation by 80 to 100% and NNLO by 25% more. As a consequence, the uncertainties due to higher-order uncalculated corrections are larger too. So, as pure QCD processes ggF and ttH production have the largest uncertainties, while VH and VBF, involving exclusively weak interactions, have the smaller theoretical uncertainties.

At the LHC, the Higgs is produced mainly by fusion of gluons due to gluon availability, the cross-section of this process being larger by more than a factor of 10 relative to the second most frequent: vector boson fusion. The enhancement of the $\sigma(ggF)$ at $m_h \sim 2m_t$ is due to the possibility of having a real top loop. In vector boson associated production, an off-shell W or Z radiates a Higgs. Since the W production cross-section is larger than the Z production and the Higgs couples more strongly to the W, $\sigma(WH)$ is larger than $\sigma(ZH)$. The ttH mode is, from the dominant mechanisms, the one expected to happen less frequently.



(a) (b)

Figure 2.16: NLO and NNLO QCD corrections (ratio to the LO prediction) to the total cross-section of the WH production at $\sqrt{s} = 7$ TeV pp collisions, as a function of the Higgs mass. Taken from [15].

2.3 ATLAS and CMS measurements

The e^+e^- collider LEP and the $p\bar{p}$ collider Tevatron searched directly for the Higgs boson from the 1980s to 2011. Having not discovered the Higgs, the experiments excluded the Higgs mass range of $m_H < 114.4$ GeV [16] and from 156 to 177 GeV [17]. This established the starting point for the Higgs search at the LHC when it began operating in 2011, colliding protons at the center-of-mass energy of $\sqrt{s} = 7$ TeV in 2011 and of $\sqrt{s} = 8$ TeV later in 2012. In July 2012, the ATLAS and CMS experiments announced the discovery of a Higgs-like particle at the LHC, with a mass of approximately 125 GeV [7, 8]. Later on the same month, the Tevatron experiments have published a combined analysis of the $p\bar{p}$ collision data yielding an evidence for a new particle in Higgs searches [18].

The principal reason why it took so long to observe the Higgs boson is better explained by Figure 2.17. One can see that the cross-section lines for the processes involving the production of a Higgs sit many orders of magnitude below multiple competing reactions. In other words, there is a huge background to overcome. For this reason, the searches for the Higgs combine a specific production mode with a specific decay mode, in what is called a channel, such that the signal event has a well-defined final state of particles and a clear signature. In this manner, the analyses can be tailored to fit the signal search and deal more efficiently in terms of background rejection.

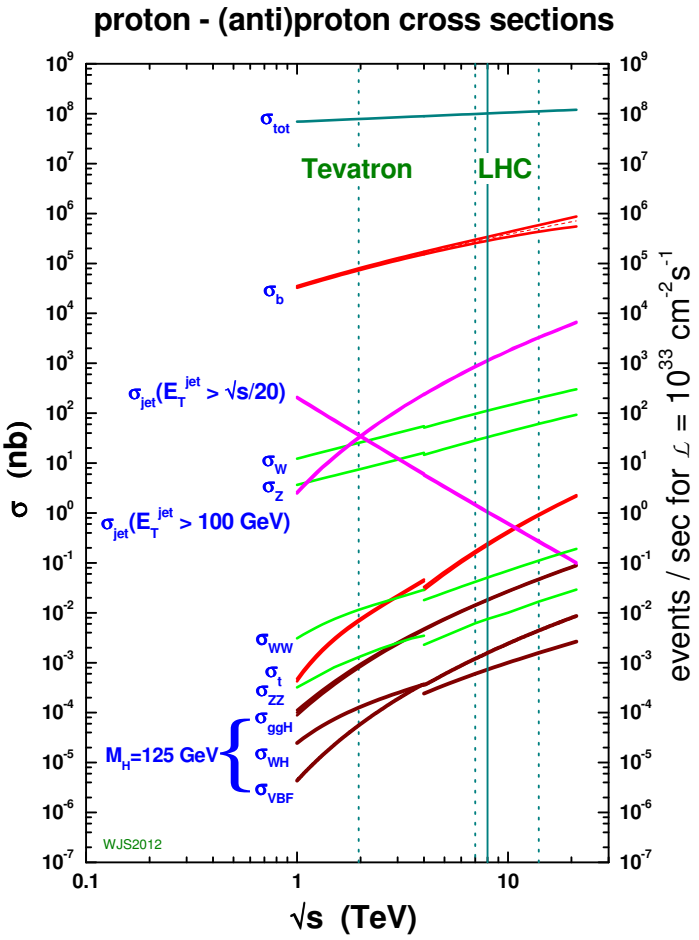


Figure 2.17: Expected cross-sections for proton-proton (LHC) and anti-proton-proton (Tevatron) collisions as a function of the center of mass $\sqrt{s}$ energy. The vertical solid line is drawn at $\sqrt{s} = 8$ TeV. From J. Stirling [19].

2.3.1 Search channels

The search channels containing the Higgs decay to two photons, Z bosons and W bosons were the first bets of the ATLAS and CMS experiments to discover the Higgs with the LHC early data because of their clean signatures in the detector. For the same reason, these searches were performed in an inclusive way considering the production mode, meaning that they did not targeted any specific production mechanism although, as seen, ggF dominates. The $H \to \gamma\gamma$ and $H \to ZZ$ (with $Z \to e\bar{e}/\mu\bar{\mu}$) have the additional advantage of providing the best resolution for the Higgs mass measurement.

The $H \to b\bar{b}$ decay, which from first principles should be easier to observe due to the large branching ratio (for $m_H = 125$ GeV the $BR(H \to b\bar{b})$ is 57.7%) is, in fact, one of the most difficult ways to observe the Higgs. This is the case because, in hadron colliders, jet production happens in nearly every event, and therefore the signal-to-background ratio is a prohibitive limitation. To overcome the issue, this decay mode is searched for in channels where the Higgs is produced in association with other particles, as is the case of WH, ZH, ttH, VBF or VBF+γ,

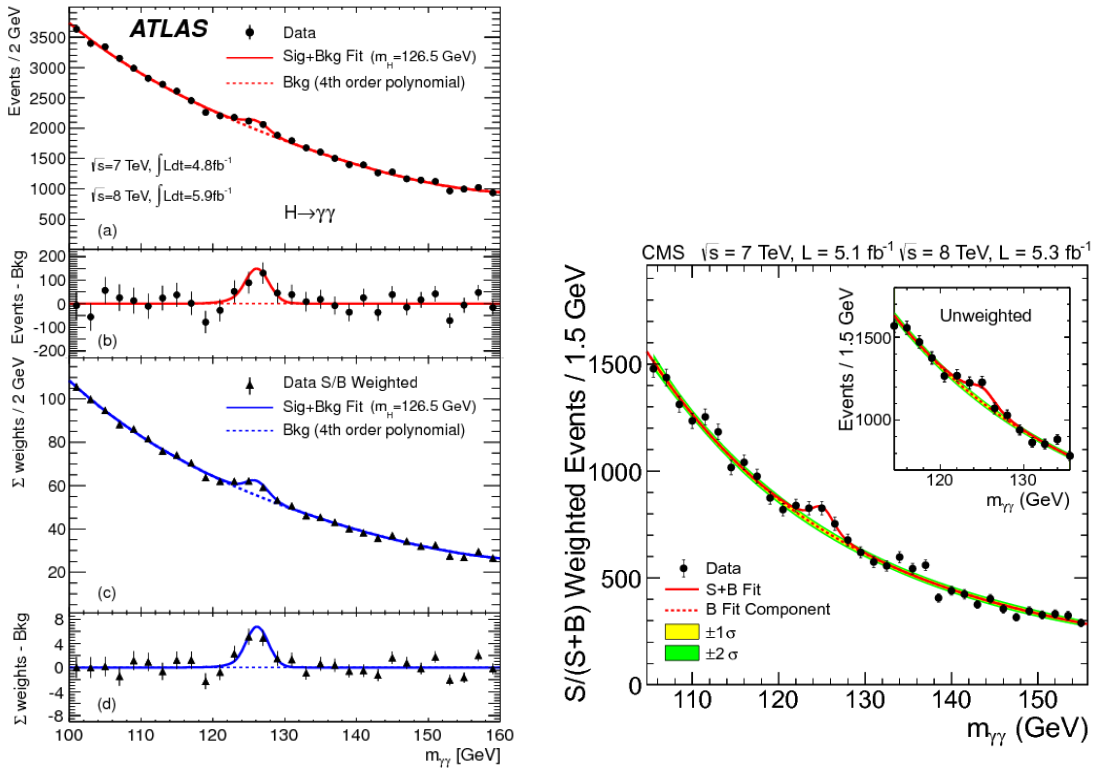


Figure 2.18: Distribution of the di-photon invariant mass obtained by the (left) ATLAS and (right) CMS experiments for the Higgs searches using ∼5 fb^{-1}of $\sqrt{s} = 7$ TeV pp collision data and ∼5 fb^{-1}at $\sqrt{s} = 8$ TeV. The lines represent the fits to the background plus signal components and to the background component only. For ATLAS, the corresponding distributions with the background subtrated are additionally exhibited in the bottom panels. Taken from [7] and [8] respectively.

able to trigger the signal process and to substantially reject background.

In what follows, the list of the channels searched for at the LHC by both the ATLAS [20] and CMS [21] experiments, is discussed.

$H \to \gamma\gamma$ looks for the Higgs resonant peak on the di-photon mass spectrum of events with two isolated photons. The background is mostly due to prompt di-photon production and the search is divided in several categories of jet multiplicity, to accommodate the different production modes according to their final state topology. Although the $H \to \gamma\gamma$ was expected to be visible only for a small Higgs mass range, the $\gamma\gamma$ signature was very appealing for being much distinct from the major multijet background produced by the LHC. In addition, since both the ATLAS and CMS detectors have good energy resolution for photons, this channel is useful to make the most precise measurements of the Higgs mass. Figure 2.18 contains the di-photon mass spectrum observed by ATLAS and CMS where the Higgs peak was seen for the first time. This particular decay is also interesting for it provides access to the Htt coupling via the dominant top loop.

$H \to ZZ^* \to 4\ell$ This channel is more successful in the charged leptonic decay of the Z bosons. If the hadronic mode is considered, the much larger multijet background spoils the signal-to-background ratio, making the signal observation impracticable. On the contrary, the four charged lepton final state produces a clear signal on the detectors and a much finer mass resolution. The Higgs signal was expected to appear as a narrow peak on top of a continuum falling background composed essentially of non-resonant ZZ production. As for the previous case, the search is explicitly made inclusively on the Higgs production mechanism. This channel contributed to the first observation of the Higgs by ATLAS and CMS in 2012.

$H \to WW^*$ This decay is dominant for the heavy Higgs scenario and, considering the leptonic decay of at least one of the W bosons, most of the jet background is ruled out. Therefore, it is one of the most sensitive Higgs search channels. The drawback is that having neutrinos in the final state does not allow to determine the Higgs invariant mass. Since neutrinos practically never interact with the detector, their four-momenta can not be measured. The Higgs signal is not observed as a clear resonant peak but rather a spread-out excess of events on the spectrum of the transverse mass, on top of the background dominated by WW continuum production. The search is made for all the Higgs main production mechanisms and was the third channel included in the first Higgs observation report.

$H \to \tau\tau$ Although having a branching fraction of only $\sim$6%, this decay provides the best chance of probing the Higgs coupling to charged leptons at the LHC. But since the τ decays immediately after production, either to lighter leptons through $\tau \to \ell\nu_\ell\nu_\tau$ or to hadrons, the identification of these particles is not straightforward. In particular, the hadronic mode produces a narrow jet that can be faked by quark-originated jets. The search is designed to fit the variety of final states of the τ decay mode topology and the Higgs production modes considered: ggF, VBF and VH.

VH $(H \to bb)$ As stated before, the search for the Higgs decay to $b-$quark pairs uses the associated production with a vector boson or a top pair. In the decay chain resulting in leptons, these are used to trigger the signal and reduce the overwhelming amount of jet background. But even using this technique, the signal-to-background ratio is too poor and the decay is still to be observed. Given the larger cross-section and the simpler final state, the VH production is a better option when compared to ttH. The analysis strongly relies on the jet flavour identification to discard the important $V+c$ or light jets background. This decay channel is of particular importance since given the magnitude of the rate, it can set serious constraints on the Higgs total width. Besides, if observed, it allows to directly access the Higgs coupling to quarks.

$H \to \mu\mu$ This channel offers a more clear way to probe the Higgs coupling to charged leptons but the branching fraction is only 0.02% for $m_h = 125$ GeV. The analysis looks for

a narrow signal peak in the di-muon invariant mass spectrum presenting a falling background corresponding essentially to the Drell-Yan Z/γ^* production.

ttH production This search targets the final state topologies associated with $H \to bb$, $H \to WW^*/ZZ^*/\tau\tau \to$leptons and $H \to \gamma\gamma$ decay modes. It is particularly interesting for probing directly the coupling of the Higgs to top quarks, since, although the $H \to \gamma\gamma$ and ggF contain the Htt vertex, this happens within a loop, and therefore has less sensitivity to the coupling. A clear disadvantage is the complexity of the final state, composed of a high multiplicity of particles where difficult combinatorial problems can arise in order to identify the Higgs decay products.

Besides these, other searches are carried out by the ATLAS and CMS experiments. One example is the $H \to Z\gamma$ decay and the Higgs decay to invisible particles. If observed, the latter would reveal evidence of physics beyond the SM (BSM) and specially, it could bring light to the dark matter question. Since dark matter candidates are in principle very weakly interacting and massive, they could provide an explanation for an observation of the invisible decay of the Higgs.

The searches described so far target the SM Higgs discovery and study. However, the LHC program for the Higgs experimental research field is far more extensive. Multiple Higgses models as the two Higgs Doublet Model (2HDM), that proposes a second Higgs field doublet resulting in five Higgs bosons, are under investigation by ATLAS and CMS. Under this context, the observed scalar would be the lighter of the multiple Higgs set. Since the 2HDM is necessary to generate the spontaneous symmetry breaking of Super Symmetric (SUSY) models, the observation of a second Higgs could establish a bridge with SUSY.

2.3.2 Results

The most precise experimental value of the Higgs mass combines the ATLAS and CMS results obtained with the complete $\sqrt{s}$ =7 and 8 TeV pp collisions datasets [22], with corresponding integrated luminosities of approximately 5 fb^{-1}and 20 fb^{-1}, comprehending what is called the LHC Run I data. It uses only the $H \to \gamma\gamma$ and $H \to ZZ$ search channels, as these provide the best mass resolution. Figure 2.19 shows the individual mass measurements per channel and experiment, and the combined measurement obtained from a simultaneous statistical analysis of all the data. The result has the central value of $m_H = 125.09$ GeV and the total uncertainty is of the order of 0.2%. The latter is still dominated by the statistical component, leaving room for improvement when the data from the LHC second run is included.

This result is used to obtain more precise theoretical predictions of the Higgs production cross-section and decay rates. The latter are confronted with their experimental measurements through the signal strength parameter μ, defined as the ratio of the observation to the SM prediction. μ is typically the outcome of a statistical analysis of data and prediction, where

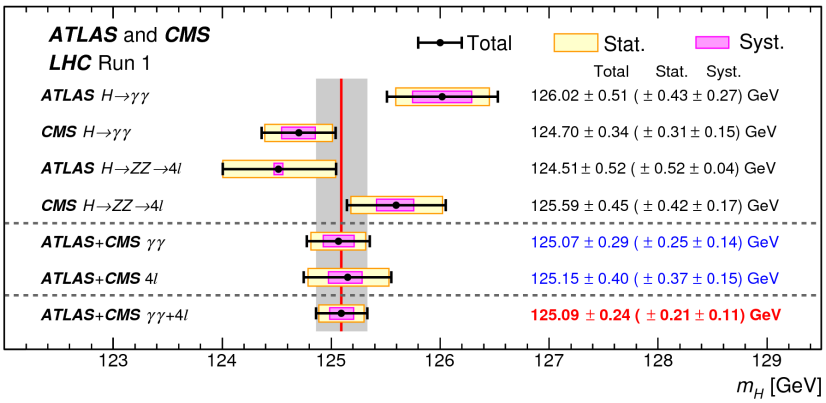


Figure 2.19: ATLAS and CMS combined Higgs mass measurement using the complete $\sqrt{s}$ =7 and 8 TeV pp collisions datasets provided by the LHC and the $H \to \gamma\gamma$ and $H \to ZZ$ channels. The individual channel results are also displayed. Taken from [22].

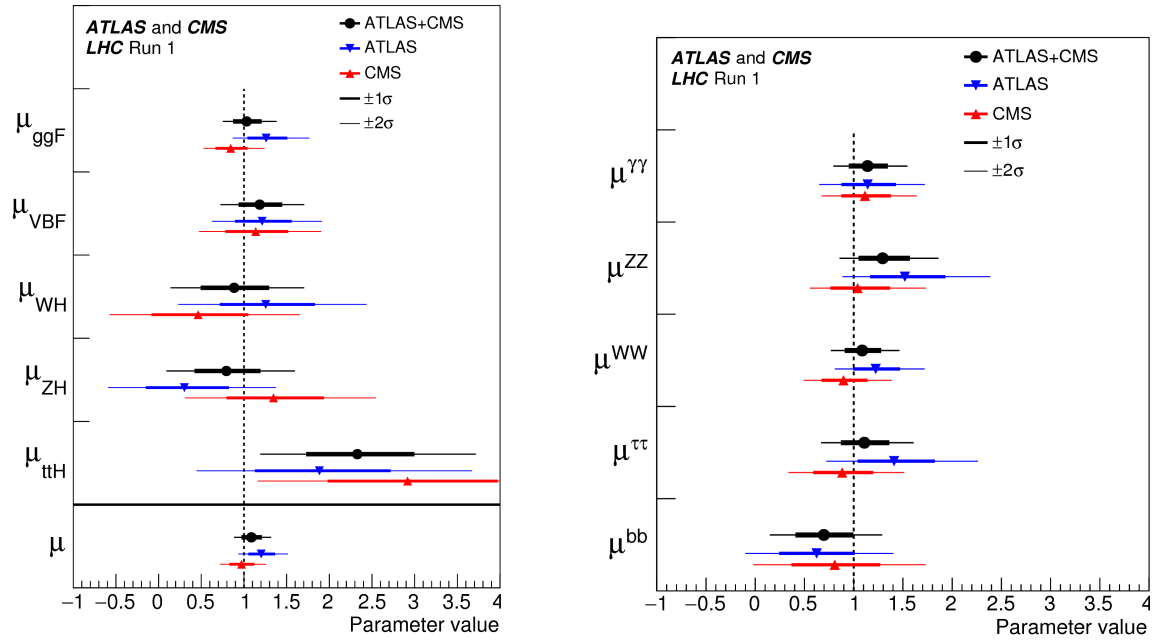
systematic uncertainties of both theoretical and experimental nature are taken into account.

As for the Higgs mass, the most precise measurements of the signal strength take advantage of the entire LHC Run I dataset, recorded and analysed by ATLAS and CMS. All the search channels described above are inputs to a combined measurement of the Higgs signal strength. The results are shown in Figures 2.20(a) and 2.20(b). The former displays the signal strength measurement of the production process, here $\mu = \sigma_{obs}/\sigma_{SM}$, for the individual experiments and their combination. The latter is based on the decay rates, where the signal strength is defined as $\mu = BR_{obs}/BR_{SM}$. With the exception of ttH production, all the experimental results are compatible with the SM prediction within a 1σ uncertainty, indicating that the new particle observed is indeed compatible with the SM Higgs.

Many of the signal strength values have large uncertainties and this is related to the fact that some of the analyses have poor sensitivity to the signal. In some cases, the sensitivity is so small that the observation was not yet accomplished. The signal observation is usually only claimed when the probability of observing the data under the background-only hypothesis is less than the gaussian probability at 5σ. This probability is usually designated significance and the 5σ convention is what is considered sufficient to securely rule out the background fluctuation faking the signal case.

Table 2.2 summarises the Run I measurements of significance for the various Higgs production mechanisms and decays, obtained from the combined analysis of the ATLAS and CMS results. The values are listed only for the processes that were not observed individually by the experiments and therefore ggF, $H \to \gamma\gamma/WW^*/ZZ^*$ are not included. Both the expected and observed significance are shown. The combined analysis resulted on the observation of $H \to \tau\tau$ and VBF. The former constitutes the first observation of the Higgs coupling to leptons. ttH provides evidence of the Higgs coupling to quarks while $H \to bb$ remains unobserved. There is also evidence of the VH associated production.

The same data was used to measure the strength of the Higgs couplings to the particles. A combined statistical analysis of all the search channels is performed to determine the coupling



(a) (b)

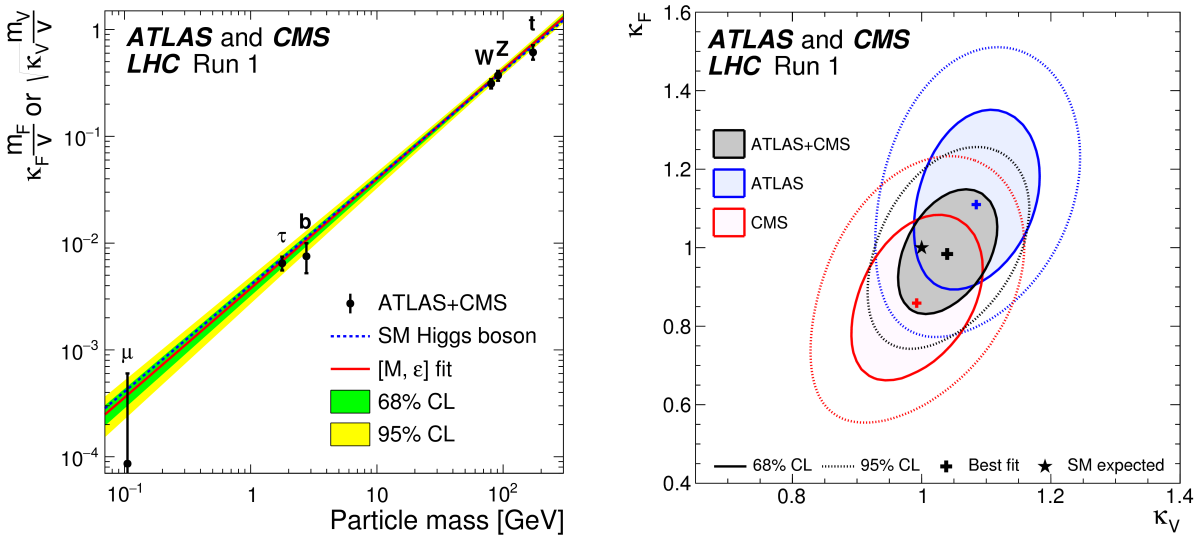
Figure 2.20: Signal strength of the Higgs (a) production mechanisms and (b) decay modes measured by ATLAS and CMS individually and their combination, obtained by analysing the Run I pp collisions at $\sqrt{s}$ =7 and 8 TeV complete dataset. Taken from [23].

modifiers κ, defined as the ratio between the observed and predicted Higgs coupling strength. Figure 2.21(a) exhibits the results obtained for the $\kappa-$factors. According to the SM, the coupling strength of the Higgs to fermions is proportional to m_f/v and to m_V^2/v for massive bosons. The graph displays $\kappa \times m/v$ as a function of the SM particles mass m. For bosons, $\sqrt{\kappa}$ is used instead of κ. The SM predicts that this representation should yield a straight line of slope $1/v$. The experimental values exhibit the same trend within uncertainties.

Figure 2.21(b) displays the Higgs-fermion coupling modifier κ_F as a function of the Higgs-vector boson coupling modifier κ_V. These values were obtained by imposing a single $\kappa-$factor modifying all fermionic couplings, and similarly for bosons, during the data fit. SM predicts $\kappa_F = \kappa_V = 1$ and the experimental results are shown for the ATLAS and CMS individual and

Production process	Observed significance (σ)	Expected significance (σ)
VBF	5.4	4.6
WH	2.4	2.7
ZH	2.3	2.9
VH	3.5	4.2
ttH	4.4	2.0
Decay mode		
$H \to b\bar{b}$	2.6	3.7
$H \to \tau\bar{\tau}$	5.5	5.0

Table 2.2: ATLAS and CMS combined significance of the Higgs production and decay modes. Adapted from [23].



(a) (b)

Figure 2.21: (a) ATLAS and CMS combined measurement of the Higgs coupling modifiers κ. The graph shows $\kappa \times m/v$ as a function of the SM particle mass m, where v is the vacuum expectation value and $\sqrt{\kappa}$ is used for bosons instead of κ. (b) Higgs-fermion coupling modifier κ_F as a function of the Higgs-vector boson coupling modifier κ_V and uncertainty contours. The results were obtained by analysing the Run I pp collisions at $\sqrt{s}$ =7 and 8 TeV complete dataset. Taken from [23].

combined analyses. The respective 68% and 95% confidence levels contours are drawn. Both the individual results and combination are compatible with the SM with 68% confidence level. The combination suggests a larger coupling strength for vector bosons than predicted.

The SM predicts a Higgs boson with spin-0 and even parity. The new particle observed was characterised in terms of these quantum numbers by ATLAS and CMS using the $H \to \gamma\gamma$, $H \to ZZ^*$ and $H \to WW^*$ analyses of the Run I data [24, 25]. The observation of a resonance with these decay channels already excludes the hypothesis of a spin-1 parent, and therefore only the spin-0 and 2 scenarios are tested. All the alternative hypothesis tested are ruled out with 99% confidence level when compared to the SM.

All the measurements made so far indicate that the new particle discovered by ATLAS and CMS in 2012 is compatible with the SM Higgs boson, but are insufficient to unequivocally confirm that or other hypothesis. The results presented and discussed here correspond to the LHC Run I dataset but the first data of $\sqrt{s}$ =13 TeV pp collisions were already analysed leading to similar conclusions. So, up to now, the LHC data was not able to rule out the SM identity of the Higgs.

The LHC Run II is expected to deliver $\sim$ 100 fb^{-1}of new data along with the possibility of making more precise measurements of the Higgs, reduce the statistical uncertainties of the measurements and increase the sensitivity of the non-observed search channels. The VH search with $H \to bb$ is one of the most awaited for its potential to probe the Higgs field. This thesis describes the $H \to b\bar{b}$ search in the associated production channel with a W boson using the

20.3 fb^{-1}of $\sqrt{s} = 8$ TeV pp collisions data collected with the the ATLAS detector.

Chapter 3

The ATLAS Experiment at the LHC

ATLAS is one of the four experiments operating at the LHC accelerator at CERN. It is a general-purpose detector designed and constructed to explore a broad range of subjects at the edge of High Energy Particle Physics, from the Higgs boson to beyond Standard Model searches.

In this Chapter, a brief introduction to the LHC is presented in Section 3.1, together with a description of the proton beam structure and proton-proton collisions. Afterwards, in Section 3.2, the ATLAS detector is described in detail in its components, with highlights to the technology, geometry and main characteristics. The ATLAS trigger system and the data acquisition and preparation chains are also presented.

3.1 The Large Hadron Collider

The Large Hadron Collider (LHC) [26] is a 26.7 km-long circular hadron accelerator and collider located at CERN that operates at the highest energies ever. It is the last unit in a chain that accelerates a beam of particles at successively higher energies, installed approximately 100 m underground. The LHC was designed to provide proton-proton collisions at a maximum nominal centre-of-mass energy of $\sqrt{s}$ =14 TeV and peak luminosity of 10^{34} cm^{-2}s^{-1}. Besides these, the accelerator can also provide lead nuclei collisions or proton-lead collisions. Four main experiments operate in the four collision points of the LHC, as shown in Figure 3.1. ATLAS and CMS are general purpose experiments, while ALICE and LHCb are dedicated to lead collision studies and B-physics, respectively.

The LHC is a two-ring accelerator, with each beam of particles travelling in opposite directions. The two beams are injected through two points close to the ATLAS experiment, one point for the clockwise travelling beam, the other for the anti-clockwise, both equipped with beam collimators. A specific point in the LHC allows the beams to be independently expelled from the accelerator, using deflecting magnets. Two regions perform the beam cleaning, where particles deviated from the beam are scattered, collimating the beam. The accelerating part consists of radio-frequency (RF) cavities.

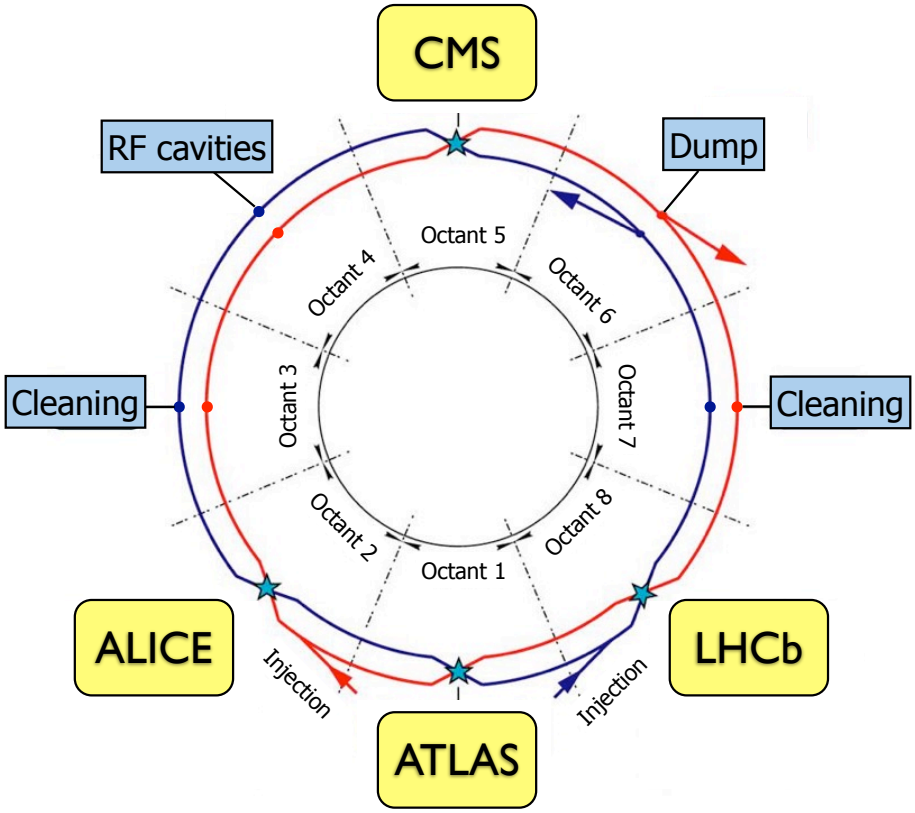


Figure 3.1: Layout of the LHC showing its main experiments and installations for its elementary functionality. Adapted from [26].

The LHC integrates the CERN accelerator complex, depicted in Figure 3.2. It is the last step in an accelerator chain that begins by injecting protons stripped out of hydrogen atoms into the linear accelerator Linac2. This step accelerates the protons up to 50 MeV providing the beam for the Proton Synchrotron Booster (PSB), that accelerates protons up to 1.4 GeV. The beam follows through the Proton Synchrotron (PS) and Super Proton Synchrotron (SPS) that increase the energy of the particles to 25 GeV and 450 GeV, respectively. The SPS injects the protons into the LHC through the two injection points in the opposite directions. At last, the LHC accelerates the two beams up to 7 TeV and may provide proton collisions up to $\sqrt{s}$ =14 TeV.

This accelerating machine has 8 arcs and 8 straight sections. The straight sections lodge the experiments or access points to the LHC tunnel, while the arcs consist of superconductor magnets based on Nb/Ti cables cooled down to 2 K by superfluid helium. The magnetic field provided has an intensity above 8 T, and guide the beam in its circular path. Two vacuum tubes where the beams travel are placed inside the magnets, creating a magnetic flux in two opposite directions for the two counter-circulating beams. Acceleration of particles is ensured by RF cavities. Here, an RF 400.05 MHz power generator supplies an electrical field that is made to resonate by the particular shape and dimension of the cavity. Particles can be either accelerated or decelerated by the oscillating field, depending on their timing. This phenomenon generates the bunch structure of the beam, each bunch spaced by 25 ns following the pace of the field

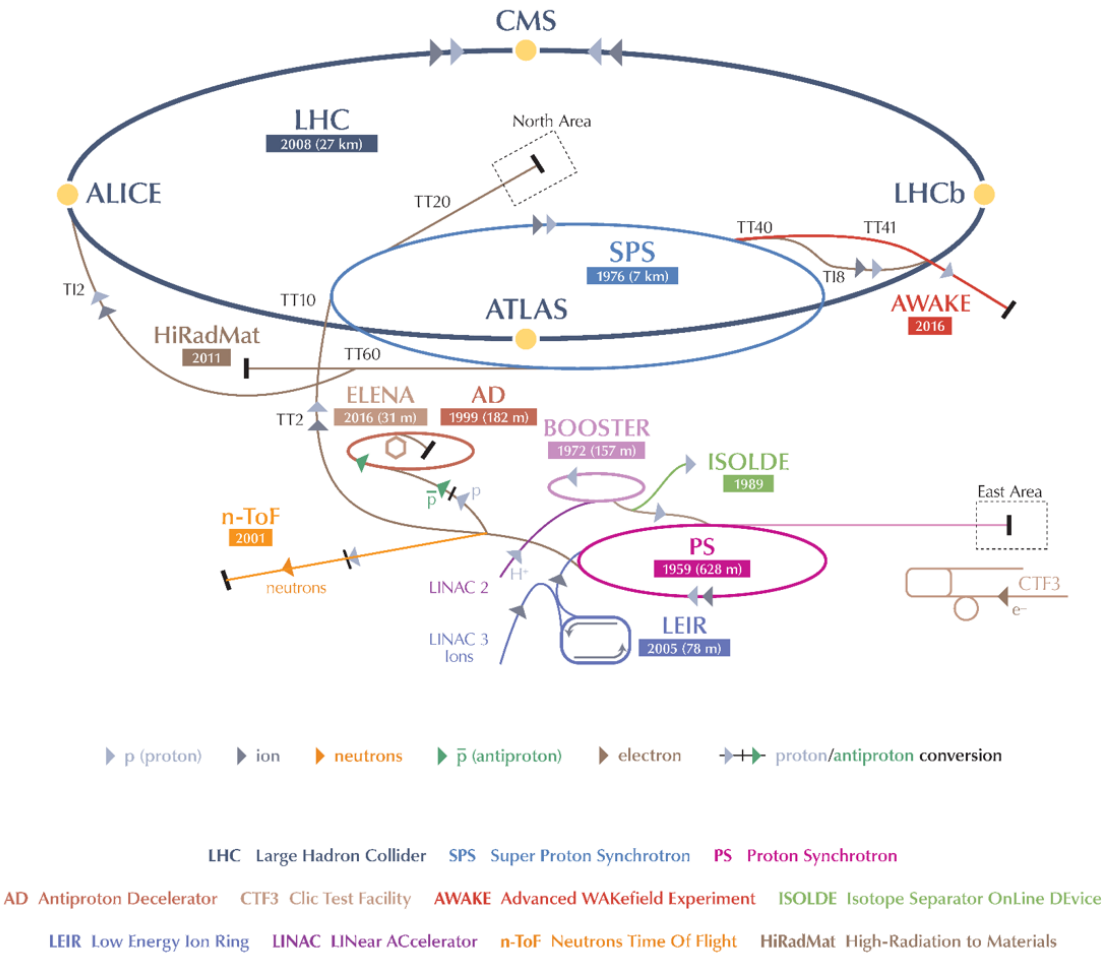


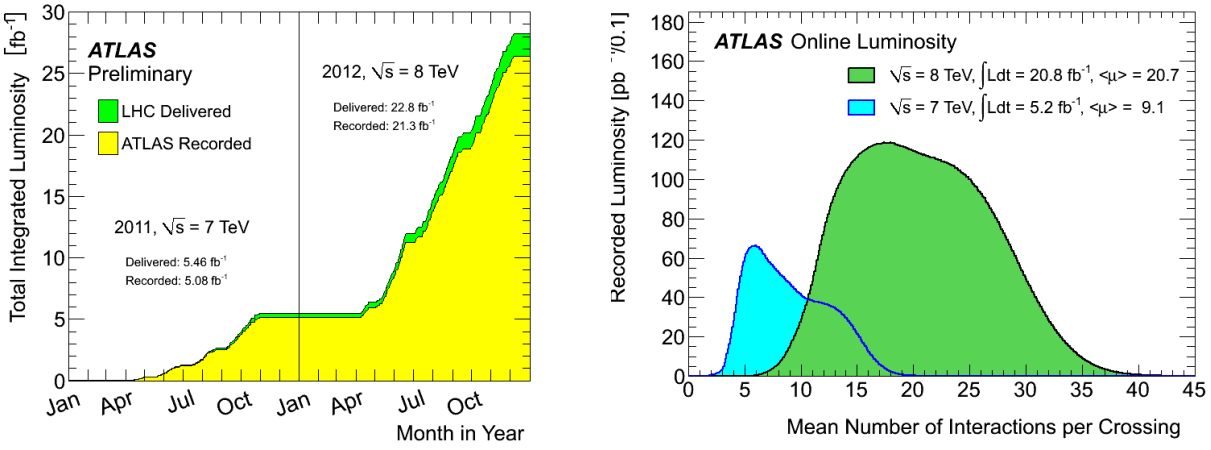
Figure 3.2: Diagram of the CERN accelerator complex. Taken from [27].

oscillation. The RF cavity system is thus responsible for the beam capture and acceleration.

The LHC purpose is to survey particle physics at unprecedently high energy, testing the Standard Model with precision and searching for new physics. These searches generally involve rare processes of small production cross-section. As shows Eq. 2.27 in Section 2.2, a way to maximise the number of such rare events is to increase the instantaneous luminosity $\mathscr{L}$ of the accelerator. The latter is given by

$$\mathscr{L} = \frac{N_b^2 n_b f_{rev} \gamma}{4\pi\varepsilon_n \beta^*} F \tag{3.1}$$

where N_b is the number of particles per bunch, n_b the number of bunches per beam, f_{rev} the frequency of revolution, γ the relativistic factor and ε_n the normalised transverse emittance of the beam. β^* is the beta function at the collision point that is related to the transverse size of the beam. F is the geometrical reduction factor of the luminosity due to the crossing angle at the interaction point (IP). To maximise the event rate, the parameters in the numerator must be as large as possible and this guided the design principles of the LHC. Then, during operation, several factors influence the stability of the beam by causing luminosity losses. The main ones are losses from collisions, particle scattering in residual gas in the vacuum tubes and intra-beam



(a) (b)

Figure 3.3: (a) Integrated luminosity as a function of time in 2011 and 2012 for pp collisions at $\sqrt{s} = 7$ TeV and $\sqrt{s} = 8$ TeV centre-of-mass energy as delivered by the LHC and recorded by the ATLAS detector. (b) Luminosity recorded by ATLAS as a function of the number of interactions per bunch crossing at the LHC for the same period. Taken from [28].

scattering. Nevertheless, the LHC is able to guarantee beams with lifetime of approximately 15 h, ensuring the stability needed during data taking.

At the LHC, the beams have 2808 bunches, and the bunch crossing happens every 25 ns. In a typical collision, several pp collisions take place, and this effect is known as collision pile-up. Most of the pp interactions in a bunch-crossing are soft and result in low energy jets, while interesting hard-scatters are rare. Being able to distinguish the hard-scatter process in the massive pile-up environment is of crucial importance to physics studies, and motivated many design guidelines of the LHC detectors. Pile-up can even affect a different bunch-crossing event. Some features of the detector, such as having an electric output signal width larger than 25 ns, allow signals resultant from neighbouring bunch crossings to overlay. This effect is known as out-of-time pile-up, whereas the former, and most common one, is oppositely referred to as in-time pile-up.

During the Run I, that took place in 2011 and 2012, the LHC collided protons at the centre-of-mass energy of $\sqrt{s}$ =7 and 8 TeV, delivering an integrated luminosity of 5.46 fb^{-1}and 22.8 fb^{-1}, respectively [28]. As Figure 3.3(a) shows, the ATLAS detector recorded about 93% of the data. The pile-up conditions of this data set are presented in Figure 3.3(b). The average number of collisions per bunch crossing $< \mu >$ for the $\sqrt{s}$ =7 TeV pp collisions was 9.1. This value raised to 20.7 for $\sqrt{s}$ =8 TeV.

After a shut-down period, the LHC initiated the Run II operation in 2015, with increased pp collision energy of 13 TeV and $< \mu >$ of 22.9 [29]. At the end of 2016, the total integrated luminosity delivered by the LHC was 43.1 fb^{-1}, from which ATLAS recorded 39.9 fb^{-1}.

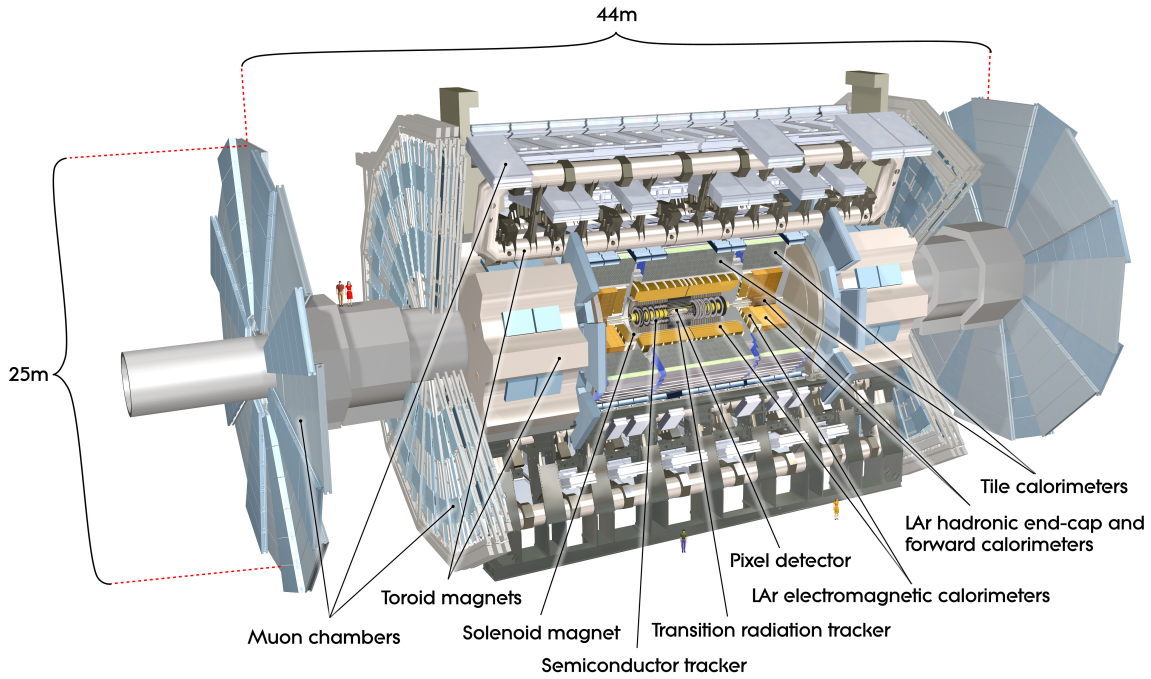


Figure 3.4: Layout of the ATLAS detector. Taken from [30].

3.2 The ATLAS Detector

The ATLAS (A Toroidal LHC Apparatus) detector [30] is one of the four detectors of the products of the LHC hadron collisions. It is a cylindrical detector made of different layers around the collision interaction point, designed to be hermetic and provide the best spherical coverage needed for the fiducial measurements.

Figure 3.4 shows the scheme of the ATLAS detector unveiling its three sub-detector systems. The innermost sub-detector, the Inner Detector (ID), composed of a pixel detector, a semiconductor tracker and a transition radiation tracker, surrounded by a solenoid magnet, is responsible for the tracking and momentum measurement of electrically charged particles. The ID is followed by the Electromagnetic (EM) and Hadronic calorimeters, that provide energy measurements of electrons, photons and hadrons. ATLAS also comprises a Muon Spectrometer (MS), consisting of muon chambers and toroid magnets, dedicated to muon identification and momentum measurements. Finally, an online trigger system executes a prompt real-time event selection, rejecting uninteresting events, reducing the event rate from around 40 MHz to about 400 Hz, adequate for the data acquisition system and data storage capacity.

The ATLAS collaboration uses the right-handed coordinate system convention, with its origin in the centre of the detector and nominal interaction point, the x-axis pointing to the LHC centre, the y-axis directed upwards and the z-axis tangent to the beam line. It is often useful to use spherical coordinates (r, ϕ, θ), where the azimuthal angle ϕ is defined in the $x-y$ plane, transverse to the z-axis, while the polar angle θ is measured from the z-axis.

The pseudorapidity η, written as a function of the polar angle by $\eta = -\ln \tan(\theta/2)$, and

Detector component	Required resolution	η coverage	
		Measurement	Trigger
Tracking	$\sigma_{p_T}/p_T = 0.05\% \, p_T \oplus 1\%$	$\pm$ 2.5	
EM calorimeters	$\sigma_E/E = 10\%/\sqrt{E} \oplus 0.7\%$	$\pm$ 3.2	$\pm$ 2.5
Hadronic calorimeters			
barrel and end-cap	$\sigma_E/E = 50\%/\sqrt{E} \oplus 3\%$	$\pm$ 3.2	$\pm$ 3.2
forward	$\sigma_E/E = 100\%/\sqrt{E} \oplus 10\%$	$3.1 < \lvert\eta\rvert < 4.9$	$3.1 < \lvert\eta\rvert < 4.9$
Muon spectrometer	$\sigma_{p_T}/p_T = 10\%$ at $p_T = 1$ TeV	$\pm$ 2.7	$\pm$ 2.4

Table 3.1: Performance goals of the ATLAS detector and pseudorapidity covarage for particle measurement and trigger system. Energy E and transverse momentum p_{T} are given in units of GeV. Taken from [30].

the rapidity y, that is function of the particle energy E and longitudinal momentum p_L

$$y = \frac{1}{2} \ln \left(\frac{\mathrm{E} + \mathrm{p_L}}{\mathrm{E} - \mathrm{p_L}} \right) \tag{3.2}$$

are commonly used in accelerator particle physics because the rate of the collision products is approximately constant over these quantities. Besides, it can be proven that differences in pseudorapidity and rapidity are Lorentz invariant. In addition, the angular distance ΔR between two points is defined as $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$.

The multiple systems of the ATLAS detector ensure a high tracking resolution, efficiency in pattern identification and energy and momentum resolution requested for vertex finding, jet reconstruction, b-tagging and particle identification. ATLAS was designed to achieve the performance goals and the pseudorapidity coverage listed in Table 3.1, using radiation-hard materials and technology, compatible with the radiation-loaded environment where it operates. The $\eta < 2.5$ region is dedicated to precision measurements, achievable due to the high-resolution ID. The good resolution of the electromagnetic calorimeter compensates the coarser resolution of the hadronic calorimeter, assuring high-quality jet measurements. In the following, these sub-systems of ATLAS are presented in more detail.

3.2.1 Inner Detector

The innermost detector system of ATLAS, the Inner Detector (ID), was specifically designed and built to provide excellent momentum resolution of charged particles, allowing to determine the position of the primary vertice, that identify the collision points, and possible secondary vertices associated with particle decays.

Its cylindrical overall envelope is about 7 m-long and has a radius of nearly 2.3 m. The ID tracks electrically charged particles with momentum above 0.5 GeV within the $|\eta| < 2.5$ range, with particles leaving a typical number of 36 hits per track in the detector. Among the particularities of its design and construction, the most relevant ones were the high precision alignment criteria during installation, and resistance against the extreme radiation environment encountered at the LHC collision point. Being the ATLAS detector first layer, the ID is the most

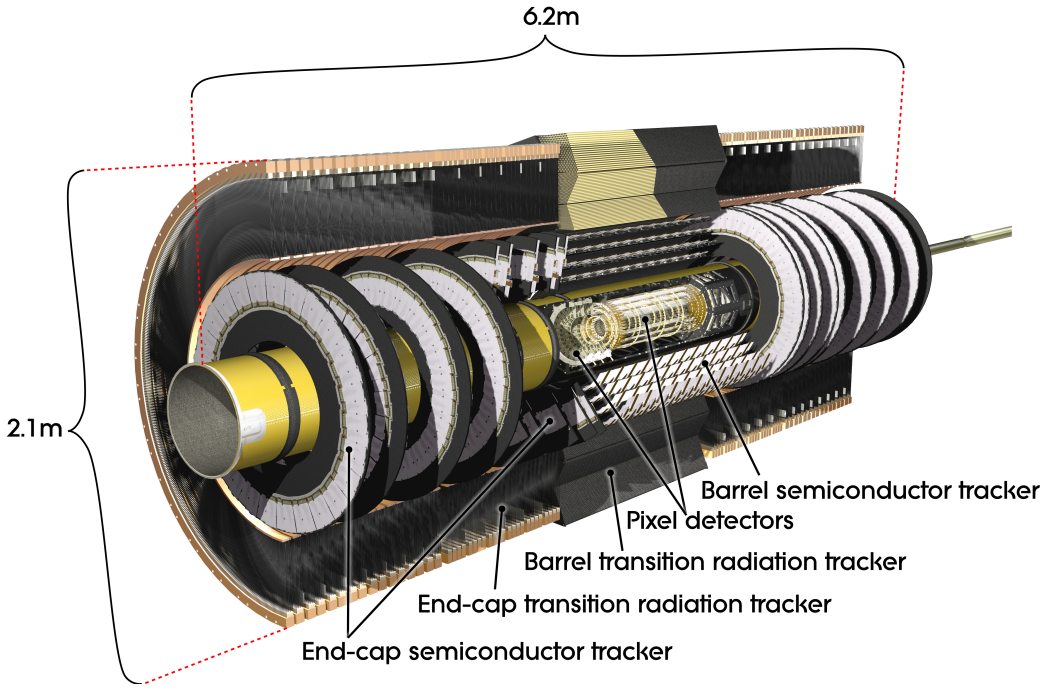


Figure 3.5: Sketch of the ATLAS inner detector. Taken from [30].

exposed to radiation, and still, it has to offer performance stability over time. During the LHC long shut down periods, many components of the ID are replaced to provide the detector with the full recovery needed for the next period of data taking.

When particles hit the ID, they are detected through ionisation of the sensitive material arranged in the components listed in Table 3.2. The ID starts with a Pixel detector made of silicon pixels layers, followed by the SemiConductor Tracker SCT composed of silicon microstrips layers, and the (TRT), made of layers of gaseous straw tubes interleaved with transition radiation material. Figures 3.5 and 3.6 present a graphical view of the ATLAS inner detector exhibiting these elements.

The readout of the ID must be synchronous with the LHC bunch crossing, and for that, a 40.08 MHz clock signal time-stamps the generated signals. The signal in the front-end electronics is then stored in buffers for the 2.5 μs compatible with the trigger system latency. Following a first trigger decision, in which the ID information does not take part, the buffer content is transferred to a readout driver (ROD) out of the detector.

Pixel

The pixel detector is located only 5 cm away from the beam pipe. It comprehends a centred barrel with 3 cylindrical layers and two end-cap regions, with 3 disk layers each, perpendicular to the z-axis. It has in total 1744 pixel sensor modules, arranged as shown in Table 3.2, with 46080 readout pixels in each sensor module, resulting in over 80 million readout channels. The sensor unit consists of oxygenated $n-$type wafers with readout pixels on the n^+-implanted side of the detector. The module comprises the front-end electronics chip and a flexible Printable

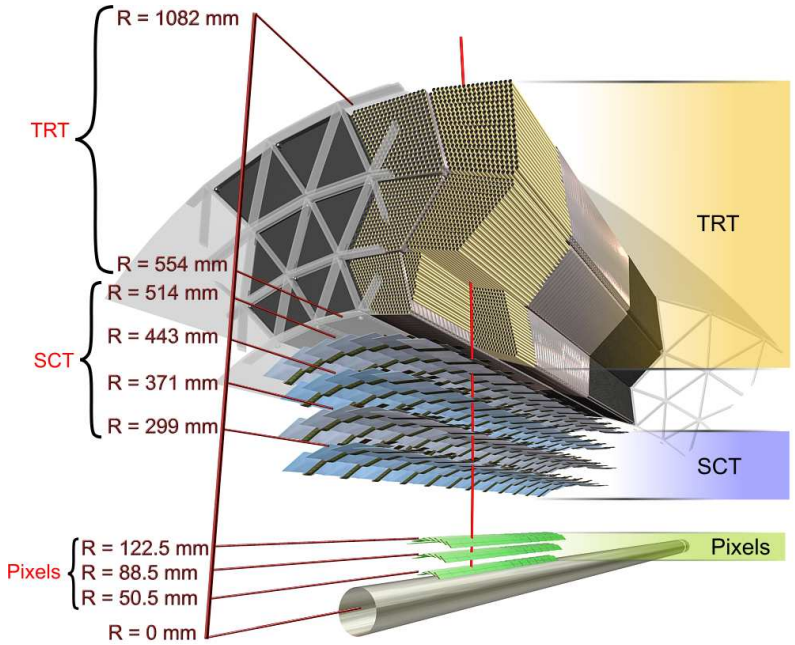


Figure 3.6: Drawing of the ATLAS inner detector sensors. Taken from [30].

Circuit Board (PCB) supporting the module control chip.

The layers of the detector are segmented in rectangles with a minimum size of $50 \times 400\ \mu\text{m}^2$ ($\Delta R \times \Delta z/\Delta\phi$) in the barrel/end-cap. This segmentation results in an intrinsic accuracy of 10 μm in $R \times z$ ($R \times \phi$) and 115μm in ϕ (z) in the barrel (end-cap), conferring to the ID the high granularity needed for high-performance tracking. A charged particle generally leads to three hits in the pixel layers, allowing to reconstruct a 3-hits segment of the particle track that then seed the full track reconstruction algorithms.

SCT

The SCT detector, made of 80 μm pitch silicon microstrips, is arranged in 4 cylindrical barrel layers and 2 end-caps formed by 9 disk layers each. The unit sensor is a single-sided p-in-n silicon-based detector, 285 ± 15 μm thick, with coupled readout bands. The barrel has 6 cm-long rectangular sensors daisy-chained in cylindrical layers, while the end-caps employ trapezoidal sensors oriented radially. A barrel module consists of 4 sensors, 2 at each side and an electrically conductive base-board providing HV supply. In total, the SCT has 15912 sensors and 6.3 million readout channels offering a combined spatial resolution of 16 μm in the $R-\phi$ space.

TRT

Finally, the TRT gives coverage to the $|\eta| < 2.0$ region with 4 mm diameter polyimide drift tubes filled with a gas mixture of 70% Xe, 27% CO_2 and 3% O_2 as the basic detector elements. The straw walls are the detector cathodes, and the anodes are 31 μm diameter tungsten wires directly connected to the front-end electronics and readout at each side of the straw. As in the

Item			Radial extension (mm)	Length (mm)	$\lvert\eta\rvert$ coverage
Overall ID envelope			$45.5 < R < 1150$	$\lvert z\rvert < 3512$	< 2.5
Beam-pipe			$29 < R < 36$		
Pixel	Overall envelope		$45.5 < R < 242$	$\lvert z\rvert < 3092$	< 2.5
3 cylindrical layers	Sensitive barrel		$50.5 < R < 122.5$	$\lvert z\rvert < 400.5$	< 1.7
$2\times$ 3 disks	Sensitive end-cap		$88.8 < R < 149.6$	$495 < \lvert z\rvert < 650$	$\in [1.7, 2.5]$
SCT	Overall envelope	barrel	$255 < R < 249$	$\lvert z\rvert < 805$	< 1.4
		end-cap	$251 < R < 610$	$810 < \lvert z\rvert < 2797$	$\in [1.4, 2.5]$
4 cylindrical layers	Sensitive barrel		$299 < R < 514$	$\lvert z\rvert < 749$	
$2\times$ 9 disks	Sensitive end-cap		$275 < R < 560$	$839 < \lvert z\rvert < 2735$	
TRT	Overall envelope	barrel	$544 < R < 1082$	$\lvert z\rvert < 780$	< 0.7
		end-cap	$617 < R < 1106$	$827 < \lvert z\rvert < 2744$	$\in [0.7, 2.0]$
73 straw planes	Sensitive barrel		$563 < R < 1066$	$\lvert z\rvert < 712$	
160 straw planes	Sensitive end-cap		$644 < R < 1044$	$848 < \lvert z\rvert < 2710$	

Table 3.2: Main components and geometrical characteristics of the ATLAS Inner Detector. Taken from [30].

pixel and SCT detector cases, this sub-system is divided into a barrel and two end-cap regions. The barrel layers 144 cm-long straws parallel to the beam direction interleaved with fibres. Since the straws are aligned with the beam line, the TRT barrel does not provide information about the z position of the track hit. Each end-cap employs planes of 37 cm-long straws oriented radially, and interleaved with foils. In total, the TRT comprehends 351000 readout channels.

Although this is not the most precise system of the ID, it contributes significantly to the track momentum measurement due to its radial length, with tracks leaving approximately 36 hits in the TRT. Other important feature of the TRT is to contribute to the identification of electrons through the detection of transition radiation photons in the Xe-mixture, produced by electrons in the material interleaved with the straws. This increases substantially the signal of electron hits when compared to hits of particles with larger mass.

Solenoid Magnet

The inner detector is immersed in a solenoid magnet, generating a longitudinal magnetic field approximately uniform and with an intensity of 2 T. The trajectory of charged particles bend under the magnetic field, and the particle momentum and its electrical charge is measured from the curvature. Table 3.3 lists the key features of the solenoid. The magnet, shown in Figure 3.7, is a superconductor, with 5.8 m length and 10 cm thick. One of the primary demands of the design was to minimise the amount of material in front of the subsequent detector layers of ATLAS in the electromagnetic calorimeter. The plot in Figure 3.8 shows the R and z dependence of the magnetic field radial and longitudinal components. The z-projection of the field is approximately constant for $|z| < 2$ m at 2 T, with little dependence on the radius, falling to 0.5 T closer to the solenoid aperture. The radial projection of the field is null for almost all volume, except closer to the solenoid edges, where it can reach an intensity of 0.7 T.

Size	Inner diameter (m)	2.46
	Outer diameter (m)	2.56
	Length (m)	5.8
	Number of coils	1
Mass	Conductor (t)	3.8
	Cold mass (t)	5.4
	Total assembly (t)	5.7
Coils	Turns per coil	1154
	Nominal current (kA)	7.73
	Peak field in the windings (T)	2.6
Conductor	Overall size (mm^2)	30×4.25
	Ratio Al:Cu:NbTi	15.6:0.9:1
	Temperature margin (K)	2.7

Table 3.3: Main specifications of the ATLAS central solenoid magnet. Taken from [30].

Figure 3.7: Photograph of the ATLAS solenoid magnet. Taken from [30]

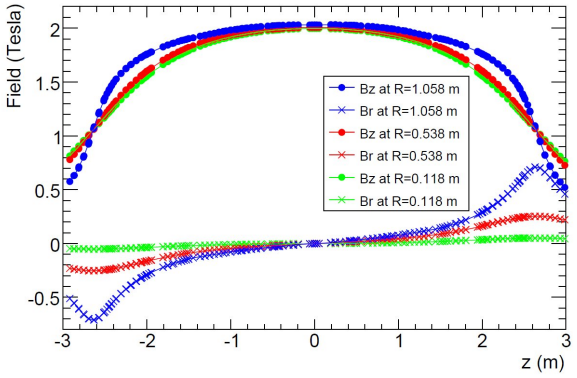


Figure 3.8: z and R dependence of the radial and longitudinal projections of the solenoid magnetic field at the ATLAS inner detector. Taken from [30].

This non-uniformity is taken into account by the reconstruction algorithms used in ATLAS and has little impact on the performance of the inner detector.

3.2.2 Electromagnetic and Hadronic Calorimeters

The ATLAS calorimetry system sketched at Figure 3.9 and detailed at Table 3.4 is composed of an electromagnetic (EM) calorimeter, layered immediately next to the solenoid magnet, followed by a hadronic calorimeter. These systems are devoted to the measurement of the energy deposited by particles and its topology. The missing transverse energy is also determined with the calorimeter, making of hermeticity a necessary design condition. For this reason, the ATLAS calorimeter occupies the $|\eta| < 4.9$ range.

The ATLAS calorimeters are sampling detectors, meaning that layers of sensitive material are interleaved with dense absorber material. Particles crossing the calorimeter lose energy in the absorber by interacting with their nuclei and form particle showers/cascades. It is the

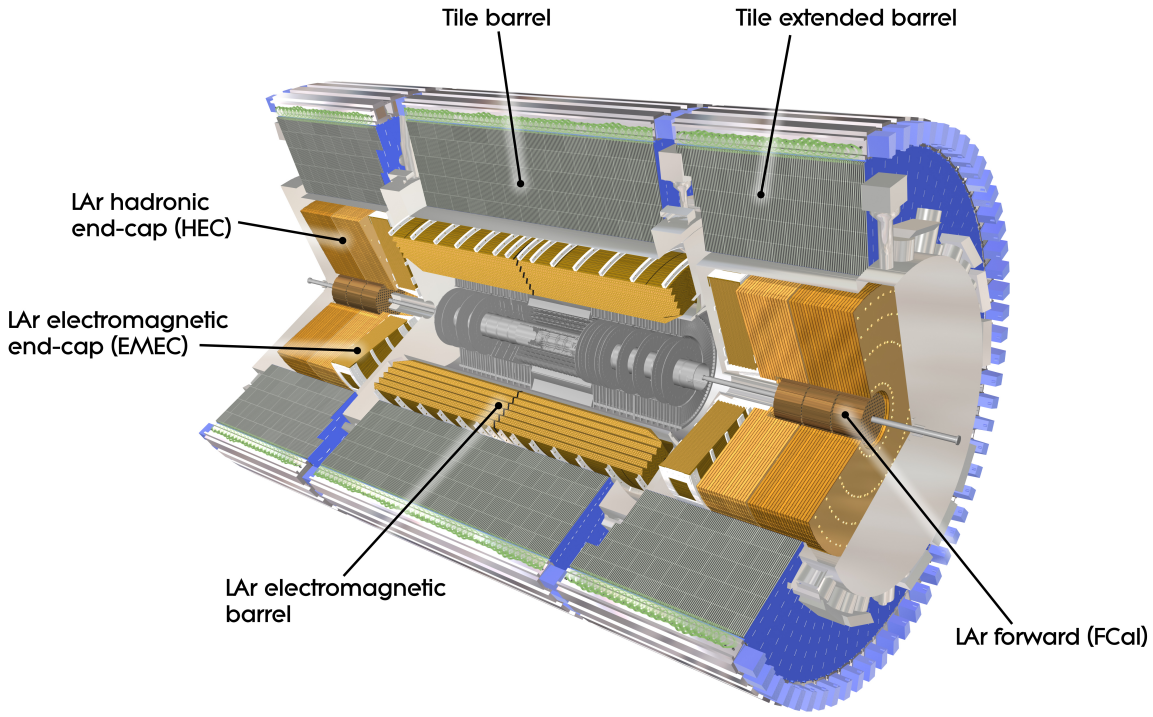


Figure 3.9: Sketch of the ATLAS calorimeter system showing the LAr electromagnetic calorimeter, and the hadronic calorimeter composed of the LAr hadronic end-cap, the forward calorimeter and the Tile barrel and extended barrels. Taken from [30].

sensitive medium that generates the signal proportional to that energy loss.

This detector system comprises an electromagnetic calorimeter barrel (EMB) and two end-caps (EMEC), two hadronic end-caps (HEC), two forward calorimeters (FCal) and a hadronic tile calorimeter (TileCal) with one barrel and two extended barrels. The EM, HEC and FCal systems employ liquid argon as the active medium, whereas the TileCal have scintillator tiles combined with steel.

The ATLAS calorimeters are non-compensating, i.e. their electromagnetic e and hadronic h energy responses are different, $< h/e > \neq 1$. This is not the ideal design but their effect, mainly in the energy measurement of jets and hadronically decaying tau-leptons, is corrected posteriorly at the energy calibration phase as will be discussed in Section 4.4 for the particular case of jets.

Electromagnetic Calorimeter

The Liquid Argon EM calorimeter have a 3-layer barrel spreading over $|\eta| < 1.475$ and two end-caps, of 2 wheels each, located at $1.375 < |\eta| < 3.2$. It has an accordion-like geometry enabling uniform azimuthal coverage without gaps. LAr is an ionisation detector, consisting of an accordion-shaped lead absorber and electrode plates interleaved with liquid argon as sketches Figure 3.10. In the barrel, the accordion waves are positioned axially comprising a total of 2048 absorbers while the two EMECs have the waves parallel to the radial direction.

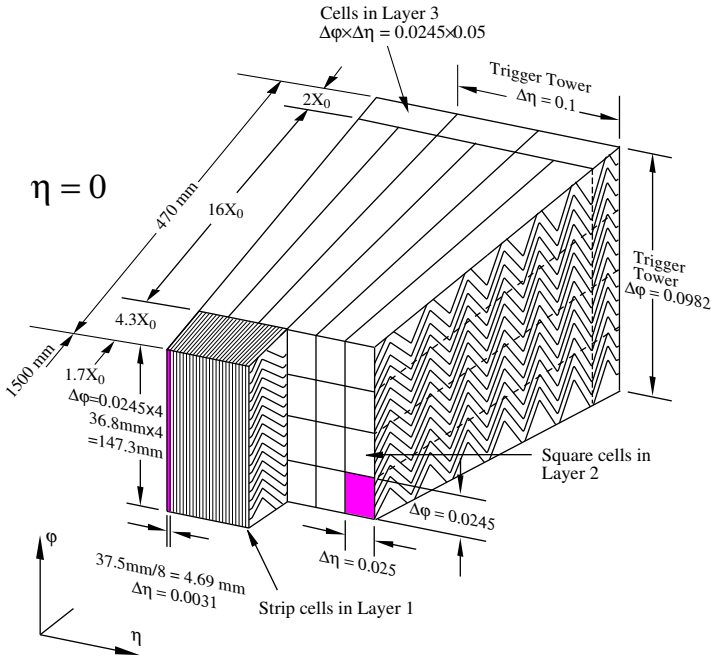


Figure 3.10: Detail view of the accordion geometry of the ATLAS Electromagnetic calorimeter. Taken from [30].

		Typical granularity $\Delta\eta \times \Delta\Phi$	$\lvert\eta\rvert$ coverage	Number of readout channels
EM Calorimeter	**Barrel**	0.025×0.025	< 1.475	109568
	End-caps	0.025×0.1	$\in [1.375, 3.2]$	63744
LAr Hadronic	**End-caps**	0.1×0.1 to 0.2×0.2	$\in [1.5, 3.2]$	5632
LAr Forward (FCal)		1×1 to 5×5(*)	$\in [3.1, 4.9]$	3524
Tile Calorimeter	**Barrel**	0.1×0.1	< 1.0	5760
	Extended barrels	0.1×0.1	$\in [0.8, 1.7]$	4092

Table 3.4: Main parameters of the ATLAS calorimeter system. Taken from [30]. (*) The granularity of the FCal refers to the $\Delta y \times \Delta x$ space and is given in cm^2.

In total, the EM calorimeter is more than 22 radiation lengths (X_0)[1] thick in the barrel and more than 24 in the end-caps. The thickness of the EM calorimeter was determined by its purpose of containing the electromagnetic shower.

The typical granularity of the detector in the $\Delta\eta \times \Delta\phi$ space ranges from 0.025×0.025 in the barrel to 0.025×0.1 in the end-caps, as summarised in Table 3.4. In its innermost region, the granularity is finer, lengthening to match the inner detector and to suit the precision measurements of electrons, photons and jets. The rest of the detector has sufficient granularity for the resolution requirements of the electromagnetic and hadronic cascades measurement.

[1]The radiation length X_0 is the mean distance over which the electron energy is reduced by a factor of $1/e$ through radiation loss, and $7/9$ of the mean free path before pair production by high-energy photons.

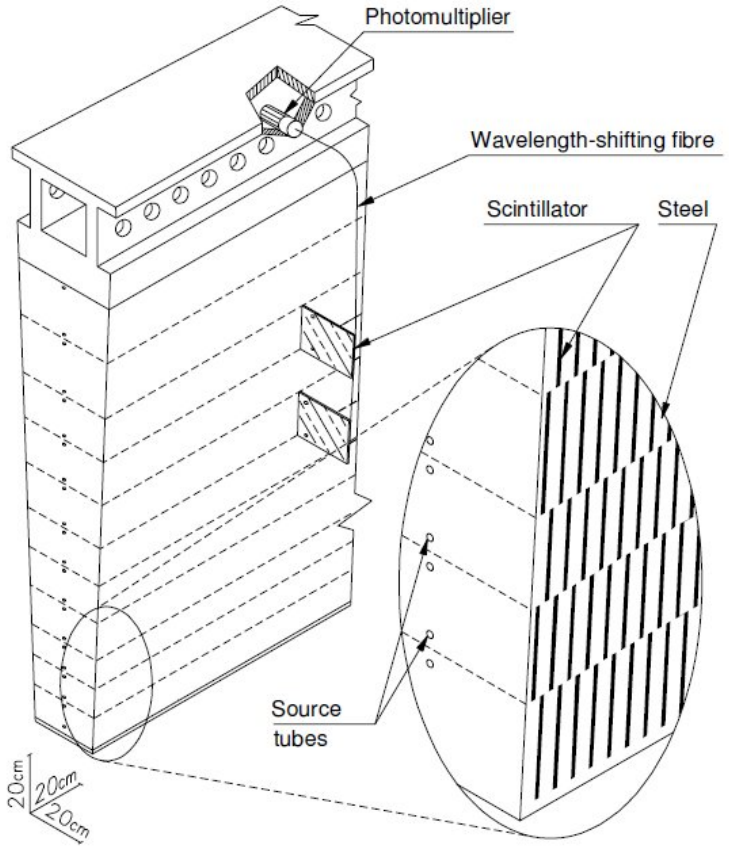


Figure 3.11: Scheme of the optical readout of an ATLAS tile calorimeter module. Taken from [30].

End-Cap and Forward Hadronic Calorimeters

The LAr Hadronic End-Cap (HEC) and Forward Calorimeter (FCal) cover the forward region of $1.5 < |\eta| < 3.2$ and $3.1 < |\eta| < 4.9$, respectively. These are concentric wheels in which the FCal occupies the inner radius area followed by the HEC. Both systems use liquid argon as the active material and HEC (FCal) employs copper (tungsten) absorber. These systems are purely hadronic calorimeters, with the exception of the FCal first layer, closest to the IP, that is also used for electromagnetic shower measurements. The typical granularity of these detectors decreases with pseudorapidity, ranging from $\Delta\eta \times \Delta\phi = 0.1 \times 0.1$ in the HEC first layers to $\Delta y \times \Delta x = 5 \times 5$ cm^2 in the FCal largest η layers, as shown in Table 3.4.

Tile Calorimeter

The Tile Calorimeter (TileCal) is the outermost system of the ATLAS calorimeter central region. The design met the specific intent of performing jet measurements and to fully contain the hadronic cascade. For that it has an overall thickness of 7.4λ[2] in the direction perpendicular to the beam line, and covers $|\eta| < 1.7$ with one barrel unit and two extended barrels.

[2]The nuclear interaction length λ is the distance required to reduce the number of high-energy particles by a factor of $1/e$.

This sampling detector has a coarser granularity since the most precision measurement of jets is devoted to the EM calorimeter. It uses scintillator tiles made of polystyrene, and steel as the passive medium in a sandwich-like configuration. The tiles and steel plates are radially oriented, perpendicularly to the beam direction. The scintillation light emitted by the tiles at the passage of ionising particles is collected on both tile edges by wavelength-shifting optical fibres as shows Figure 5.2. These fibres receive the light emitted by the scintillator at the ultraviolet region of the electromagnetic spectrum, converting it to visible light by the wavelength-shifting fluor doping the fibres, and connecting it to the readout Photo-Multiplier Tubes (PMT). The fibres are aluminised in the top opposite to the PMT, mirroring the light and increasing the light collection efficiency at the PMT photocathode.

A TileCal cell is constituted by grouping several optical fibres into the same readout channel, thus assembling a set of tiles to define the detector granularity of $\Delta\eta \times \Delta\phi = 0.1 \times 0.1$ in the barrel. Since each tile is read from two sides by independent fibres and PMTs, the energy deposit in each cell is also measured independently by two readout channels. The detector is arranged in 3 radial layers, uniformly segmented in ϕ to form 64 modules. In total, TileCal has 5182 cells read by 9852 channels.

Three calibration systems provide information at different stages of the TileCal readout chain: a ^{137}Cs γ source, a laser and a charge injection system. The modules were longitudinally drilled to accommodate a tube for the radioactive source passage, as sketched in Figure 5.2. The TileCal calibration system will be addressed in more detail in Chapter 5.

3.2.3 Muon Spectrometer

The Muon Spectrometer (MS) is involved in a toroid magnet that deflects the trajectory of muons and measures the position of their hits. It is located at the outermost part of the ATLAS detector and is composed of Monitored Drift Tubes (MDT) and Cathode Strip Chambers (CSC) for high precision measurements in the pseudorapidity range of $|\eta| < 2.7$, and Resistive Plate Chambers (RPC) and Thin Gap Chambers (TGC) dedicated to triggering purposes for $|\eta| < 2.4$. These components are sketched in Figure 3.12 and the MS main parameters are summarised in Table 3.5. The muon momentum and electrical charge are determined from the muon track curvature reconstructed from the hits.

The MDTs and CSCs are track precision chambers, that measure the muon hits coordinate in the bending plane, η. The triggering system of the MS, composed of the RPCs and TGCs, serve three purposes: provide information of prompt muon tracks to the ATLAS first level real-time trigger, identify the bunch-crossing and measure the coordinate orthogonal to the muon track bending plane, ϕ, to be combined with the η measurement from the precision chambers in offline reconstruction.

The measurements strongly rely on the alignment precision of the chambers and knowledge of the position of the MDT and CSC components, with a tolerance less than 30 μm . The ATLAS MS employs a high precision optical alignment system to fulfil these requirements.

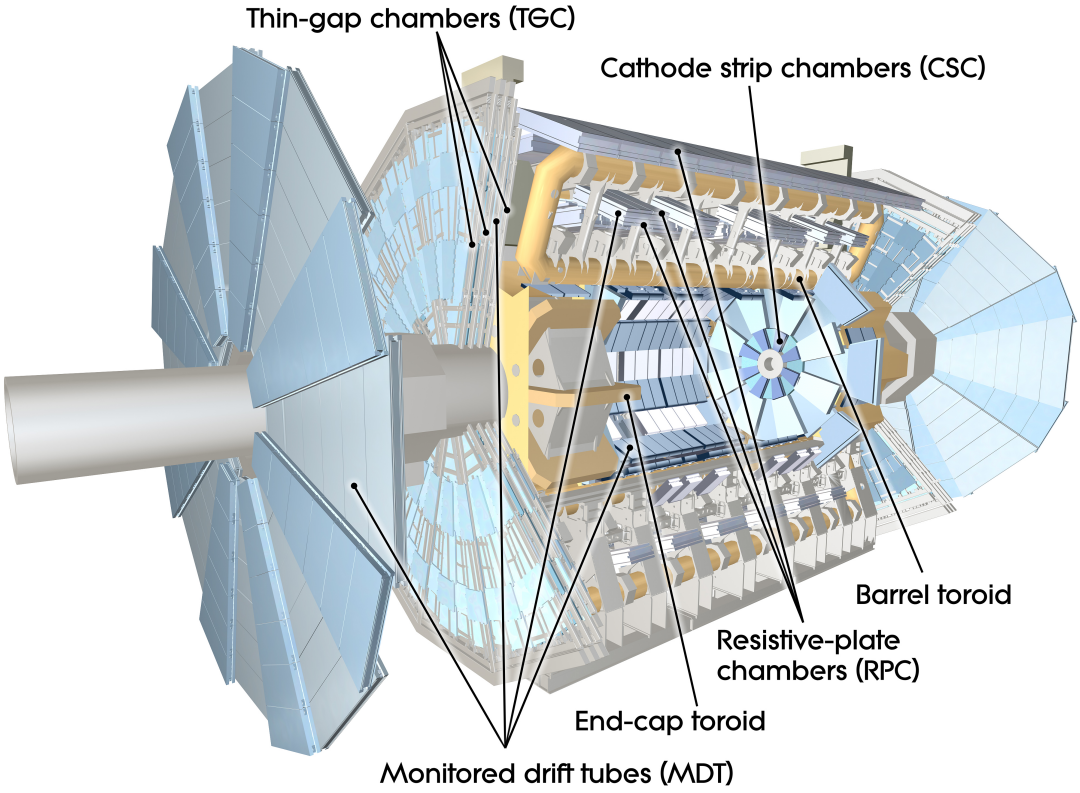


Figure 3.12: Layout of the ATLAS Muon Spectrometer. Taken from [30].

Precision Tracking Chambers

Monitored Drift Tubes (MDT) are installed in the barrel and end-cap regions covering $|\eta| < 2.7$, and measure precisely the muon momentum due to the high accuracy and the construction simplicity of this technology. The drift tubes are disposed along ϕ in layers configured in cylindrical rings in the barrel, and in wheels orthogonal to the beam pipe in the end-cap. The basic detector unit is a pressurised cathode tube of approximately 30 mm diameter, filled with Ar/CO_2 (93/7) gas and an anode wire made of tungsten-rhenium readout at the tube end. This configuration achieves a resolution of 80 μm per tube. Since the tubes in the barrel and end-caps are aligned in ϕ, this detection system does not provide the ϕ coordinate of the muon hit. This coordinate is obtained by matching the η measurement of the MDT with the (η, ϕ) coordinates determined by the trigger chambers.

The innermost layer of the MS in the forward region, covering the pseudorapidity range of $2 < |\eta| < 2.7$, is additionally equipped with Cathode Strip Chambers (CSC) to face the high counting rate needs at higher η. The layout consists of 2 disks perpendicular to the beam line, with multi-wire proportional chambers. The chambers have 4 planes with stripped copper cathodes and 2.5 mm pitch anode wires. Each of the two cathodes is segmented perpendicularly and parallel to the wires providing the measurement of the two track hit coordinates. The resolution obtained with this system is of 40 μm in the track bending plane and 5 mm in the transverse plane.

Monitured Drift Tubes (MDT)	
Coverage	$\|\eta\| < 2.7$
Number of Chambers	1150
Number of Channels	354000
Function	Precision tracking
Cathode Strip Chambers (CSC)	
Coverage	$2.0 < \|\eta\| < 2.7$
Number of Chambers	32
Number of Channels	31000
Function	Precision tracking
Resistive Plate Chambers (RPC)	
Coverage	$\|\eta\| < 1.05$
Number of Chambers	606
Number of Channels	373000
Function	Triggering, second coordinate
Thin Gap Chambers (TGC)	
Coverage	$1.05 < \|\eta\| < 2.7$
Number of Chambers	3588
Number of Channels	318000
Function	Triggering, second coordinate

Table 3.5: Main parameters of the ATLAS Muon Spectrometer. Taken from [30].

Trigger Chambers

The trigger chambers were designed to discriminate on muon transverse momentum, identify the bunch-crossing and provide coarse tracking information to the first level trigger in a fast manner. Besides, they provide the second coordinate measurement of the hit to complement the MDT measurement. In the barrel, the trigger chambers are composed of cylindrical layers of Resistive Plate Chambers (RPC) aligned with the beam line and covering $|\eta| < 1.05$. The end-caps, located at $1.05 < |\eta| < 2.4$, are equipped with circular disks of Thin Gap Chambers (TGC). Both have adequate space and time resolution and counting rate capability. The RPC is a gaseous electrode plate detector with η and ϕ readout pitch of 23-35 mm, and the TGC is a multi-wire proportional chamber spatial resolution of 1.8 mm. A chamber consists of readout wire planes, cathode planes and readout strip planes to read the ϕ coordinate.

Toroid Magnets

The magnetic field system used in the MS comprises one toroid barrel, $|\eta| < 1.4$, and two end-cap toroids, located at $1.6 < |\eta| < 2.7$, as depicted in Figure 3.12. Table 3.6 lists their principle characteristics. Together, they produce the toroidal magnetic field orthogonal to the muon trajectory that bends charged particle trajectories along the entire spectrometer volume. In the $1.4 < |\eta| < 1.6$ region, the magnetic field results from the combination of the barrel and and-caps fields. The magnets employ air-core technology to minimise the multi-scattering of muons. The generated toroidal field covers $|\eta| < 2.6$ and its associated uncertainty contributes

		Toroid barrel	Toroid end-cap
Size	Inner diameter (m)	9.4	1.65
	Outer diameter (m)	20.1	10.7
	Length (m)	25.3	5.0
	Number of coils	8	2×8
Mass	Conductor (t)	118	2×20.5
	Cold mass (t)	370	2×140
	Total assembly (t)	830	2×239
Coils	Turns per coil	120	116
	Nominal current (kA)	20.5	20.5
	Peak field in the windings (T)	3.9	4.1
Conductor	Overall size (mm^2	57×12	41×12
	Ratio Al:Cu:NbTi	28:1.3:1	19:1.3:1
	Temperature margin (K)	1.9	1.9

Table 3.6: Main specifications of the ATLAS toroid barrel and end-cap magnets. Taken from [30].

less than 3% to the muon momentum resolution degradation.

3.2.4 Trigger and Data Acquisition Systems

The ATLAS Trigger and Data Acquisition (DAQ) systems were designed to select interesting events for physics studies among spurious ones in real-time. They keep the event output rate and data amount at a level adequate for storage, and properly flow the data from its generation until record. The LHC 25 ns bunch crossing interval places tight requirements on the trigger system, that has to quickly decide to reject an event or send it to storage.

The trigger decision is based on the presence of high transverse momentum electrons or photons, jets, muons, and hadronically decaying tau leptons. In the search for flexibility, many of the ATLAS trigger parameters are configurable. A diagram of the trigger and DAQ systems is shown in Figure 3.13. In the Run I, the design strategy was to implement a three level trigger, where events are first evaluated by the hardware-based Level 1 (L1) trigger, then by the Level 2 (L2) trigger that runs over the Regions of Interest (RoI) seeded by the L1, and finally by the Event Filter (EF), where algorithms similar to the offline ones are executed. These last two levels are software-based, running in computing clusters. When an event is accepted by the whole trigger chain it is categorised according to the fulfilled trigger condition.

Two important requirements on the trigger levels are the latency, defined as the maximum allowed time to an algorithm to reach a decision, and the event output rate, the rate at which events are accepted. The L1 was designed to have a maximum processing time of 2.5 μs , the L2 of 40 ms and the EF of 4 s. Concerning the output rate, the L1 has the difficult task of reducing the event rate from 40 MHz to 75 kHz, subsequently reduced to 3 kHz by the L2 and to around 400 Hz by the EF. At this stage, the DAQ registers events at an approximate speed of 320 MB/s.

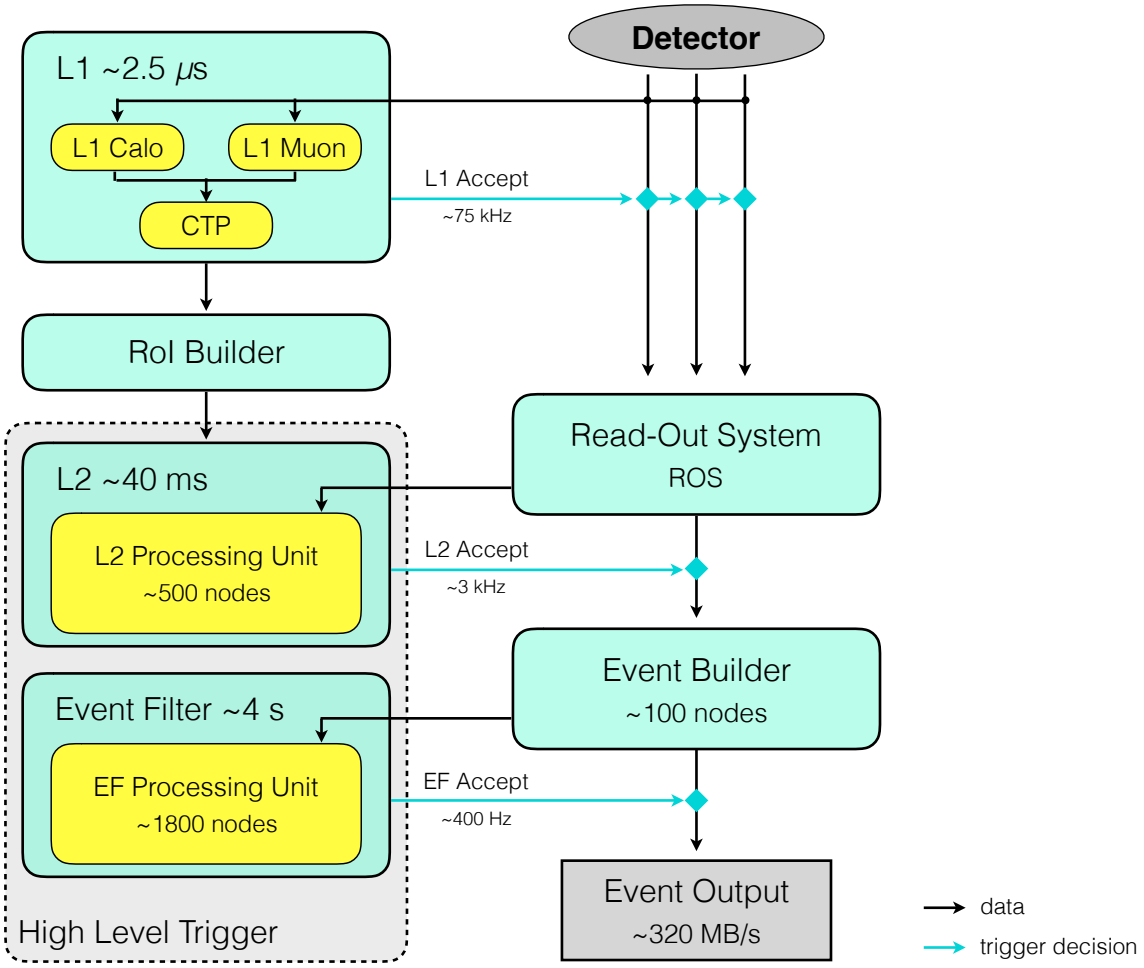


Figure 3.13: Diagram of the ATLAS trigger and data acquisition systems. Adapted from [30].

L1 Trigger

To strictly achieve the latency of 2.5 μs , the Level 1 trigger does not make use of the full granularity of the detector. It is divided in a muon trigger (L1 muon), a calorimeter-based trigger (L1 calo) and a Central Trigger Processor (CTP). The L1 calo dedicates to the selection of electrons, photons, hadronically decaying taus (τ_{had}), jets and event energy sums, starting by pre-processing the detector information. At this stage, the output signals are calibrated and compensated for different time-of-flight and path-lengths and associated to the correct bunch-crossing. The detector granularity is reduced by the analogue sum of the outputs of several cells, forming energy-integrated towers of $\Delta\eta \times \Delta\phi = 0.1 \times 0.1$ called trigger towers. The L1 calo trigger runs the electron/photon and the jet/energy trigger algorithms to evaluate the event. These algorithms consist of:

L1 calo e/γ and τ_{had} algorithm: This algorithm, illustrated by Figure 3.14, searches for 2×2 clusters of electromagnetic trigger towers, where the energy deposit in two adjacent towers exceeds a particular threshold. This leads to two possible configurations in the 2×2 cluster: 2×1 or 1×2. An isolation-veto threshold is set to the 12 electromagnetic trigger towers

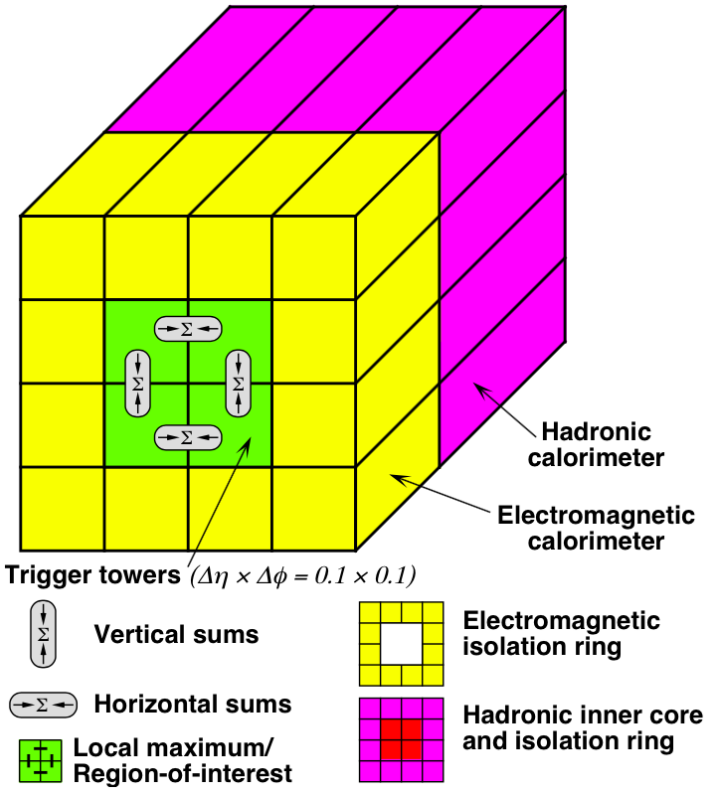


Figure 3.14: Illustration of the L1 e/γ and τ_{had} trigger algorithm based on $\Delta\eta \times \Delta\phi = 0.1 \times 0.1$ trigger towers showing the local maximum corresponding to the RoI. Taken from [30].

surrounding the 2×2 cluster and to the 4×4 hadronic cluster behind it. The algorithm scans the detector within $|\eta| < 2.5$ through a sliding window technique. It similarly searches for narrow jets corresponding to τ_{had} leptons, summing the hadronic towers behind the electromagnetic towers under evaluation. For the isolation-veto condition, only the outermost ring of 12 trigger units are considered. The centre of the triggered 2×2 clusters are the coordinates of the RoI to seed the L2 trigger.

L1 calo Jet/Energy algorithm: In the search for high transverse momentum jets, a similar algorithm is used. Instead of $\Delta\eta \times \Delta\phi = 0.1 \times 0.1$ trigger towers, it uses jet elements which are sums of 2×2 trigger towers in both the electromagnetic and hadronic calorimeters. The Figure 3.15 illustrates the procedure. It runs with different window sizes in the $\Delta\eta \times \Delta\phi$ plane: 0.4×0.4, 0.6×0.6 and 0.8×0.8, to identify 2×2 jet elements with transverse energy above the triggering threshold. The location of these local maxima defines the RoI coordinates. During the scan of the detector, the total transverse energy deposit and missing transverse energy are evaluated and thresholds applied to it.

On the contrary of the L1 calo that is seeded by the entire calorimeter system, the L1 muon only uses information from the dedicated and high-segmented RPC and TGC trigger chambers of the muon spectrometer, as described before in Section 3.2.3. These detectors have the adequate timing accuracy to associate the muon candidate with the right bunch crossing in

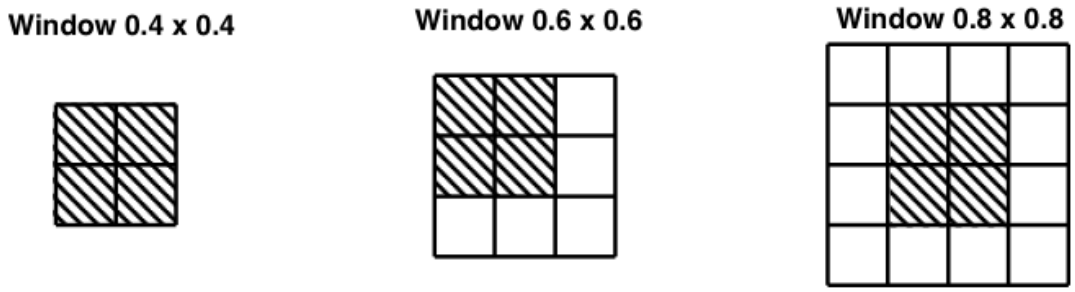


Figure 3.15: Illustration of the L1 jet trigger algorithm based in $\Delta\eta \times \Delta\phi = 0.2 \times 0.2$ jet elements for the three different window sizes in the $\Delta\eta \times \Delta\phi$ plane: 0.4×0.4, 0.6×0.6 and 0.8×0.8. The shaded 2×2 jet elements correspond to the RoIs. Taken from [30].

an unambiguous manner. The following describes the algorithm responsible for the muon-based trigger:

L1 muon algorithm: This algorithm requires coincidence hits across the trigger chamber layers within a trajectory hypothesis, tracking the muon path from the interaction point through the detector. The width of the trajectory tested is inversely proportional to the muon transverse momentum threshold to apply. In total, 6 operational p_T thresholds are available, 3 for low p_T triggers (ranging from 6 to 9 GeV) and 3 for high p_T triggers (from 9 to 35 GeV). This algorithm is performed independently in the η and ϕ projections to avoid accidental triggers of fake hits originated by noise. The η and ϕ trigger information is combined to form the RoI's to be sent to the L2 trigger.

Both the L1 calo and L1 muon processors report to the L1 central trigger processor. The information sent is a summary on the multiplicity of different triggered object types, electrons/photons, τ_{had}, jets and muons, and passed thresholds with corresponding RoI's. The CTP also receives flags indicative of the thresholds passed by the total and missing transverse energy. This processor combines the previous conditions to form up to 256 trigger menu items. Each trigger item has an associated pre-scaling factor. This pre-scale controls the fraction for which a given item will actually trigger the event, keeping the output triggered event rate within the required 75 kHz. Pre-scales are designed to record part of the statistics of events that happen at very high rate. The L1 accept signal is then the logical OR operation of all trigger items. The L1 outputs the RoIs to the RoI builder, and the L1 accept signal and passed menu items to the detector front-end electronics and Readout System (ROS). The CTP is also responsible for the lumiblock counter and performs the timing task, providing the whole detector system with a clock synchronous with the LHC proton bunch-crossing.

L2 Trigger

The L2 trigger makes use of the full granularity of the detector. It receives the L1 accept signal through the ROS that used the ROI information, typically representing about 1 to 2% of the full event data. Its main component is the L2 processing farm, where an event selection aimed to have a rejection factor of 30 is carried out. Refined selections are performed by requiring more information of the detectors to the ROS, and by executing improved algorithms that can be summarised as follows:

L2 trigger algorithms: The L2 operates by finding patterns and searching for a list of physics signatures. It runs a sequence of algorithms alternating methods of feature extraction, such as calorimeter clusters or well-defined tracks, with hypothesis algorithms to judge the matching between the feature and the hypothesis. The objective is to reject events as soon as possible, leaving the more time-consuming steps to be processed at the end, at lower event rate stage.

The selection algorithms also provide real-time information for quality monitoring. The results of the trigger analysis are incorporated into the final event by the Event Builder, that assembles the event as a single formatted data structure.

Event Filter

The last element in the triggering chain is the Event Filter, that unlike the L2 trigger, is based on standard offline ATLAS event reconstruction and physics analyses techniques. The generated data during analysis is appended to the event data structure, to seed the offline data analysis. Following the event selection performed in the EF, the selected events are classified according to the assigned physics signature into different streams, such as electrons and photons (Egamma); muons (Muons); jets, τ leptons and missing transverse energy (JetTauEtmiss). A tag with this information is appended to the event data structure. If an event has more than one physics signature, it will be associated with all the corresponding streams.

Event Output

The final step of the trigger and DAQ systems is the Event Output, performing the data storage operation to CERN's central data records with a transmission speed of 320 MB/s. It has sufficient buffering power to hold the data for 24 h in case of transmission failure. The event data is recorded in a set of files that map the trigger streams. Event overlap across streams happens at a minimal rate and has to be solved by subsequent analysis. Besides the physics streams, other streams are formed with a subset of events. This is the case of the express stream, containing events selected by the EF as useful events for monitoring and quality assessment of both the detector and the data, and dedicated calibration streams that provide the necessary information to calibrate the detector.

Chapter 4

Object Reconstruction and Performance

The reconstructed physical objects are the key ingredients on which any physics analysis relies on. The search for the $WH \to \ell\nu b\bar{b}$ process, with $\ell = e, \mu$, in the pp collisions provided by the LHC, requires an event topology characterised by one electron or muon, large missing transverse energy and at least two jets identified as resulting from the fragmentation of b-quarks.

This Chapter describes how these final state objects are reconstructed, identified and calibrated in the context of the ATLAS experiment. Tracks and vertices constitute the basis of charged particle detection and collision vertex finding. Their reconstruction, using the inner detector, is a common step of many object reconstruction techniques and for this reason also addressed in this Chapter. The same is true for the clustering algorithm that search for clusters of energy deposits in the calorimeters, the basic constituents of the calorimeter jets used in the analysis.

4.1 Tracks and Vertices

The reconstruction of charged particle tracks and vertices at the inner detector is essential to identify particles and allows to determine the collision points, referred to as primary vertices (PV). Besides, the ability of associating tracks to the main PV provides a way to suppress pile-up effects. On another hand, the track displacement with respect to a given vertex, called impact parameter, offers the opportunity to distinguish particles that emerged from the main interaction from those that did not. Consequently, impact parameter-based conditions are widely used in the physics analyses. Additionally, and as shall be seen ahead, tracks and vertices are the basis of jet flavour identification.

4.1.1 Track Reconstruction

Charged particles interact with the inner detector leading to hits in the sensitive detector units. Track reconstruction intends to recover the charged particle trajectory using the measured space point hits. In terms of the track reconstruction technique used in the ATLAS experiment [31], it is useful to define two kinds of particles: primary and secondary particles.

Primary particles are produced before reaching the ID and correspond, for instance, to particles with half-life greater than 3×10^{-11} s resulting from pp interactions, or stable products of promptly decaying particles produced by the collisions. Secondary particles are produced at the ID through the interaction of primary particles. The reason for this categorisation is that, contrary to primary particles, secondary particles do not cause hits on the first ID layers. Given this important distinction, two different track reconstruction methods are utilised: the Inside-Out and the Back-Tracing algorithms, respectively, reconstruct the primary and the secondary charged particle tracks [31].

Inside-Out

- Starts with 3 point hits in the silicon detectors consistent with a track segment;
- Iteratively adds hits moving away from the interaction point using the Kalman filter technique;
- The procedure is continued until the TRT hits are added.

Back-Tracing

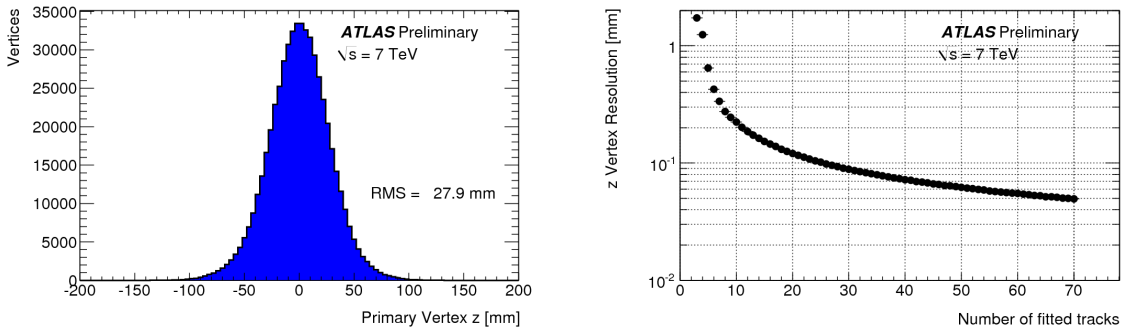
- Starts from track segments reconstructed in the TRT;
- Adds silicon hits backwards using the Kalman filter.

A common step to both algorithms is the usage of the Kalman filter. This technique uses information about the track segment under construction to predict the location of the next hit of the track according to a physical model. Once the hit has been added to the track segment, the prediction is updated. The physical model used to fit the hits is the helical path of charged particles in a uniform magnetic field, compensated by the particle energy loss in the ID. Once the track is fully reconstructed, the same model is used to determine the particle charge and momentum from the track parameters.

4.1.2 Vertex Reconstruction

The reconstruction of vertices with the ATLAS detector uses the charged particle tracks from the ID [31]. A vertex represents the point in space where the final particles of a collision emerged. In order to determine the vertex, an iterative algorithm is used to find a common origin point from several tracks. The z-coordinate of the crossing point between reconstructed tracks and the beam line are the seeds to the algorithm. The tracks used to seed the algorithm must fulfil general reconstruction quality requirements. The algorithm proceeds as follows:

- Iteratively add nearby tracks to the seed and use a χ^2 fit to find the common vertex;
- Tracks displaced from the fitted vertex by $> 7\sigma$ are removed from the current vertex and used to seed a new one;
- The process is repeated until no additional vertices can be found;



(a) (b)

Figure 4.1: (a) Distribution of the longitudinal position of the reconstructed primary vertices in $\sqrt{s} = 7$ TeV pp collision data events. (b) Resolution of the primary vertex longitudinal coordinate measurement as a function of the number of fitted tracks. Taken from [32].

- Vertices are required to have at least two tracks.

Figure 4.1(a) shows the reconstructed longitudinal coordinate of primary vertices for pp collisions at the LHC. The distribution is centred at the LHC nominal interaction point ($z = 0$). PVs can be displaced from the nominal IP up to 100 mm. The resolution of the z coordinate measurement improves substantially with the event track multiplicity as Figure 4.1(b) shows. A resolution of 2 mm is obtained for events with only 2 or 3 fitted tracks and 0.05 mm can be reached in the case of 70 tracks.

The main primary vertex of an event is defined as the vertex with larger squared p_{T} sum of all its associated tracks and identified as the collision point where the hard scatter took place.

4.2 Electrons and Photons

High energy electrons interact with matter dominantly by Bremsstrahlung emission, a process by which the electron emits a photon. High energy photons interact primarily through pair production: in the vicinity of a nucleus the photon converts into an electron-positron pair ($\gamma \rightarrow e^- e^+$). The repetition of these two processes governs the electromagnetic shower development. As stated before in Chapter 3, the ATLAS electromagnetic calorimeter was designed to stop and measure the electromagnetic cascades and is therefore an essential system to reconstruct and identify both electrons and photons. To distinguish between electrons and photons, the identification algorithms use the ID.

4.2.1 Electron Reconstruction

The reconstruction of electrons and photons with the ATLAS detector [33, 34], depicted in Figure 4.2, begins by reconstructing energy clusters in the EM calorimeter. Clusters are seeded by energy deposits above 2.5 GeV and formed using a sliding window algorithm. For central electrons located at $|\eta| < 2.5$, the algorithm window size corresponds to 3×5 of 0.025×0.025

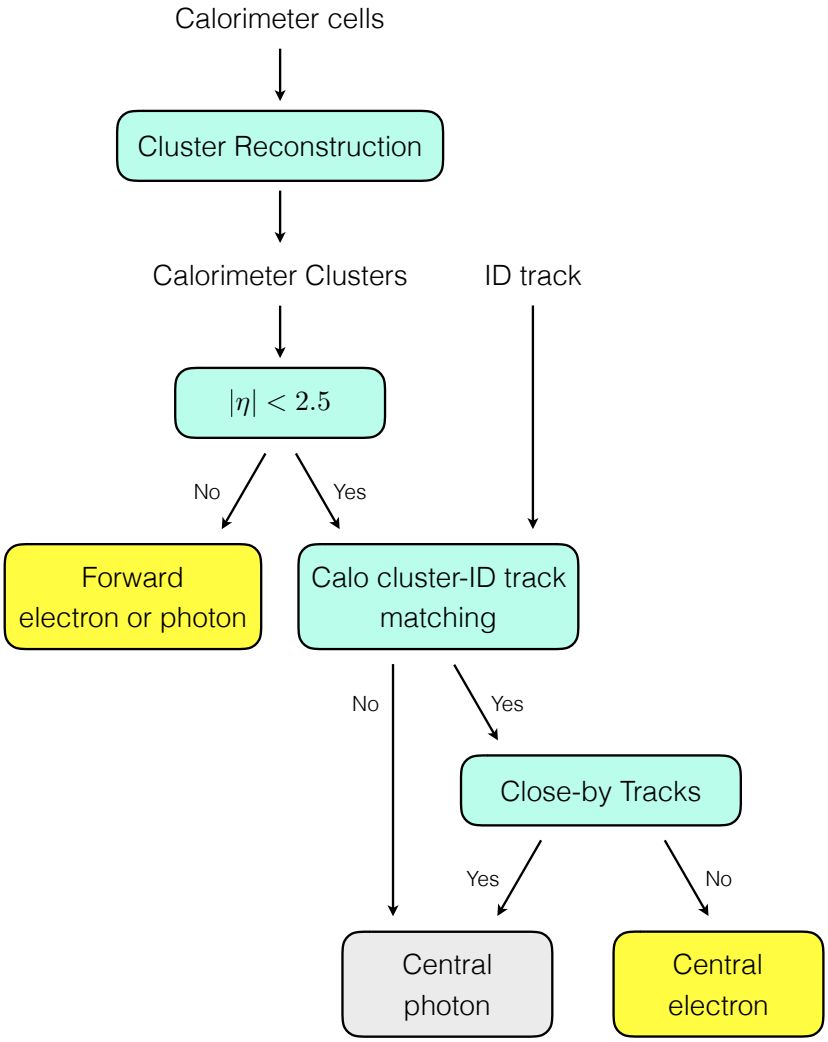


Figure 4.2: Diagram of the electron and photon reconstruction chain with the ATLAS detector.

cells in the (η, ϕ) plane. The cluster reconstruction efficiency is above 99% for MC simulation electrons of $E_{\mathrm{T}} =>15$ GeV [34].

The second step of the electron reconstruction algorithm is the spatial matching between the cluster and a track reconstructed in the ID. Since the ID does not cover the forward region, the primary difference between central and forward electrons is that forward electrons are indistinguishable from photons. Cluster-track association is done as follows: tracks with $p_{\mathrm{T}} > 0.5$ GeV are extrapolated to the EM calorimeter and associated with a cluster if the distance between the track and the cluster barycentre is $|\Delta\eta|$ <0.05 and $\Delta\phi$ <0.1. An electron candidate must have at least one associated track, while a photon does not have any. Pairs of close-by tracks with common vertex displaced from the IP are investigated to distinguish between electrons and converted photons (i.e. photons that undergo pair production in the ID). If two close-by tracks are found the two associated clusters are reconstructed as a photon.

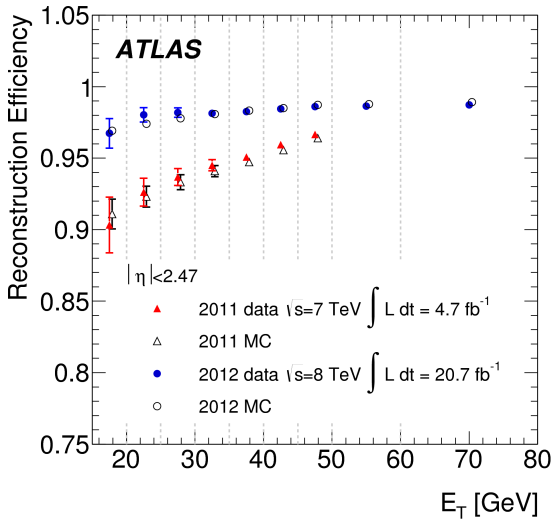


Figure 4.3: Electron reconstruction efficiency a function of the electron E_T for data and MC. Taken from [34]

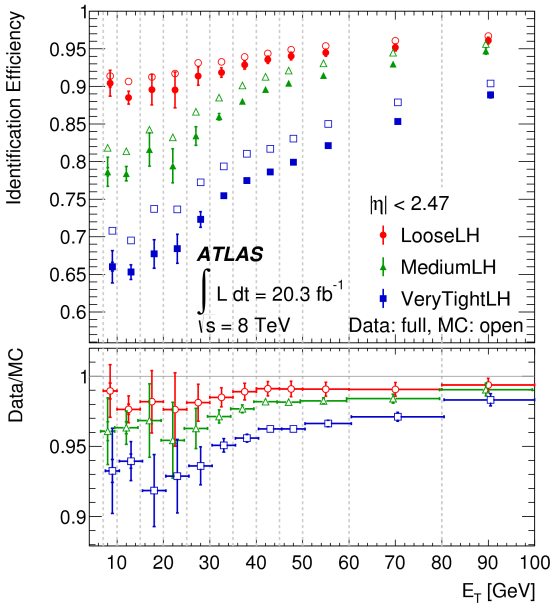


Figure 4.4: Electron identification efficiency for the LooseLH, MediumLH and VeryTightLH classifications, as a function of the electron E_T for data and MC. The ratio plot shows the data-to-MC efficiency ratio. Taken from [34].

The efficiency of the reconstruction, shown at Figure 4.3, is 97% (99%) for electrons with transverse energy E_T of 15(50) GeV. This measurement employs tag-and-probe methods using samples of $Z \to e\bar{e}$ and $J/\psi \to e\bar{e}$. Residual differences in efficiency for data and MC, not greater than 0.5%, are corrected by applying event scale factors to simulated events containing electrons.

4.2.2 Electron Identification

After reconstruction, dedicated algorithms are used to identify signal electrons among background electrons resulting from light jets, semi-leptonic decays of heavy flavour hadrons and electrons from photon conversions.

A MVA likelihood (LH) method is used to identify electrons according to different quality criteria and purity levels: VeryLooseLH, MediumLH and VeryTightLH [34], in ascending order of background rejection and decreasing order of electron identification efficiency. The inputs to the method are Probability Density Functions (PDFs) of discriminant observables for signal and background electrons. Using the set of PDFs, the LH method determines an overall probability for the object to be a signal or background electron.

The electron identification is established by a cut on the LH discriminant. The three purity levels are achieved by setting three different LHs with different input variables. The VeryLooseLH classification makes use of electromagnetic shower shape variables measured with the EM calorimeter, quality parameters of the track and of the cluster-track matching, and

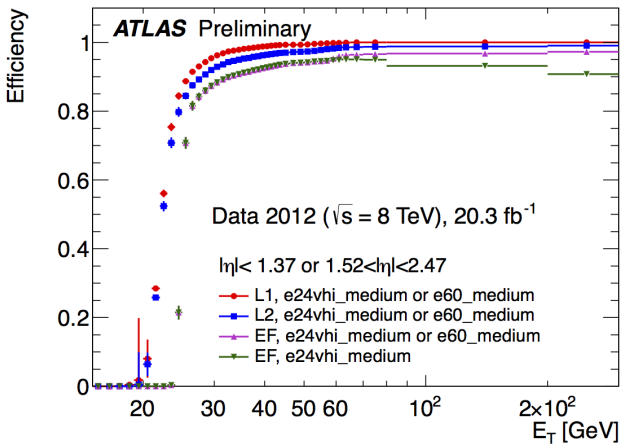


Figure 4.5: Efficiency of the `e24vhi_medium1` or `e60_medium1` electron triggers at the ATLAS trigger Level 1 (L1), Level 2 (L2) and Event Filter (EF) as a function of the offline electron E_{T}. Taken from [35].

hadronic leakage information. The hadronic leakage is evaluated from the fraction of energy deposited in the hadronic calorimeter. The MediumLH classification adds to these variables the number of hits in the innermost layer of the pixel detector to reject electrons from photon conversions, the transverse impact parameter d_0 and transition radiation information from the TRT to reject heavy charged hadron backgrounds. The VeryLooseLH and MediumLH input variables are used in the VeryTightLH classification that adds to those the veto on reconstructed photon conversions resulting from the investigation of close-by tracks as described before. Discriminators related to electron isolation, aiming to reject background electrons from jets for instance, are not used by the identification algorithms because the isolation requirements depend on the physical process under analysis, and are not incorporated into the LH for sake of flexibility.

Figure 4.4 shows the electron identification efficiency measured in tag-and-probe electron samples for data and MC. The efficiency is larger for the LooseLH classification, always above 90% and increasing with the electron E_{T}. The more pure VeryTightLH selection has a substantially lower efficiency, 65% for 10 GeV electrons. The efficiency differences between data and MC are used to correct MC through event weights attributed to simulation. The magnitude of the correction do not overcome 7% for electrons with $E_{\mathrm{T}} > 20$ GeV.

4.2.3 Performance of the Electron Trigger

Electron-based triggers employ the Level 1 and Level 2 trigger algorithms described in Section 3.2.4 and the offline reconstruction technique described above at the Event Filter (EF) level of the ATLAS trigger system. The performance of the electron triggering chain was measured in data using pure samples of high quality electrons, and is shown in Figure 4.5 for the lowest E_{T}-threshold un-prescaled electron triggers, `e24vhi_medium1` and `e60_medium1`, of E_{T} thresholds of 24 GeV and 60 GeV, respectively. The various curves reach a stable plateau after

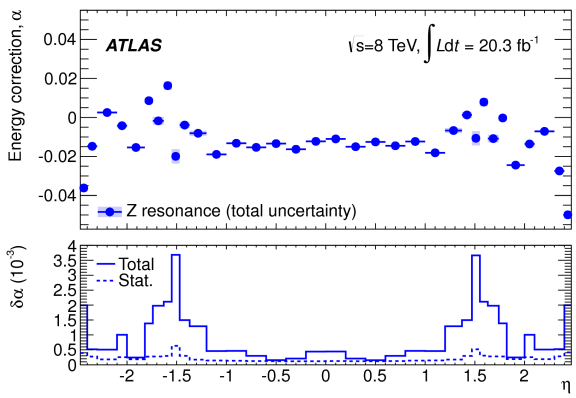


Figure 4.6: Electron energy scale corrections α, defined as $E_{data} = (1+\alpha)E_{MC}$, derived from data and Monte-Carlo simulated samples with $Z \to ee$ events, as a function of η. Taken from [36].

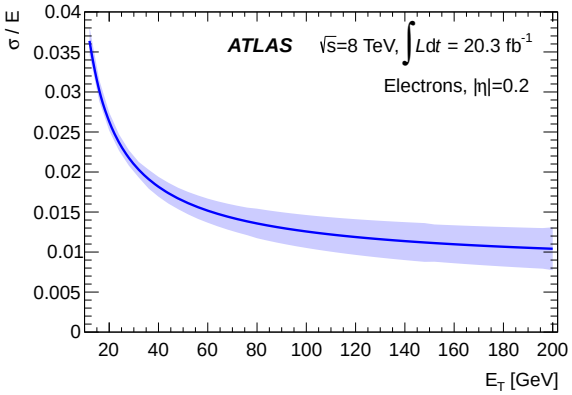


Figure 4.7: Electron energy resolution as a function of the electron E_{T}. Taken from [36].

the turn on curve between 20 and 30 GeV. For electrons with $E_{\mathrm{T}} > 30$ GeV, the EF efficiency of these triggers is above 90%.

The corresponding efficiencies for MC events differ slightly from data. For that reason, when electron-based triggers are used for event selection, MC data is corrected to account for this feature. The correction is applied through event scale factors that do not differ more than 3% from the unity.

4.2.4 Electron Energy Scale, Resolution and Calibration

The electron reconstructed energy is the total cluster energy, at the EM scale, corrected for the estimated energy loss in passive material in front of the calorimeter and outside the cluster, while the electron spatial coordinates η and ϕ are taken from the matched track in the case of central electrons and from the cluster barycentre in the case of forward electrons.

Then, the absolute electron energy scale is determined through calibration procedures that compare the electron energy measurement for data and Monte-Carlo (MC) simulation. High quality electrons from the $Z \to e\bar{e}$, $J/\psi \to e\bar{e}$ and $W \to e\nu$ processes are used for this purpose. The measured energy scale calibration factors are shown and defined in Figure 4.6 [36]. These factors are only applied to experimental data electrons. They do not exceed a 5% correction for forward electrons and 2% for the central ones.

The energy resolution for electrons and its uncertainty is shown in Figure 4.7. The plot is compatible with the design requirements of the energy measurements of the ATLAS electromagnetic calorimeter. This resolution is determined for data, mainly by studying the $Z \to ee$ peak, and the uncertainties related to this approach are the ones that most contribute to the uncertainty on the resolution. In addition, the resolution uncertainty comes from pile-up effects, and uncertainties related to the distribution of the passive material in front of the EM calorimeter.

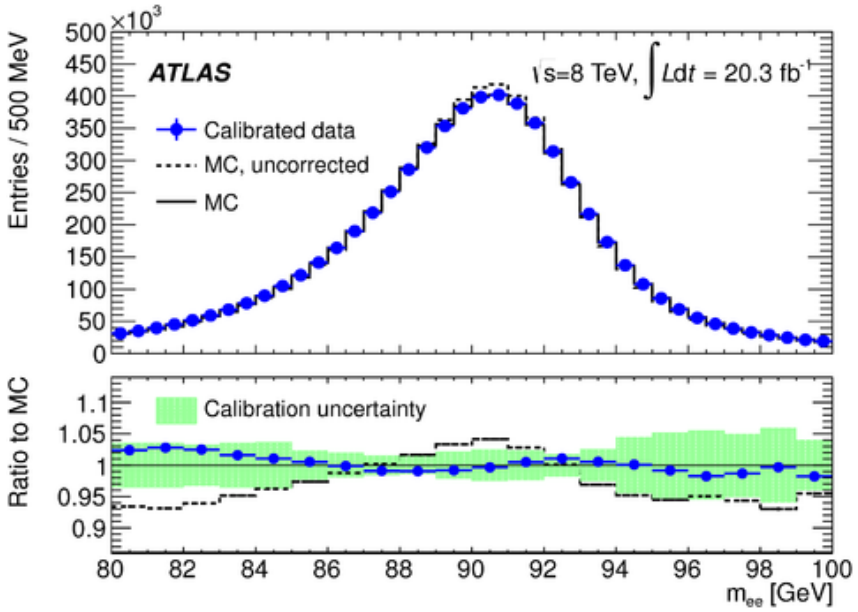


Figure 4.8: Electron pair invariant mass distribution for $Z \to ee$ decays in data and simulation. The distribution for simulation is shown with and without resolution smearing corrections. For data, energy scale corrections are applied. Taken from [36].

Moreover, the energy resolution for data and MC differ slightly. A smearing factor taking into account their difference is applied to the simulated electron energy to correct this feature. Figure 4.8 shows the effect of this correction in the Z boson mass peak.

4.3 Muons

Muons do not decay promptly after production. Having a mean life time of 2.2×10^{-6} s, ultra relativistic muons can travel about 650 m before the decay takes place [37]. Within a momentum range of 1 to 100 GeV, muons are fairly approximated to minimum ionising particles. Therefore, the muons produced by the LHC proton collisions can traverse the entire detector without losing their total energy through interaction with matter. They are detected and measured by the ATLAS detector combining the information from the MS, the ID and also the calorimeters to provide maximum pseudorapidity coverage and momentum resolution [38].

4.3.1 Muon Reconstruction and Identification

The muon reconstruction procedure depends on the tracks reconstructed in the MS and in the ID, and on the energy deposits in the calorimeter. Track measurements in the ID are limited by the coverage of the detector size itself that spans for $|\eta| < 2.5$. From the MS side, two regions have pronounced acceptance losses. In $\eta \sim 0$, the spectrometer is not totally equipped with chambers to provide space for the ID and calorimeters services. In $1.1 < \eta < 1.3$, certain regions in ϕ have only one layer of chambers installed. To overcome these features, ATLAS combines the information from the three subdetectors to recover efficiency in the reconstruction

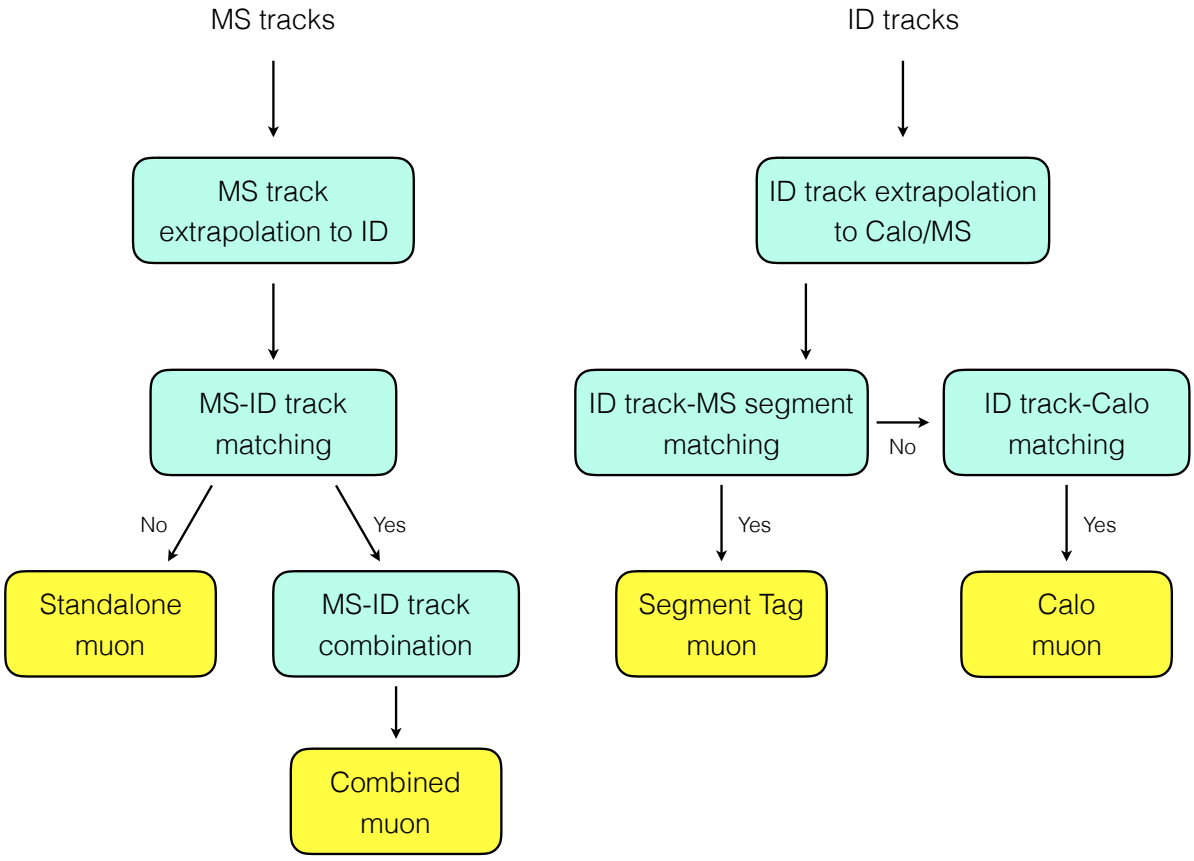


Figure 4.9: Diagram of the muon reconstruction chain with the ATLAS detector.

of muons, as Figure 4.9 sketches. Five reconstruction types of muons are established: Stand-Alone (SA), Combined (CB), Segment-tagged (ST), Calorimeter (Calo) and Silicon Associated Forward (FW).

Stand-Alone Muons SA muons are reconstructed only by the MS. The MS track is then extrapolated to the beam line to determine the transverse and longitudinal impact parameters. The extrapolation procedure takes into account the energy loss in the calorimeters. In the MS, the track reconstruction algorithm starts by finding track segments in each spectrometer layer. A MS track must have at least two track segments, meaning that the muon traversed at least two chamber layers. These are then combined into a single track that is afterwards extrapolated to the ID.

Combined Muons Combined muons have matching tracks independently reconstructed in the ID and MS, that are then combined into a single track. Several algorithms are available for the combination. The one used in the WH analysis starts by propagating the MS track back to the beam line taking into account the effects of the magnetic field and energy loss in the calorimeter. For isolated muons, the energy loss is the measurement from the calorimeter cells traversed by the propagated MS track. Non-isolated muons rely on parametrisations of the energy loss. The combined track is formed by a global ξ^2 fit to the hits of the MS and

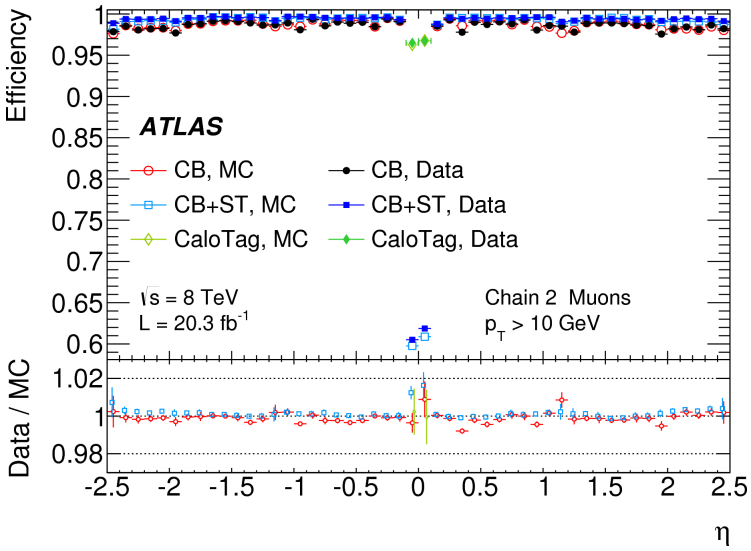


Figure 4.10: Reconstruction efficiency of Combined (CB), Combined + Segment-tag (CB+ST) and Calorimeter (CaloTag) muons as a function of η. Chain 1 refers to the ID-MS track combination algorithm used. Taken from [38].

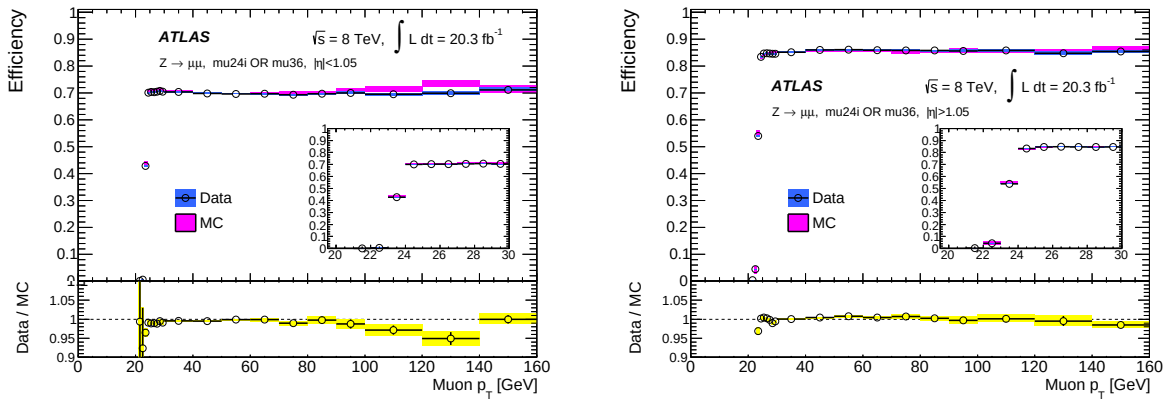
ID stand-alone tracks. The four-momentum and electrical charge of the CB muons is obtained from the combined track.

Segment-tagged Muons ST muons do not have a full reconstructed track in the muon chambers and only the ID track is available. ST muons are those whose ID track is extrapolated to the MS precision chambers, MDT or CSC, and match a track segment formed from only one layer of hits.

Calorimeter Muons For Calo muons, only the ID track is available and so there is no possible ID and MS track association. The muon identification relies in the association of the ID track with an energy deposit in the calorimeter consistent with a minimum ionizing particle.

Silicon Associated Forward Muons Silicon associated forward (FW) are stand-alone muons reconstructed in the forward region that spatially match an ID track segment.

The majority of muons are reconstructed as combined muons and this is also the type with highest muon purity. SA muons, for instance, have as background non-isolated muons from π or K decays in the calorimeter, and therefore have lower purity. In fact, muons other than CB are more important to extend the detector acceptance to regions where the detector configuration prevents the combined identification. The SA type recovers the $2.5 < |\eta| < 2.7$ region not covered by the ID, while the Calo type renders possible the muon reconstruction for the uninstrumented central region of the MS located at $|\eta| < 0.1$. ST muons are used in cases where the muon only traverses one layer of the MS and therefore the MS track is not reconstructed. This happens to low p_{T} muons or in the $1.1 < \eta < 1.3$ region only covered by



(a) (b)

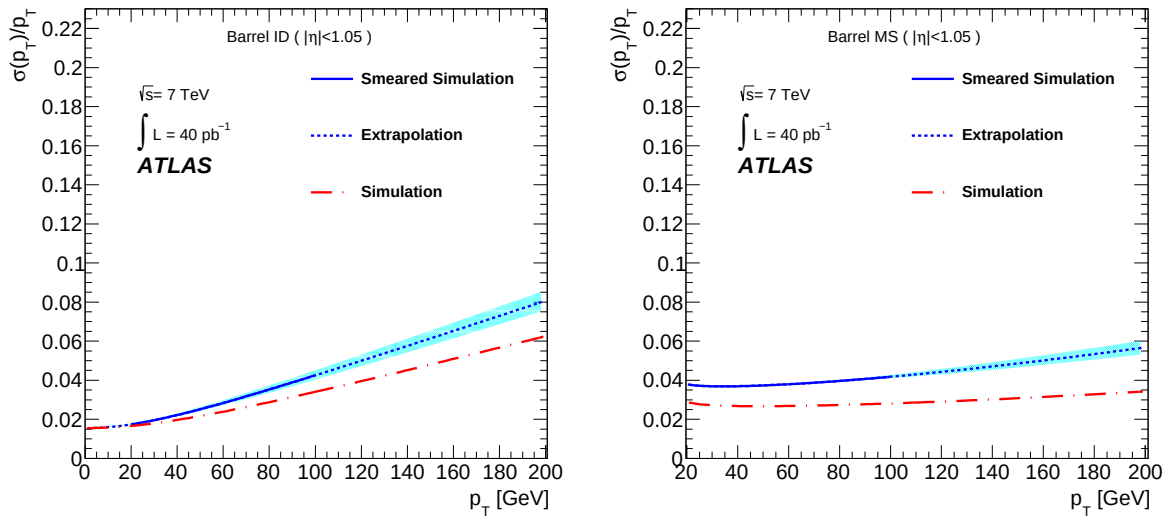
Figure 4.11: Efficiency of the `mu24i` OR `mu36` muon triggers as a function of the probe muon p_{T} for $Z \to \mu\mu$ events in data and simulation in the (left) barrel region ($|\eta| < 1.05$) and (right) end-cap region ($|\eta| > 1.05$). The bottom plots show the data-to-MC efficiency ratio. Taken from [39].

one chamber layer.

The reconstruction efficiency of the various types of muons and their momentum scale and resolution is determined from simulated and real data samples of $Z \to \mu\bar{\mu}$ events using tag-and-probe methods [38]. The reconstruction efficiency of various types of muons is shown as a function of η for data and MC in Figure 4.10, proving how the inclusion of the Calo type recovers the efficiency at $|\eta| \sim 0$. Combining all the reconstruction types, the efficiency is 99% and uniform across η. Residual efficiency differences between MC and data, shown in the ratio plot of Figure 4.10, must be considered when using events containing reconstructed muons, usually by weighting MC events to correct for this small discrepancy.

4.3.2 Performance of the Muon Trigger

Figure 4.11 shows the efficiency of the muon triggering chain for the lowest p_{T}-thresholds un-prescaled muon triggers, `mu24i` and `mu36` of 24 GeV and 36 GeV p_{T}-thresholds, respectively. The L1 and L2 muon trigger algorithms were described in Section 3.2.4, and the Event Filter uses the offline tracking algorithms described above [39]. In the end-cap region of $|\eta| > 1.05$, the trigger reaches a stable plateau of 85% efficiency at p_{T} =24 GeV. In the barrel, the efficiency at the plateau is degraded to 70% due to the uncovered region around $\eta \sim 0$. The efficiency differences between MC and data are shown in the bottom panels. The values are used to correct MC through event scale factors when muon-based triggers are employed. This correction is at the $\mathcal{O}(1\%)$ level.



(a) (b)

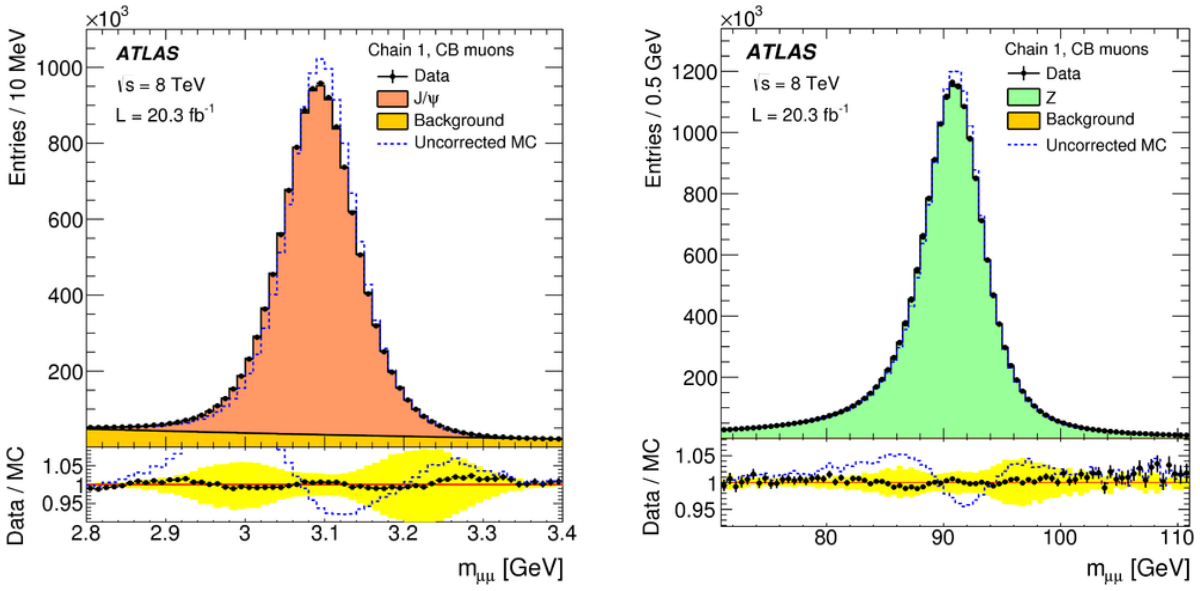
Figure 4.12: Muon momentum resolution for the (a) inner detector and (b) muon spectrometer for MC, before and after resolution smearing. Taken from [40].

4.3.3 Muon Momentum Scale, Resolution and Calibration

For the momentum scale and resolution assessment, samples of lighter resonance decays $J/\psi \to \mu\bar{\mu}$ and $\Upsilon \to \mu\bar{\mu}$ are also used, in addition to $Z \to \mu\mu$, as sources of lower momentum muons. The simulated muon momentum is corrected to match the determined momentum scale. The correction factors are parametrised according to the pseudorapidity of the muon and are typically 0.1% for the MS reconstructed momentum and -0.1% for the ID. Only for regions of $|\eta| \sim 1.05$, the momentum scale correction is larger, up to 0.3% in the MS [38].

Figure 4.12 shows the p_{T} resolution for muons, measured from the ID and MS tracks. As expected, the ID has better resolution than the MS for low-p_{T} muons due to the finer granularity. For large-p_{T} muons, the MS performs a more precise measurement of the curvature of the track, and hence of the muon p_{T}, because of its dimension. For muon momentum measurement involving the combination of the two tracks, a resolution of 1 to 3%, depending on the muon momentum and on the pseudorapidity region, is obtained in the di-muon invariant mass[38].

The real data has worse resolution on the muon momentum than simulation due to residual misalignments of the ID and MS not taken into account in the detector simulation. A resolution smearing correction to account for the difference is applied to the simulation. The smearing is below 10% for the ID measurement and below 15% for the MS [38]. The scale and resolution corrections effect on the MC muons of the J/Ψ and Z decays is shown in Figure 4.13 for the invariant mass of the di-muon system. The distribution is only slightly changed by these corrections and agrees better with the real data measurement of the resonance peaks afterwards.



(a) (b)

Figure 4.13: Di-muon invariant mass distribution for (a) $J/\psi \to \mu\mu$ and (b) $Z \to \mu\mu$ decays in data and corrected and uncorrected simulation. The MC momentum correction includes both the scale and the smearing corrections. Taken from [38].

4.4 Jets

Quarks and gluons exist confined in bound states called hadrons. This confinement prevents them to be observed as free particles and whenever a prompt parton is produced it hadronises, i.e., it radiates gluons and $q\bar{q}$ pairs in a shower that ends when all the particles produced are bound into colourless hadrons. These hadrons may be unstable and many decay shortly producing more hadrons, leptons or photons. This shower of particles is the experimental signature of quarks and gluons and is referred to as jet. Unlike lighter quarks, the top-quark does not hadronise since it decays promptly after production and its decay products are the detector signature of tops.

The hadronisation mechanism responsible for jet formation is not well described by theory but several models exist to fit experimental observations to a physical description able to predict their main features. They mainly imply that the momentum and quantum numbers of the closest hadron to the jet axis follow the momenta and quantum numbers of the parton originating the hadron. In this way, the parton physical properties are reflected on the hadronic shower. The string model described in Section 2.2 is one of the models commonly in use due to its ability to describe many of the jet observables such as energy and shower shape and substructure properties.

4.4.1 Overview of the Jet Reconstruction and Calibration chains

The jets used in the $WH \to \ell\nu b\bar{b}$ analysis are reconstructed with the ATLAS detector primarily through the electromagnetic and hadronic calorimeters measurements of the energy

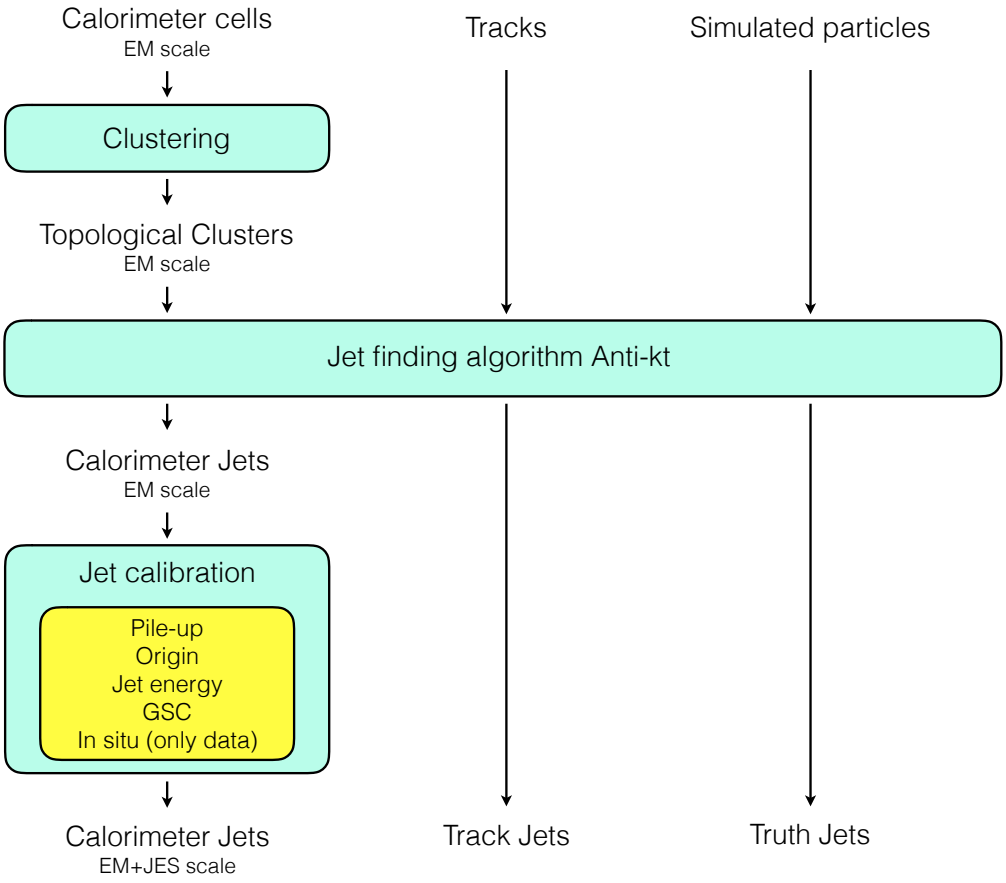


Figure 4.14: Diagram of the jet reconstruction and calibration chain with the ATLAS detector.

deposited by the particles of the shower. Information from tracks reconstructed by the ID and MS is incorporated only at the energy calibration stage, and is less relevant for the basis of jet reconstruction [41, 42].

Due to the non-compensating nature of the ATLAS calorimeters, the jet energy measurement needs to be corrected for the detector response difference with respect to the electromagnetic response, in which the calorimeter is calibrated - called EM scale.

The jet reconstruction and calibration chain is organised as Figure 4.14 shows [41, 42]. The key mechanism to reconstruct jets is the jet finding algorithm that can have different inputs: clusters of energy deposited in the calorimeters are the inputs to reconstruct calorimeter jets, particle tracks are used to form track jets and simulated particles are used to build truth jets. Calorimeter jets are the main jets used in the ATLAS physics analyses, while truth jets, available only at simulation level, provide the truth reference base to calibrate the properties of calorimeter jets. On the contrary of energy clusters, the tracks can be associated with the collisions PVs, and for that reason jets built from tracks are less affected by pile-up effects and constitute a useful reference to study the effect of pile-up in jets. However, since track jets do not include the energy of neutral particles, these are not used for physics analyses.

For calorimeter jets, the reconstruction chain is more complex than for truth or track jets. First, the calorimeter cells are grouped to form topological clusters in the calorimeter using

the Topological Clustering algorithm. These then feed the jet finding algorithm that forms calorimeter jets, with energy measured at the EM scale. A series of calibrations follow, in order to match the properties of calorimeter jets to the properties of truth jets, in average.

4.4.2 Topological Clustering Algorithm

A calorimeter jet corresponds to a reconstructed shower of particles that deposit energy in the calorimeter. The identification and energy measurement of the particles constituting the shower is therefore the first step of the calorimeter jet reconstruction. To do so, an algorithm joins the energy measured by the calorimeter cells to form clusters representing the energy deposited by a single particle of the shower.

The cells are grouped by the Topological Clustering algorithm [42], a method that suppresses cell noise by considering three different energy-to-noise ratio thresholds to the cells measurement. It starts by classifying the cells according to these thresholds:

- cells with energy-to-noise ratio E/σ above 4 are classified as seeds;
- cells with $E/\sigma > 2$ are classified as growing cells;
- cells with $E/\sigma > 0$, i.e. that have a positive energy measurement, are classifies as terminal cells.

Then the algorithm groups the cells starting by the seeds and adding neighbouring cells. The procedure is iterated until all neighbouring growing cells are clustered. At the end of the cluster growing phase, all terminal cells located at the cluster frontier are added to the cluster. At this stage, clusters resulting from close-by particles may have been joined into a single large cluster. In order to separate these clusters, a splitting algorithm identifies the energy valleys between energy maxima corresponding to different particles, and splits the cluster accordingly.

The formed topo-clusters are defined to have a null mass and an energy corresponding to the sum of the energy of their cells, therefore at the EM scale. Some jet energy calibration schemes used in ATLAS start by calibrating the cluster energy, using the Local Cluster Weighting method for instance, but that is not the case of the VH analysis. The direction of the cluster is the energy weighted average of the direction of the cells relative to the ATLAS nominal interaction point.

4.4.3 Jet Reconstruction

Several algorithms exist to find jets from different objects such as energy clusters, particle tracks or truth particles. Generally, the jet finding methods can be assigned to one of the following two families: cone algorithms or sequential algorithms. Cone algorithms are seeded by a local hardest object defining the cone axis. All surrounding objects falling within a predefined sized cone around the axis are added to form the jet. Following a first iteration,

a split step is implemented to resolve overlapping jets. Sequential algorithms, on the other hand, sequentially add objects together, without targeting a predefined shape.

Each class has its advantages. In general, cone algorithms have more regular boundaries and shapes and consequently are better for calibration, but the adaptable jet shape offered by sequential algorithms suits better the branching nature of hadronisation and shower development. Important features of jet finding algorithms are how the presence of soft objects changes the reconstructed jets, and how is the jet affected when its hardest object splits collinearly during fragmentation. The stability with respect to these effects is known as infrared safety and collinear safety, respectively, and these are both desired properties. Cone jets are collinear unsafe, since the seeding objects for the algorithm change substantially while splitting, and can also be infrared unsafe, depending on how the jet overlap step is conducted. Most of the sequential algorithms are infrared unsafe due to their adaptable essence.

The Anti-kt Jet Reconstruction Algorithm

The anti-kt algorithm [43] is the default jet reconstruction algorithm used in ATLAS. This sequential algorithm gathers the best of both jet reconstruction classes, giving origin to jets of almost conical shape that are not much influenced by soft radiation. This last feature is of particular importance under the severe pile-up conditions of the LHC. The algorithm starts by defining the energy weighted distance d_{ij} between each combination of two objects i and j:

$$d_{ij} = \min\left(\frac{1}{k_{\mathrm{T}i}^2}, \frac{1}{k_{\mathrm{T}j}^2}\right)\frac{\Delta_{ij}^2}{R^2} \tag{4.1}$$

where $\Delta_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$ and $k_{\mathrm{T}i}$, y_i and ϕ_i are respectively the transverse momentum, rapidity and azimuthal angle of the object i. R is the radius parameter governing the resulting jet size. To reconstruct jets the algorithm proceeds as follows:

- Identify the pair of objects yielding the smallest d_{ij}
- If $d_{ij} < \min\left(\frac{1}{k_{\mathrm{T}i}^2}, \frac{1}{k_{\mathrm{T}j}^2}\right)$ merge the objects into a pseudo-jet being formed
- Otherwise, if $\frac{1}{k_{\mathrm{T}i}^2}$ ($\frac{1}{k_{\mathrm{T}j}^2}$) is smaller than d_{ij}, i (j) is called a jet and removed from the list of objects in the event (this is the stopping criterion for a single jet definition)
- Proceed to the first step until no objects are left in the event

With this method, pairs of objects that are either very close together or the hardest object and its closest one are successively merged, such that objects within a distance R to the pseudo-jet are iteratively added to form the jet. The stopping criterion prevents the jet to grow with the addition of soft objects outside the radius parameter, making it infrared safe. The jet overlap is solved by method construction with the anti-kt: the more energetic jet will be reconstructed first and will therefore be conical while the less energetic overlapping jet will be missing

the overlapping piece. The inputs to the algorithms can be parton or hadron-level simulated particles or detector-level tracks or calorimeter clusters.

This analysis uses calorimeter jets resulting from applying the anti-kt jet finding technique, with chosen radius parameter of 0.4, using topological calorimeter clusters as inputs. The four-momentum of the jet is determined from the sum of all its clusters four-momenta.

4.4.4 Jet Energy Scale and Calibration

The jet calibration scheme is applied to calorimeter jets measured at the EM scale aiming mainly to restore the jet energy scale. The EM scale was initially determined by a beam test of the calorimeters, and is regularly maintained through calibration of their response over time. It reconstructs the energy deposited by electrons and photons correctly but does not include any corrections for the loss of signal for hadrons. The jet particle shower consists of several components: electromagnetic showers resulting, for instance, from $\pi \to \gamma\gamma$, hadron energy deposited in the calorimeter, escaped energy due to hadron decays to neutrinos and muons, and invisible energy associated with nuclei break up and nuclear excitation. The last two components are not detected by the calorimeter and this leads to a jet energy measurement that is typically 15 to 55% lower than it truly is [44]. This effect is referred to as the non-compensating hadronic energy measurement.

A chain of calibration procedures, schematically exhibited in Figure 4.14, is applied to the calorimeter jets used by the VH analysis. The very first step of the calibration chain consists in the correction of pile-up effects. Afterwards, an origin correction adjusts the jet direction taking into account the measured coordinates of the PV where the jet had origin. Then, the jet energy is calibrated to account for the effects of the non-compensated measurement of the calorimeters. The jet energy resolution is then improved by correcting the dependence of the calorimeter response to different jet flavours. Finally, residual differences in the jet energy response between real data and simulated jets are mitigated by calibrating real data jets using in-situ techniques. These procedures are described in what follows.

Pile-up correction

For the high luminosity LHC program, the effects of collisions pile-up cannot be neglected, specially in what comes to jets that are broad objects and multiple particle compounds. The pile-up can give rise to background clusters that do not match any of the signal shower particles or can simply overlap signals resulting from the hard scatter. A pile-up correction is applied to the calorimeter jet energy measurement to subtract the pile-up contribution. In the technique used, these contributions that do not originate from the main collision event are quantified and measured as an energy area density [45]. Then, the pile-up energy density times the reconstructed jet area is subtracted from the calorimeter jet energy measurement.

This correction is effective in mitigating effects of in-time pile-up that increase the cell signal, but are ineffective in what comes to out-of-time pile-up that generally results in lowering

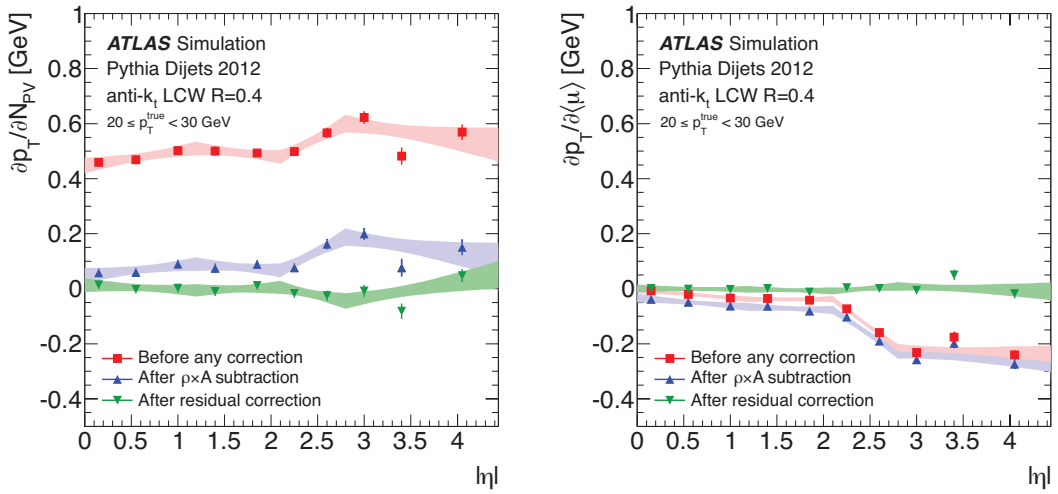


Figure 4.15: Dependence of the jet p_{T} on the (left) in-time pile-up measured through the number of primary vertices N_{PV} and on the (right) out-of-time pile-up measured through the average number of interactions per bunch crossing $<\mu>$, as a function of the jet η. This dependence is shown for three cases: before pile-up correction, after the pile-up energy density correction and after the residual correction. A simulated sample of di-jet events, calibrated at the EM+LCW scale, from pp collisions at $\sqrt{s} = 8$ TeV corresponding to an integrated luminosity of 20.3 fb^{-1}is used. Taken from [45].

cells signal. The effect of out-of-time pile-up is much smaller. Nevertheless, an additional correction derived from simulation is used to compensate for its residual effects [45].

Figure 4.15 shows the dependence of the reconstructed jet p_{T} on both the in-time pile-up, measured by the number of primary vertices in an event, and on the out-of-time pile-up, proportional to the average number of interactions per bunch crossing $<\mu>$. Pile-up affects more low-p_{T} jets, and for that reason only jets within the 20 to 30 GeV p_{T} interval are included in the plots. Jets at the EM+LCW scale, i.e. calibrated using the local cluster weighting calibration method, are shown. The jet p_{T} dependence on pile-up is shown before the correction, after the pile-up density correction and after the correction of the residual effects caused by out-of-time pile-up. The pile-up density correction is effective in eliminating the jet p_{T} dependence on $<\mu>$, that on average increases the jet energy in $\sim$0.4 GeV per additional interaction in the event. It is however incapable of eliminating the effects due to out-of-time pile-up, more pronounced for forward jets. When both corrections are applied, the jet energy measurement is independent of pile-up.

Origin correction

Following pile-up correction, the jet origin is corrected. The direction of the calorimeter jets is determined using the ATLAS nominal interaction point as reference but the actual primary vertex (PV) where the jet had origin is usually displaced from this point. In order to properly determine the origin of the jet, the momentum of each topo-cluster forming the calorimeter jet is corrected such that it points back to the main PV of the event. Afterwards, the jet momentum

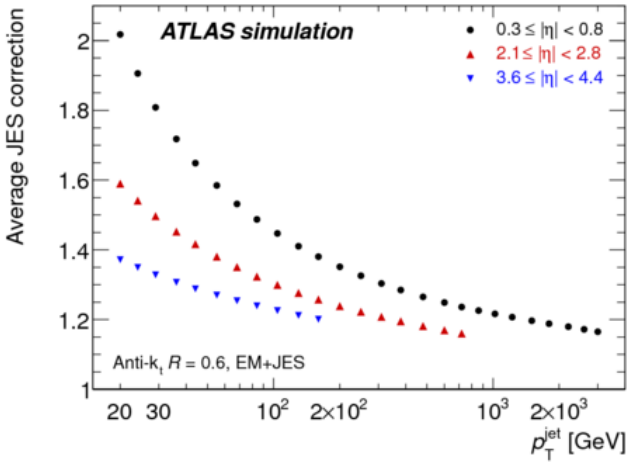


Figure 4.16: Average JES calibration factor as a function of the jet p_T for various intervals of the jet η. Both taken from [42].

is redefined by the vector sum of the corrected momentum of the topo-clusters constituting the jet. This correction improves angular resolution while leaves the jet energy unaffected [42].

Jet Energy Scale Calibration

After origin correction, an energy calibration technique brings the jet energy from the EM scale to the jet energy scale (JES) [42].

The aim is mainly to correct for the non-compensating nature of the hadronic energy measurement. It was derived from MC samples of isolated jets, by comparing the measured jet energy with the MC truth jet energy. The truth jets used to determine this calibration are composed of simulated truth particles, excluding muons and neutrinos, because the calibration is designed to correct only the non-compensated measurement of the hadrons energy. The spatial matching between calorimeter and truth jets is done searching for the closest calorimeter and truth jet pair and allows the determination of the energy response, defined as the ratio between the calorimeter jet energy measured at the EM scale and the truth jet energy E^{EM}/E^{truth}. The average energy response is calculated as a function of the calorimeter jet η and energy to determine a calibration factor. That is applied to the calorimeter jets to restore on average the truth jet energy scale. Figure 4.16 shows the average JES calibration factor as a function of the calorimeter jet p_T for various intervals of η. The correction factor decreases with increasing jet p_T and and increases with jet centrality reaching a value of 2 for central low p_T jets [42].

Global Sequential Calibration

A good jet energy resolution is fundamental for the $H \to b\bar{b}$ search since the Higgs candidate corresponds to a di-jet system and its invariant mass - $m_{b\bar{b}}$ - to its mass. $m_{b\bar{b}}$ is one of the variables that most effectively discriminates signal and background events in this analysis, and the better the jet energy resolution the narrower is the signal peak and the better

the resolution on the Higgs mass. The global sequential calibration (GSC) is a technique designed specifically to improve the jet energy resolution and therefore has an essential role in the maximisation of the $H \to b\bar{b}$ signal sensitivity.

GSC follows the JES calibration and explores correlations between the jet energy response and detector observables, to address issues of response dependence on the jet flavour [46]. Several measurements from the calorimeters but also from the inner detector and muon spectrometer are used for this purpose. The topology of energy deposits in the calorimeter proved to be useful. A larger fraction of energy deposits in the hadronic calorimeter indicates more hadrons in the jet, while a larger energy fraction in the first layers of the EM indicates that the shower initiated before the calorimeter. Both cases correspond to a lower calorimeter response. On the other hand, tracking information is used to characterise the particle content of the shower: gluon initiated jets tend to have more particles that are also softer resulting in a lower calorimeter response with respect to light jets.

Figure 4.17(a) shows how the GSC uses the correlation between the jet energy response and the jet track width. This observable is defined as

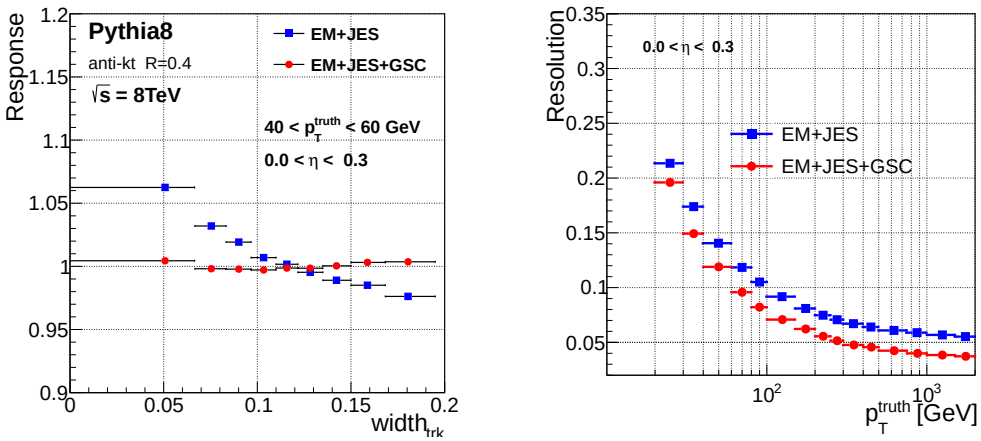
$$width_{trk} = \frac{\sum_i p_{\mathrm{T}}^i \Delta R_{i,jet}}{\sum_i p_{\mathrm{T}}^i} \tag{4.2}$$

where the sum runs over every track i pointing to the calorimeter jet, p_{T}^i is the transverse momentum of the ith track and $\Delta R_{i,jet}$ is the radial distance between that track and the jet axis. A jet with large track width is more affected by out-of-cluster energy deposits resulting in a poor response of the calorimeter as seen in Figure 4.17(a). Since the JES calibration corrected the response on average, the opposite effect is seen for narrow width jets. The GS calibration corrects for both effects leaving the average response unaffected.

The GSC method employs a sequence of corrections as a function of the different observables, defined such that the average jet energy scale remains invariant at each step. By eliminating the response deviations, it improves the jet energy resolution up to 35% with respect to the JES resolution as Figure 4.17(b) shows, and reduces the calorimeter dependence on jet flavour [46].

In-situ correction

Up to this point, all the calibrations were derived using MC information exclusively. The residual difference in the calorimeter response between simulated and real data is corrected in a last calibration step that is only applied to data. The calibration is derived from data and MC comparisons using in situ transverse momentum balance. It explores the p_{T} balance between the jet and a well measured reference object in events of well defined topology. For central jets with $p_{\mathrm{T}} < 800$ GeV, photons or the di-lepton system from Z bosons decays are used as reference objects. For larger transverse momentum, multijet topologies provide a system of already calibrated low p_{T} jets recoiling against the high momentum jet for which the calibration



(a) (b)

Figure 4.17: (a) Calorimeter response for simulated jets calibrated at the EM+JES scale as a function of the jet track width before and after GSC. (b) Jet energy resolution as a function of the truth jet p_{T} at the EM+JES and EM+JES+GSC scale, for jets within the pseudorapidity interval of $0 < \eta < 0.3$. Adapted from [47].

is under determination. Di-jet events are used to calibrate forward jets with one calibrated central jet used as reference for the p_{T} balance.

4.5 $b-$Tagging

Being able to determine whether a jet originates from a $b-$quark, referred to as $b-$tagging, is crucial to select signal events in the $WH \to \ell\nu b\bar{b}$ search, and its impact on signal related measurements is therefore evident. On the other hand, b-tagging methods aim to efficiently reject jets with different flavour origin such as c-jets or light-jets. This leads to better performance in what comes to background rejection and in the case of this analysis is crucial to suppress backgrounds coming from W plus $c-$ or light-jets production. In this way, an efficient b-tagging technique has a leading role in enhancing the signal sensitivity by contributing both to the signal detection and background rejection.

4.5.1 Jet Flavour Properties

The quark and gluon fragmentation results in a shower of hadrons. At the early stage of the hadronic shower development, a single hadron carries the original quark and a substantial part of its momentum. As gluons do not form hadrons, in jets from gluons this role is more often played by a light quark resulting from gluon splitting. The fundamental physical differences between $b-$mesons and $c-$ and light mesons can be then utilised to provide the basics for jet flavour discrimination. When compared to light mesons such as pions and kaons, b and c mesons are short-lived particles. Most kaons and electrically charged pions have a mean lifetime of the order of 10^{-8} s and reach flight lengths of 7 and 3 m, respectively, before decaying. b and

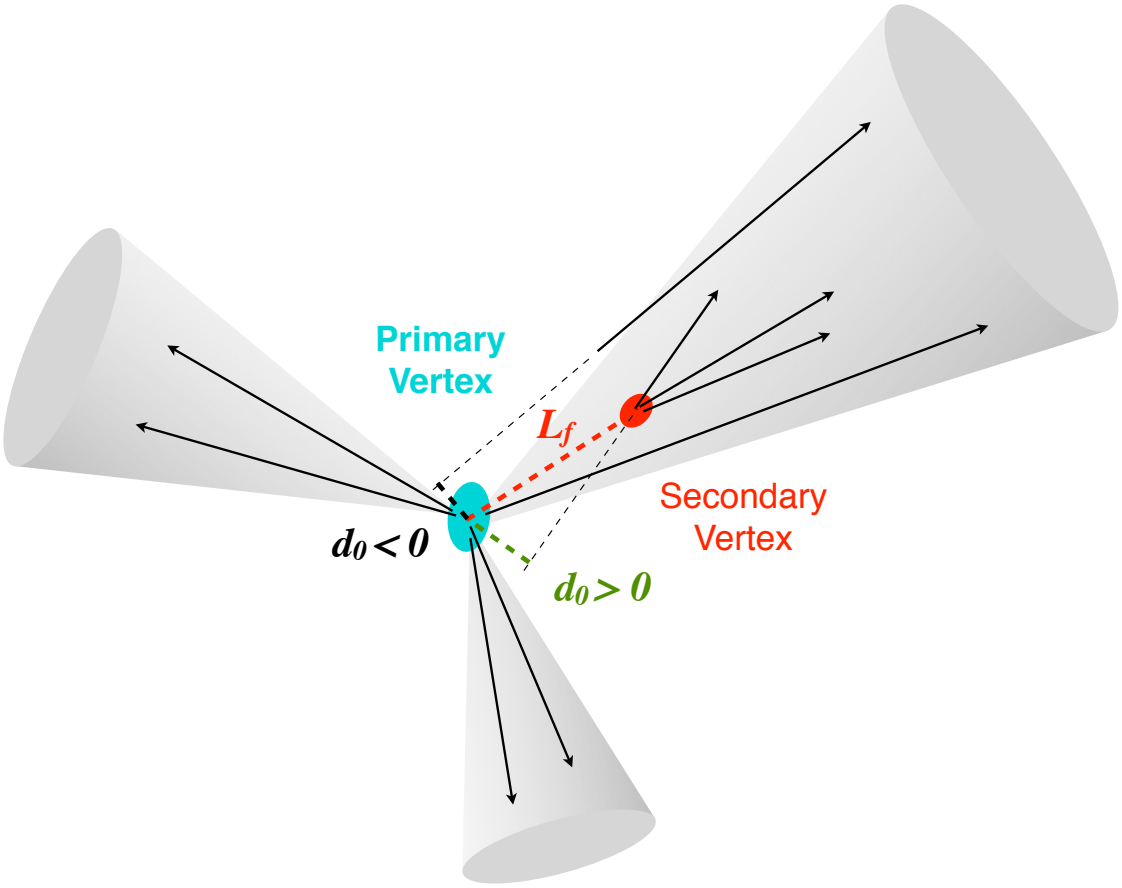


Figure 4.18: Schematic view of a three-jet event with one primary vertex and a secondary vertex. Oppositely signed transverse impact parameters d_0 for two displaced tracks.

c mesons have mean lives of about 10^{-12} s and 10^{-15} s, respectively, which is sufficient for a relativistic $b-$ or $c-$meson to travel a few mm before decaying. In the pp collisions detected by ATLAS, a $b-$meson resulting from a $b-$quark produced at a hard-scatter vertex decays before the ID first layer. The neutral pion on its turn, having a mean lifetime of $\sim 8.5 \times 10^{-17}$ s, decays immediately after being produced and does not have a flight path length nearly as significant as b or c hadrons. The $b-$meson decay cascade originating a muon, at a rate of about 20%, can be used complementary to discriminate b-jets for it provides the clear muon signature in the MS.

4.5.2 $b-$Tagging Algorithms

There are several algorithms for b-tagging [48] to explore the distinctive signatures of these physical properties in the detector: lifetime-based algorithms and muon-based algorithms. Jets are reconstructed with the calorimeter but, except for the muon-based tagging algorithms where the use of the MS is indispensable, b-tagging relies mostly on the ID information and on the reconstruction resolution of tracks and vertices.

Lifetime-based b-taggers are further split into two categories: based on impact parameters or on secondary vertices. That can be understood from Figure 4.18: the secondary vertex (SV) is reconstructed from jet tracks displaced from the primary vertex (PV). The track impact

parameter is defined as the distance of minimum approach between the track extrapolation and the PV. The transverse and longitudinal projections of the vector separating the impact parameter of the track and the PV, d_0 and z_0 respectively, are typically considered. To enhance $b-$ and light-jets discrimination, d_0 is attributed a positive sign if the extrapolated track intersects the jet axis in front of the PV and a negative sign if the intersection occurs behind the PV.

Impact Parameter-based Algorithms

The JetProb and IP3D are two examples of impact parameter-based algorithms used in the ATLAS experiment [48].

JetProb The JetProb uses the signed impact parameter significance d_0/σ_{d_0} to distinguish between tracks coming from the PV and tracks displaced from the PV, produced on a B hadron decay. σ_{d_0} is the uncertainty of the d_0 measurement. The d_0 and d_0/σ_{d_0} distributions are shown in Figure 4.19. Tracks from b or c hadrons tend to have larger and positive impact parameters while tracks from light-jets jets do not present any asymmetry. JetProb exploits these distributions to extract a probability for a jet to be a light or a heavy flavour jet. By comparing the d_0/σ_{d_0} of each track of the jet with a pre-determined probability function obtained from data for prompt tracks, the probability of the track to have origin in the PV is measured. The individual track probabilities are then combined into the probability for the jet to be a light or a heavy flavour jet. The method is independent of simulation since all its inputs can be extracted from data, and that constitutes its main advantage.

IP3D On the other hand, the I3PD is a more efficient algorithm, essentially because it also takes advantage of the longitudinal impact parameter significance z_0/σ_{z_0}, obtaining a three-dimensional measurement of the track displacement to the PV. The algorithm implements a log-likelihood ratio (LLR) that compares the measured impact parameters of tracks with two-dimensional PDFs of the $b-$ and light-jet hypothesis obtained from simulation. In this manner, the I3PD also profits from z_0 and d_0 correlations.

Secondary Vertex-based Algorithms

Secondary vertex-based algorithms, such as SV0, SV1 or JetFitter, explicitly attempt to reconstruct the secondary vertex [48]. One of the weaknesses of these algorithms lies precisely on the efficiency of reconstructing the SV of approximately 70% [48]. The SV reconstruction, for each jet, uses tracks within the jet that are significantly displaced from the PV that best matches that jet, to form a vertex through a χ^2 fit technique. Long life-time hadrons decays, for instance, K_s or Λ^0 with $\tau \sim 10^{-10}$ s, and products of photon conversions can originate background SVs. Since b hadrons are heavier, these backgrounds are excluded by imposing a

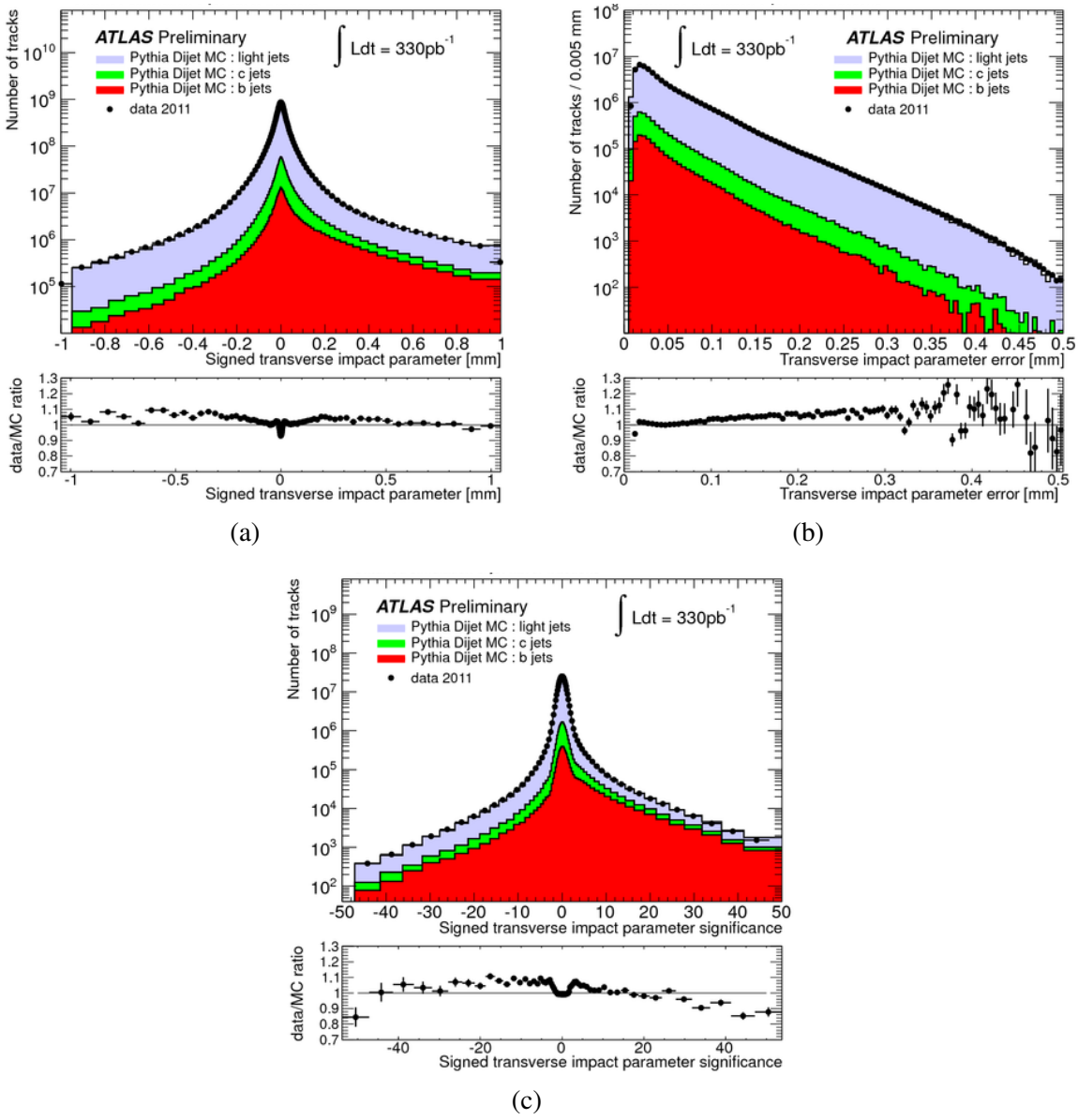


Figure 4.19: Data and simulated distributions of the (a) transverse impact parameter d_0, (b) its uncertainty σ_{d_0} and (c) significance d_0/σ_{d_0} for displaced tracks inside simulated light-, $c-$ and $b-$jets. Taken from [49].

minimum threshold on the invariant mass of the system composed of the tracks associated with the formed SV, M^{vtx}_{track}.

SV0 Having encountered a signal SV, the SV0 method uses the significance of the flight length measurement L_f/σ_{L_f}, where L_f is the distance between the primary and secondary vertices and σ_{L_f} its uncertainty, as the final $b-$tag discriminant to distinguish between b and light-jets. The flight length is signed in a manner similar to d_0 to further increase the separation between light and $b-$jets.

SV1 SV1 has a better performance than SV0 since in addition to the L_f significance, it takes advantage of other four variables and uses the LLR technique to provide the final output discriminant. The additional input variables to SV1 are the $\Delta R(axis_{jet}, axis_f)$ between the jet and the flight path axis; M^{vtx}_{track}, as defined previously; the number of two-track vertices, $N_{vtx} : n_{track} = 2$; and the fraction of the sum of the energies of these tracks to the sum of the energies of all tracks in the jet, E_f. For a $b-$jet, $\Delta R(axis_{jet}, axis_f)$ should be low and E_f close to the unity, since most of the $b-$quark momentum is carried by the $b-$hadron and transferred to the full jet itself with the shower development. M^{vtx}_{track} is compatible with the b hadrons masses in the case of $b-$jets. More than relying on the separation capabilities of these variables in an isolated fashion, the LLR benefits from their correlations and makes SV1 a more efficient $b-$tagger.

JetFitter The decay of b mesons into other hadrons results primarily in at least one $c-$hadron given the $b,c-$quark mixing magnitude $|V_{cb}|$, and the decay to top-quark suppression by kinematic arguments. In turn, c hadrons decays result in flight path lengths similar to b hadrons and additional SVs can be reconstructed. The JetFitter method implements an artificial Neural Network (NN) multivariate method to search for this decay chain, receiving as inputs variables that can describe this topology: the number of vertices with at least two associated tracks, $N_{vtx} : n_{track} \geq 2$; the total number of tracks associated with these vertices, $\sum n_{track} : n^{vtx}_{track} \geq 2$; and the number of additional single track vertices on the $b-$hadron flight path $N^{1track}_{vtx} : n^{f_{axis}}_{track} = 1$. As for SV1, the vertex-related quantities M^{vtx}_{track}, L_f significance and E_f are included.

Combined $b-$Tagging Algorithms

MV1 is a combined algorithm based on a NN that uses as inputs the output weights of SV1, IP3D, and IP3D+JetFitter. Figure 4.20 shows the schematic diagram of MV1 and all its inputs. The IP3D and JetFitter combination is done by using the I3PD output as an additional input to the JetFitter NN. The distributions of the inputs to the MV1 NN are shown in Figure 4.21 for $b-$, $c-$ and light-jets. They all constitute good discriminants of $b-$jets. Furthermore, it was shown that their correlations differ between different flavour jets [48]. The MV1 NN uses this information to produce a single final discriminant, ranging from 0 to 1, where a single cut can be applied to identify a jet as resulting from $b-$quark fragmentation. As can be seen from Figure 4.22, MV1 yields values closer to 1 for signal $b-$jets and closer to 0 for the other cases.

In the $WH \to \ell\nu b\bar{b}$ analysis, jets are identified as $b-$jets using the MV1c algorithm. This algorithm uses the same scheme as MV1, differing only from the usage of the JetFitterC instead of the JetFitter. The JetFitterC is a version of the JetFitter trained specifically to distinguish $c-$jets. As a result, MV1c is more efficient in rejecting $c-$flavoured jets than MV1. For a 70% $b-$jet efficiency point, the c-jet rejection factor is 1.9 times larger than with the MV1 algorithm [50].

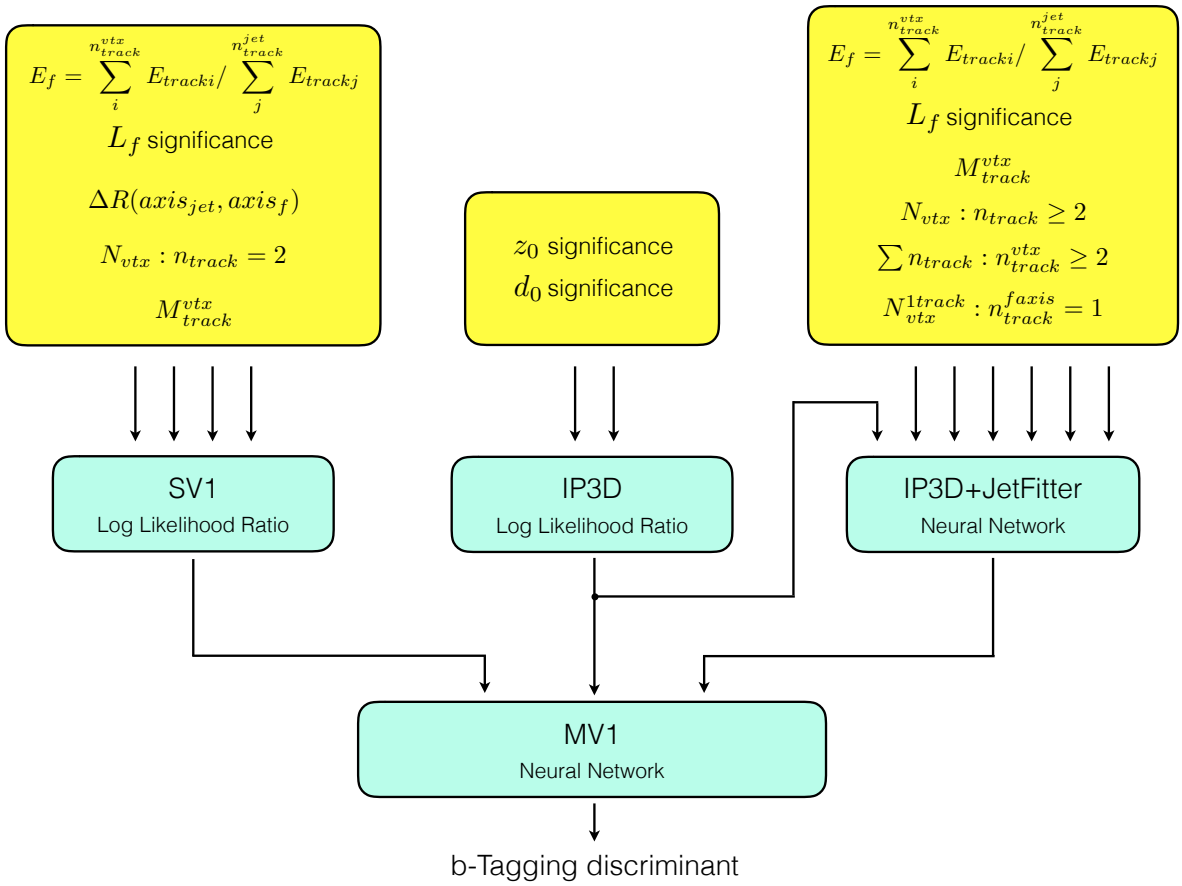


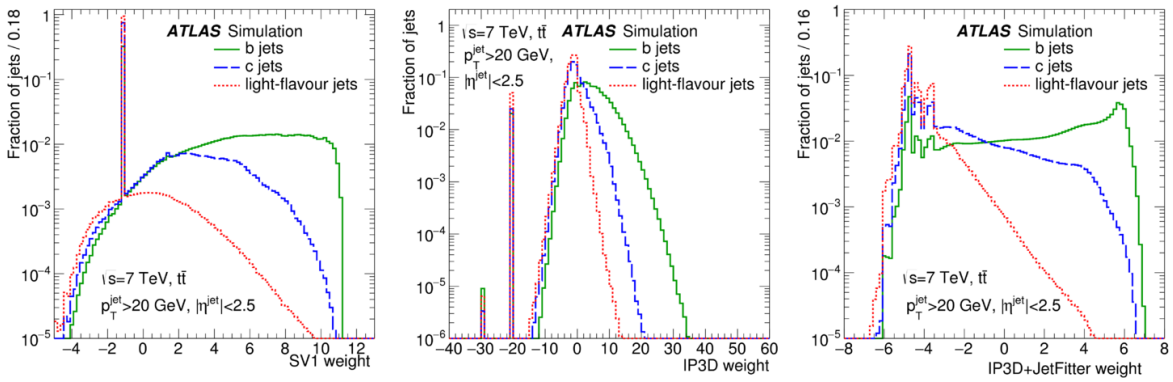
Figure 4.20: Diagram view of the MV1 b-tagging algorithm.

Performance of the $b-$Tagging Algorithms

The performance of the algorithms described can be evaluated from Figure 4.23, where the light-jet rejection factor is defined as the inverse of the mis-tag rate. For the same signal efficiency, the larger the rejection the better, for it means that less light-jets are being mis-tagged as $b-$jets. A more pronounced slope on the rejection versus efficiency curve corresponds to a better performance. JetFitter is the best performant algorithm because it uses several jet flavour discriminants and makes use of the sophisticated NN analysis. I3PD performs better than SV1, as expected from the limited SV reconstruction efficiency, and JetProb, the simplest tagging method, has the worst performance. But the main achievement results from the combination of different algorithms as in the case of IP3D+SV1, IP3D+JetFitter or MV1, all performing better than their components alone.

4.5.3 $b-$Tagging Calibration

In order to achieve the same $b-$tagging performance for MC and data, the algorithm is calibrated [48]. To do so, the efficiency for $b-$ and $c-$jets and mis-tag rate of light jets is measured for data and MC in reference samples enriched in each jet flavour. The ratio between the number of jets in the sample after and before applying $b-$tagging corresponds



(a) (b) (c)

Figure 4.21: Distributions of the output (weight) of the (a) SV1, (b) IP3D and (c) IP3D+JetFitter $b-$tagging algorithms for simulated light-, c- and $b-$jets. Taken from [48].

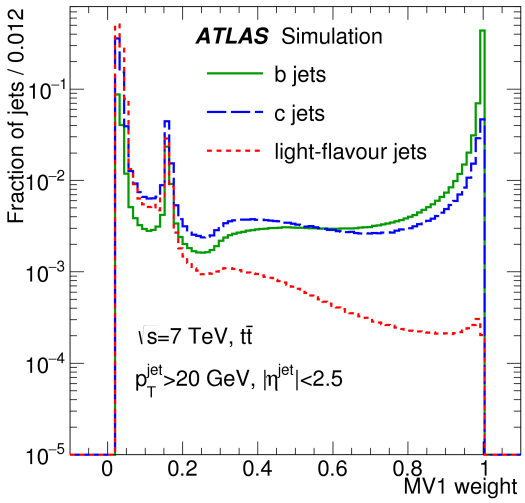


Figure 4.22: Distribution of the MV1 output (weight) for simulated light-, $c-$ and $b-$jets. Taken from [48].

to the algorithm efficiency. The fraction of other flavour jets composing the enriched sample is determined by MC and must be subtracted for this procedure. The determined efficiencies are compared for data and MC and the arising differences are used to derive data-to-MC calibration scale factors that are applied to MC as event weights. To obtain samples enriched in each jet flavour, specific event topologies and selection methods independent of $b-$tagging are considered.

$b-$jets

Event samples enriched in $b-$jets are obtained from:

- Di-jet events with one jet containing a muon, taking advantage of the $b-$ hadrons decay to muons;
- Top-quark pair events with one or two leptons from the W bosons decay, profiting from

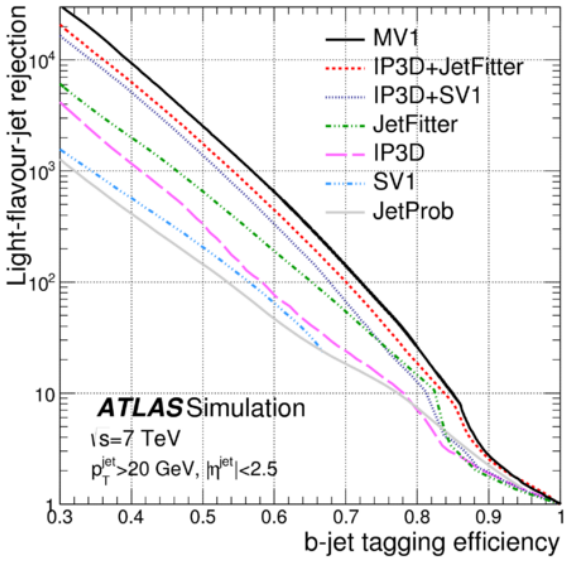


Figure 4.23: Light-jet rejection as a function of b-jet efficiency for the different b-tagging algorithms. Taken from [48].

the dominant $t \to Wb$ decay in the SM. In the one lepton case, the invariant mass of the products of the hadronic decay of the other W, $W \to q'\bar{q}$, is requested to be within the W mass window to reject background.

$c-$jets

Event samples enriched in $c-$jets are obtained from:

- W plus a $c-$jet with a soft muon from $c-$hadron decay, with the W decaying via the electron channel. The main production mechanisms of $W+c-$jet from pp collisions are $gs \to W^- c$ and $g\bar{s} \to W^+ \bar{c}$. Since the soft muon and the $c-$quark have same sign electrical charges, requiring that the muon and the electron from W decay have opposite sign charges yields a high pure $W+c-$jet sample. Background events result equally in same-sign (SS) and opposite-sign (OS) electron and muon. This feature is exploited to obtain the number of $W+c-$jet events from the difference of OS and SS events.

- Di-jet events reconstructing the $c-$meson decay chain $D*^+ \to D^0\pi^+ \to K^-\pi^-\pi^+$, by selecting two oppositely charged tracks as the D^0 candidate, and events within the D^0 mass window.

Light jets

Since light jets pass the $b-$tagging criteria mainly due to resolution effects on track and vertex reconstruction (d_0 has the same probability to be positively or negatively signed), they are selected from an inclusive sample of jets using an inverse tagging method (version of MV1c obtained from reversing the sign of d_0 and f_L significance parameters) to reject the $b-$ and $c-$flavour.

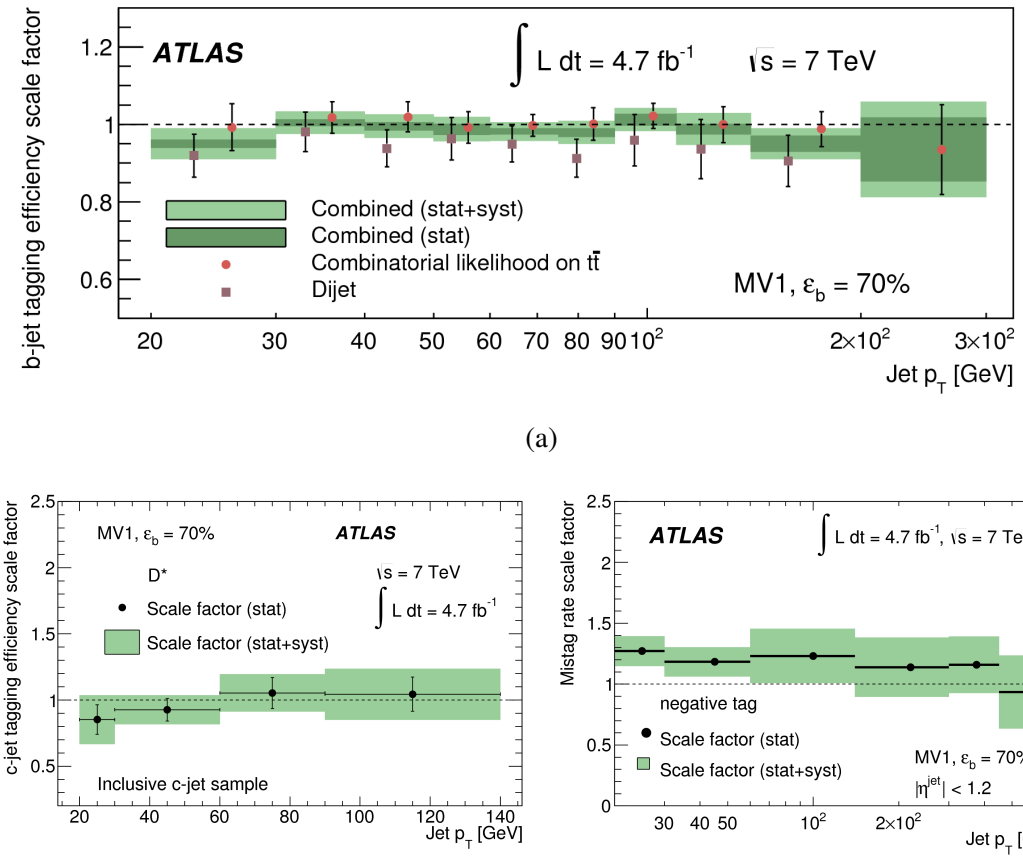


Figure 4.24: Data to MC calibration scale factors for the MV1 $b-$tagging algorithm at 70% efficiency working point as a function of jet p_{T} for (a) $b-$jets, (b) $c-$jets and (d) light-jets. Taken from [48].

As a final step, the $b-$tagging correction factors resultant from the different methods described above are statistically combined to obtain the best precision. For a MV1 operating point of 70% efficiency for $b-$jets, these are shown in Figure 4.24 for the different jet flavours. As can be seen, the resultant $b-$tagging scale factor for $b-$jets result from the combination of the scale factors obtained with the di-jet sample and with the $t\bar{t}$ sample. The calibration factors are parametrised as function of jet p_{T} and η and for $b-$ and $c-$jets. They do not differ more than 5% and 15% from the unity, respectively, while for light-jets this difference can be as large as 30%.

For the MV1c algorithm used in the analysis, these correction factors are similar and the procedure used for their determination is the same. The calibrations were derived for six operating or working points based on the cut value applied to the MV1c discriminant, each corresponding to a specific $b-$jet efficiency - 100%, 80%, 70%, 60%, 50% and 0% - and corresponding $c-$ and light-flavour rejection factors.

The final systematic uncertainty of the $b-$tagging calibration factors, displayed on the plots of Figure 4.24, have several sources. Among them, the most important are the statistics of the simulated samples used to perform the measurement, the quark fragmentation models, the

amount of initial and final state radiation, the jet energy scale and resolution uncertainties, and the uncertainty on the flavour composition of the jets samples.

4.6 Missing Transverse Energy

Missing transverse energy[1], $E_{\mathrm{T}}^{\mathrm{miss}}$, is obtained from the momentum imbalance in the plane transverse to the beam of the collider particles. In pp collisions, the projection of the momentum of the interacting partons along the beam axis in the laboratory frame is unknown before the collision. However, its projection on the plane transverse to the nominal beam line is null to a very good approximation. So, after the parton collision, the vector sum of the transverse momentum of all final state particles must be null due to momentum conservation. If a non-zero value is obtained it must be due to undetected particles.

Detector resolution, volume acceptance and insensitive regions lead to some amount of fake $E_{\mathrm{T}}^{\mathrm{miss}}$. In the SM context, real $E_{\mathrm{T}}^{\mathrm{miss}}$ arise for events where final state neutrinos are produced. These particles only interact via weak force and escape the detector without depositing their energy, so a large value of $E_{\mathrm{T}}^{\mathrm{miss}}$ in an event is usually assigned to the neutrino presence. The $E_{\mathrm{T}}^{\mathrm{miss}}$ reconstruction is then fundamental to infer the neutrino presence in the context of SM physics. In BSM physics scenarios, the $E_{\mathrm{T}}^{\mathrm{miss}}$ can be a sign of a new particle predicted by new physics models, that is not detected. The $E_{\mathrm{T}}^{\mathrm{miss}}$ importance extends, for instance, to dark matter searches with weakly interacting particles as candidates.

4.6.1 $E_{\mathrm{T}}^{\mathrm{miss}}$ Reconstruction

The $E_{\mathrm{T}}^{\mathrm{miss}}$ measurement carried out in the ATLAS experiment depends mainly on energy deposits in the calorimeters and muons momentum measurements with the MS [51]. Only in special cases the ID information is involved. The $E_{\mathrm{T}}^{\mathrm{miss}}$ is determined from the E^{miss} $x-$ and $y-$components, respectively E_x^{miss} and E_y^{miss}, as follows:

$$E_{\mathrm{T}}^{\mathrm{miss}} = \sqrt{(E_x^{\mathrm{miss}})^2 + (E_y^{\mathrm{miss}})^2} \tag{4.3}$$

where the E_x^{miss} and E_y^{miss} components are calculated by:

$$E_{x(y)}^{\mathrm{miss}} = E_{x(y)}^{\mathrm{miss,e}} + E_{x(y)}^{\mathrm{miss},\gamma} + E_{x(y)}^{\mathrm{miss},\tau_{\mathrm{had}}} + E_{x(y)}^{\mathrm{miss,jets}} + p_{x(y)}^{\mathrm{miss},\mu} + E_{x(y)}^{\mathrm{miss,SoftTerm}} \tag{4.4}$$

The various terms in 4.4 represent the total energy associated with each final state object type in the event as follows:

- $E_{x(y)}^{\mathrm{miss,e}}$, $E_{x(y)}^{\mathrm{miss},\gamma}$: total energy associated with the electrons and photons;
- $E_{x(y)}^{\mathrm{miss},\tau_{\mathrm{had}}}$: total energy associated with the hadronically decaying $\tau-$leptons;

[1] In this description energy and momentum appear as equivalent often. This equivalence is only valid in the assumption of $E_{\mathrm{T}}^{\mathrm{miss}}$ being associated with very light particles, i.e. in the $|\mathbf{p}| \to E$ limit.

- $E_{x(y)}^{\text{miss,jets}}$: total energy associated with jets of $p_{\text{T}} > 20$ GeV;

- $p_{x(y)}^{\text{miss},\mu}$: total momentum associated to muons, where energy is used instead in the case of calorimeter muons;

- $E_{x(y)}^{\text{miss,SoftTerm}}$: total energy of topological clusters (or momentum of particle tracks not matching any cluster in the calorimeter) not associated with any object reconstructed previously. The contribution from soft jets of $p_{\text{T}} < 20$ GeV is also taken into account in this term.

Each term $E_{x(y)}^{\text{miss,type}}$ is therefore the energy sum for all objects of a given type:

$$\begin{aligned} E_x^{\text{miss,type}} &= -\sum_{i=1}^{N_{\text{type}}} E_i \sin\theta_i \cos\phi_i \\ E_y^{\text{miss,type}} &= -\sum_{i=1}^{N_{\text{type}}} E_i \sin\theta_i \sin\phi_i \end{aligned} \tag{4.5}$$

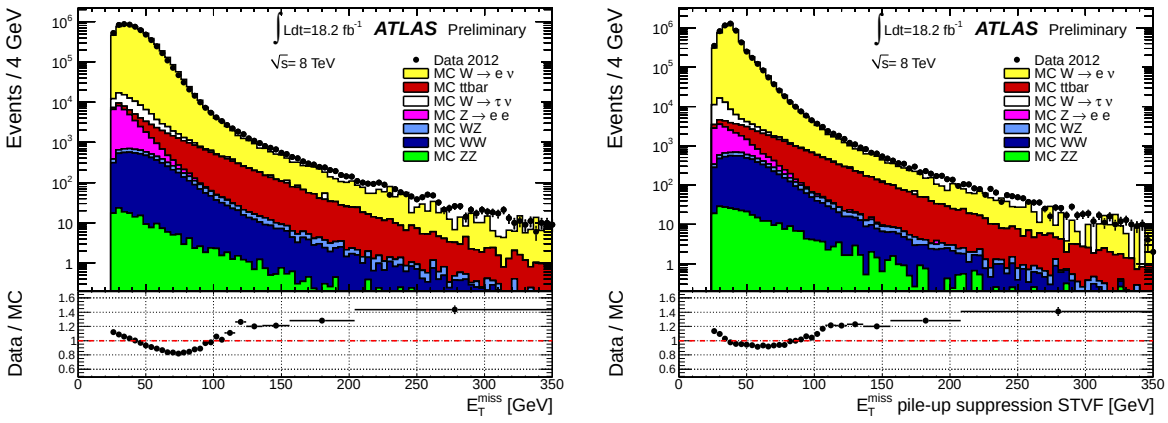
where E_i, θ_i and ϕ_i are the calibrated energy of the object, polar angle and azimuthal angle, respectively. Calorimeter noise suppression is guaranteed by using only reconstructed objects made of clusters for which an energy-to-noise ratio threshold is applied at the cell level in the clustering phase. In order to resolve overlap between objects and avoid energy multiple counting, the calorimeter energy deposits are attributed to the final state objects in the following order: electrons, photons, hadronically decaying $\tau-$leptons, jets, and muons. Energy clusters not associated with any reconstructed object are assigned to the $E_{\text{T}}^{\text{miss}}$ soft term and the contribution of low energy particles that do not reach the calorimeter is recovered by using the p_{T} of tracks not matching any energy cluster in the calorimeter.

4.6.2 $E_{\text{T}}^{\text{miss}}$ Scale and Resolution

One of the final state particles of this analysis is a neutrino, produced on the W boson leptonic decay, so $E_{\text{T}}^{\text{miss}}$ is used to select signal events. This makes the analysis very relying on the precision of the $E_{\text{T}}^{\text{miss}}$ measurement.

Pile-up interactions cause $E_{\text{T}}^{\text{miss}}$ scale and resolution degradation and performance losses, so methods are employed to suppress pile-up effects on the $E_{\text{T}}^{\text{miss}}$ reconstruction [51]. The jet and soft terms contributions are more affected by pile-up interactions than other terms, since hadrons are often produced in pp interactions. For these reasons, the $E_{\text{T}}^{\text{miss}}$ pile-up suppression methods involve exclusively these two terms and consist of the following:

- $E_{x(y)}^{\text{miss,jets}}$ is determined excluding jets of $p_{\text{T}} < 50$ GeV laying within the ID pseudorapidity coverage, $|\eta| < 2.4$, for which none of the matching tracks come from the main PV.



(a) (b)

Figure 4.25: $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution of data and MC $W \to e\nu$ events (a) before and (b) after applying pile-up suppression. Taken from [51].

- $E_{x(y)}^{\mathrm{miss,SoftTerm}}$ is scaled by the fraction of momenta of tracks associated to the soft term that come from the main PV, denominated soft term vertex fraction STVF and defined as $\mathrm{STVF} = \sum_{\mathrm{tracks}_{\mathrm{SoftTerm,PV}}} p_{\mathrm{T}} / \sum_{\mathrm{tracks}_{\mathrm{SoftTerm}}} p_{\mathrm{T}}$

Figure 4.25 shows the $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution for $W \to e\nu$ events from simulated and collision data, before and after applying the pile-up suppression techniques explained above. The data and MC agreement is improved when the pile-up effects are suppressed. For the spectrum region within 40 and 100 GeV, the discrepancies do not overcome 10%. Nevertheless, important differences are still observed. In the interval between 25 to 40 GeV, these are due to the absence of the multi-jet events simulation on the prediction side, that tend to concentrate at lower values as no real $E_{\mathrm{T}}^{\mathrm{miss}}$ is expected in such cases.

Chapter 5

Calibration and Data Quality of the TileCal

This Chapter presents a discussion about the key aspects about the performance and operation of TileCal, the hadronic calorimeter of ATLAS, that directly impact the quality of the measurements of hadrons, jets and $E_{\mathrm{T}}^{\mathrm{miss}}$ done by ATLAS.

First, the details about the TileCal readout scheme and architecture are given in Section 5.1 and the systems and strategies used to calibrate it are described in Section 5.2.

The methods used to analyse the data from the laser calibration system are described in Section 5.3. Then, a dedicated method developed to automatically identify channels with some kind of mis-functioning is presented here in Section 5.4. Variables derived from data of the laser calibration system are studied with the purpose of discriminating these channels from the ones with normal response.

5.1 The ATLAS Tile Calorimeter

The ATLAS Tile Calorimeter [52] was briefly described in Section 3.2.2. It covers the most central region of the ATLAS detector of $|\eta| < 1.7$, using iron as absorber and plastic scintillator tiles as active medium. This Section describes in more detail the segmentation of TileCal, the readout electronics and presents the algorithm used to reconstruct the energy deposit in the cells of the detector.

5.1.1 Architecture

The TileCal structure comprehends an innermost long barrel (LB), totally covering the region $|\eta| < 1.0$, and two extended barrels (EB) for the $0.8 < |\eta| < 1.7$ coverage. The barrels, with a radial size of 7.4λ [1], are divided in the azimuthal direction into 64 wedged modules of size $\Delta\phi \sim 0.1$, and are radially segmented in three layers (A, BC, and D) as drawn in Figure 5.1.

[1] λ corresponds to one interaction length defined as the mean free path of particles between two inelastic interactions.

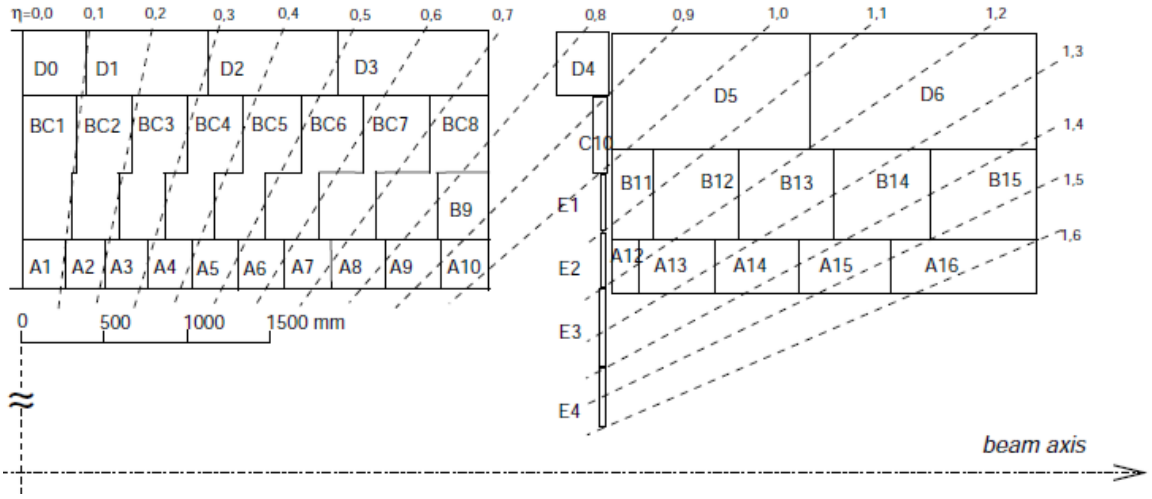


Figure 5.1: Segmentation in depth and η of the Tile Calorimeter modules in the (left) long and (right) extended barrels. Taken from [52].

Special E cells (with no iron) are placed between the barrel and long barrel in order to cover the gap region between them. The three layers are segmented in η and the resulting unit volume defines the TileCal detection cell.

Each module periodically assembles layers of iron and scintillator tiles and the cell is formed by grouping several tiles at the readout level as shows Figure 5.2. Wavelength shifting optical fibres are coupled to each side of the tiles, collecting the scintillation light and transmitting it to two Photomultiplier Tubes (PMTs), each associated to one side of the tiles. In this way, cells have two readout channels providing the redundancy needed in case of failure of a readout channel. This also prevents non-uniformity in the response to particles entering the scintillator at different azimuths since light attenuation in the tiles can be as much as 40% [52].

5.1.2 Readout Electronics

The PMTs are lodged in drawers at the outer radius of the modules, that also contain the readout electronics and low and high voltage power supplies. The front-end electronics consists of a small printed circuit board per channel with two pulse amplifiers, with high gain (HG) and low gain (LG), and a 10 bit ADC. The HG and LG amplified signals are sampled and digitised by the ADC every 25 ns, i.e. at the frequency of the LHC nominal bunch-crossing. When a global trigger signal reaches back the module electronics, seven samples are kept to be read out. A switcher, actioned by the saturation of the HG signal, determines whether the HG or LG is passed on to the remaining readout chain.

The drawers also contain adder boards to sum the analogue signals of all the cells of same η. The summed signal corresponds to the $\Delta\eta \times \Delta\phi = 0.1 \times 0.1$ trigger towers, described in Section 3.2.4, that are used by the L1 Calo trigger. Overall, TileCal has 5182 cells, 9852 PMTs and 19704 ADCs, and provides 2080 trigger signals to the ATLAS first trigger level.

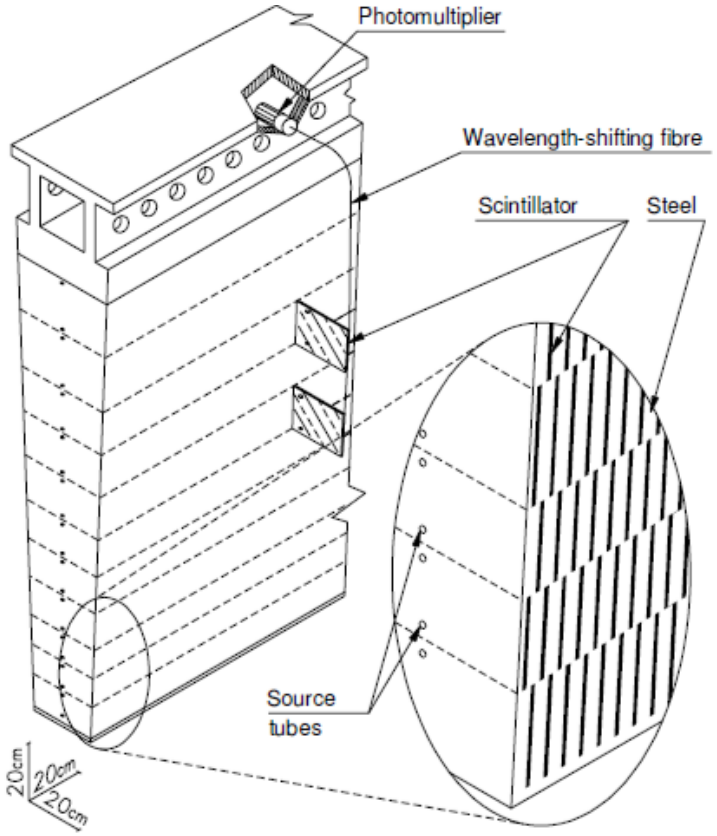


Figure 5.2: Schematic view of the optical readout of the Tile Calorimeter. Taken from [52].

5.1.3 Energy Reconstruction

The reconstruction of the energy deposited in a given TileCal cell begins by determining the signal amplitude A from the seven digitised samples S_i of each of the two PMTs pulses. This is done using the Optimal Filter (OF) algorithm employing the following weighted sum

$$A = \sum_{i=1}^{7} a_i S_i \tag{5.1}$$

where a_i are the OF weight coefficients derived from a reference PMT pulse shape. The weights take into account the noise correlation with the given sample.

Since a reference pulse is used, the ADC timing is a key aspect of the amplitude reconstruction. In that sense, the sampling timing can be adjusted in multiples of $\sim$0.1 ns such that the central sample matches the PMT pulse peak, for instance to take into account the different time of flight of particles arriving at cells located at different radius. The adjustment is configured through a time offset database that can be regularly updated. This hardware timing calibration acts simultaneously on a set of six channels and some residual time offset affects the amplitude at single channel level. These are later corrected by software at the high-level trigger and at offline reconstruction phase.

The conversion of amplitude in energy involves many conversion factors, and most of them can be provided in a regular basis by the calibration systems to account for the drift of the

detector response over time:

$$E_{channel} = A \times C_{ADC \to pC} \times C_{pC \to GeV} \times C_{Cs} \times C_{Laser} \tag{5.2}$$

where A is the signal amplitude in ADC counts and the factor $C_{ADC \to pC}$ provides its conversion to charge. $C_{ADC \to pC}$ is determined for each channel by the Charge Injection System (CIS) as will be discussed later. $C_{pC \to GeV}$ is the conversion factor from charge to energy, determined by studies with incident particles during the test beam. The test beam was performed for 11% of the TileCal modules and used muons, electrons and hadrons with known energy ranging from 3 to 350 GeV. The studies with electron beams allowed the determination of the calorimeter electromagnetic scale. The value established was 1.050 ± 0.003 pC/GeV.

C_{Cs} corrects for the non-uniform response of the cells, and is determined by the Cs calibration system. C_{Laser} corrects for PMTs response non-uniformity and is measured by the Laser calibration system.

The cell energy is given by the sum of its two readout channels. A residual number of TileCal cells have only one readout channel and in this case, the cell energy is twice the readout. The same strategy is adopted in cases of channel masking. Channel masking is the deliberate decision of removing a channel output from the cell energy reconstruction. It happens when the detector monitoring and data quality activities identify problems in a given channel.

5.2 TileCal Calibration Systems

The calibration of the calorimeter and its performance checks are crucial to ensure optimal energy resolution which has a major influence in physics analyses. About 1/3 of the transverse energy of jets is deposited in the TileCal [52]. Besides, the uncertainty on the jet energy scale due to the calorimeter response, including LAr, as a function of the jet transverse momentum, was estimated to be of the order of 1-3% [53] by test beam studies evaluating the calorimeter response to single hadrons of known energy. Despite the complex offline jet calibration chain, this still constitutes the largest contribution to this uncertainty for jets with high transverse momentum [54].

The TileCal signal has to go through a complex chain starting with cintillation light that is then guided through optical fibres to the PMTs that produce the electrical pulse, finally digitised by ADCs. All these steps may lead more easily to signal losses than other detectors where the signal is read more directly. For this reason, it is crucial to have a calibration and monitoring system that allows the independent access to all the detector components permitting at the same time to verify the couplings between them.

In order to do so, TileCal is equipped with three calibration and monitoring systems: cesium, laser and charge injection [52]. Each of these systems is dedicated specifically to a part of the readout chain and have a precision better than 1%. The combination of the different outcomes provides information about the full detector response, allowing to determine which

particular component is failing in case the readout chain presents any fluctuation.

5.2.1 Charge Injection

A charge injection system (CIS) monitors the front-end electronics, providing also the conversion factor of ADC counts to charge for all channels, defined above as $C_{ADC \to pC}$, and also known as CIS calibration constants. The CIS electronics is part of the channel front-end electronics and the measurements can be performed in HG or LG mode.

CIS calibration runs are taken nearly twice a week between LHC collision runs. This allows to monitor the readout electronics for every channel and identify bad ones. CIS constants are very stable over time, the typical channel-to-channel variation is 1.5%, and so these are updated just twice a year in a database used by the energy reconstruction algorithms to correct the energy measurement for the effect of the ADC slowly drifting gain.

5.2.2 Cesium

Scintillator tiles are irradiated by a gamma emitter ^{137}Cs radioactive source. In order to do so, the tiles are drilled along the z axis and the source is hydraulically moved through tubes that scan the entire calorimeter. Since the typical ^{137}Cs is a source of γs with energy much lower than typical high energy hadrons deposits, the readout of the PMT signal originated by the Cs system does not employ the standard electronics described before. Instead, an integrator and a separate 12-bit ADC are used. In this way, the Cs system monitors the joint response of the scintillator, the optical fibres and PMTs but not the electronics.

This system was particularly important to propagate the EM scale measured with electron beams in 11% of the modules of the calorimeter to the remaining modules. This was done by equalising the PMT gain to reproduce the same energy response to the energy deposit of the Cs source, as in the modules calibrated during the test beam.

During the LHC Run I, Cs scans were performed outside collision periods with periodicity of weeks or months. Its data is used to extract the C_{Cs} calibration constants that are then updated in the database and used to perform the cell energy reconstruction.

5.2.3 Laser

A laser system is used to perform the calibration and monitoring of the PMTs and readout electronics. It allows the measurement of the linearity of the PMTs response to light intensity and the inter-calibration of their gain.

The system is composed of a laser box, optical fibres and beam splitters. The laser box houses the laser head as well as an optical filter wheel, four photodiodes, two PMTs and an ^{241}Am radioactive α source. The laser produces short light pulses similar to those produced by ionising particles in the scintillators. Light intensity is set by attenuating the beam with an optical filter and is sufficient to simultaneously produce signals in all TileCal channels over

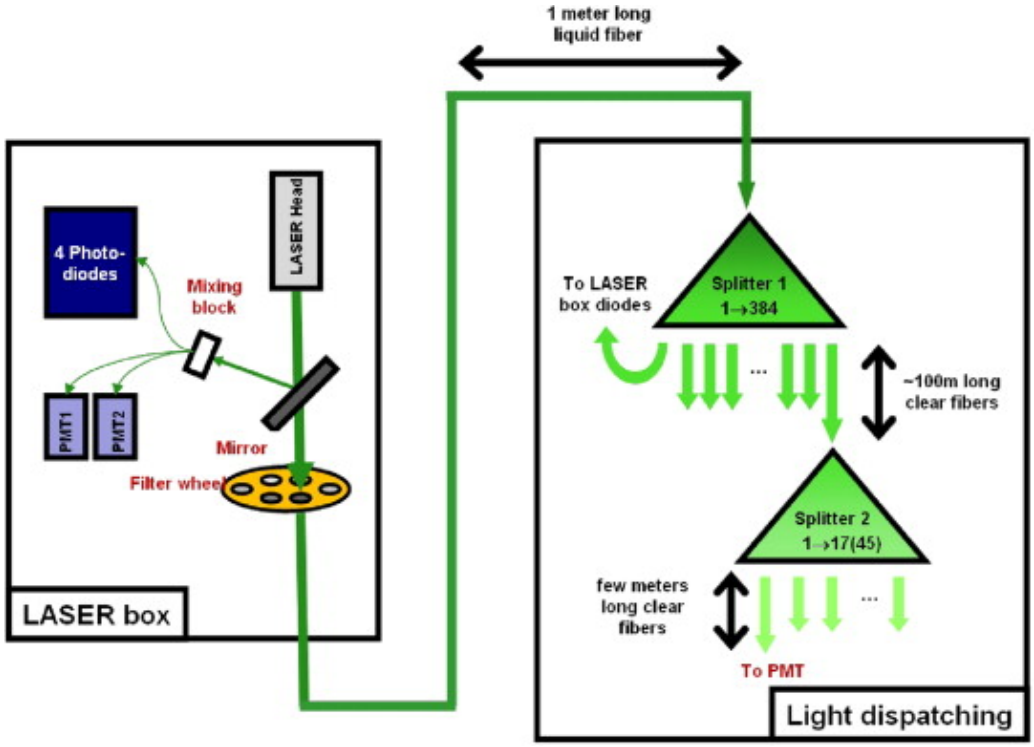


Figure 5.3: Schematic view of Laser system for TileCal calibration. Taken from [52].

their entire dynamic range. The photodiodes perform the monitoring of the laser output beam. Since their performance is very sensitive to temperature fluctuations, their monitoring is done with the radioactive source. Their absolute stability was measured to be below 0.5% [55]. Laser light pulses are transmitted outside the box and a first beam splitter distributes the light by 384 $\sim$ 100 m long optical fibres. At this stage, one of the fibres is read by a photodiode in the laser box. Light beams associated with the remaining fibres are then split and delivered to the optical fibre illuminating each of the PMTs as shown in Figure 5.3.

The stability of the PMTs is monitored every other day with dedicated laser runs, with two beam light intensities each corresponding to an ADC operating gain. These runs are taken in absence of proton beams in the LHC, but runs are also taken during the empty bunches of an LHC fill mainly to monitor the calorimeter timing. The laser constants, C_{Laser}, are determined for every PMT with laser runs data and were used to correct the cell reconstructed energy dependence on the PMT gain fluctuation over time during the LHC Run I. The data treatment employed to determine the laser constants will be described in Section 5.4.

5.3 Laser Calibration Constants and PMT gain monitoring

The cesium is the main system used to calibrate the TileCal energy scale, but these runs can not be performed very often since they need about 6 h to complete. Relative calibrations between two Cs scans are therefore accomplished with the laser. The laser calibration procedure consists of two runs:

- Low Gain run (LG) with 10000 pulses with a constant amplitude and the filter attenuation factor equal to 3
- High Gain run (HG) with 100000 pulses with a constant amplitude and the filter attenuation factor equal to 330

For each run category, the response of a channel i to each pulse p, $E_{i,p}$, is normalised to the photodiode 1 pulse measurement, D_p, defining $R_{i,p} = E_{i,p}/D_p$.

The main purpose of the laser runs is to obtain the photomultipliers gain trends over time, and determine the laser constants, C_{Laser} in Eq. 5.2. A main approach is used to analyse the laser data and determine these constants, known as the relative method [56]. The analysis is performed averaging the channel response with respect to all laser pulses in a run: $R_i =< R_{i,p} >$. Then, the method determines the gain deviation for each PMT by computing the deviation, Δ_i, of the channel response R_i with respect to the same quantity for a reference run taken just after a Cs scan, R_i^{ref}

$$\Delta_i = \frac{R_i - R_i^{ref}}{R_i^{ref}} \tag{5.3}$$

Two correction terms are applied to Δ_i to account for laser light instability and inequalities in the light transmission:

- Δ^{global}: global deviation calculated as the average of all observed Δ_i, computed iteratively by removing outliers. Known unstable channels, as those located in the innermost A layer, are initially excluded from this evaluation for being more exposed to radiation. This correction is attributed to the laser system instability and has a value of the order of 1%.
- $\Delta_{f(i)}^{fibre}$: average deviation for PMTs connected to the same distributor optical fibre, after being corrected for Δ^{global}. It uses the same iterative procedure with removal of outlier channels as before. $\Delta_{f(i)}^{fiber}$, also at the percent level, is attributed to inequalities in the beam light splitting and transmission among optical fibres.

The corrected deviation, Δ_i^{corr}, is then given by

$$\Delta_i^{corr} = \Delta_i - \Delta^{global} - \Delta_{f(i)}^{fibre} \tag{5.4}$$

And the laser constants, per channel, are defined as

$$C_{Laser} = \frac{1}{1 + \Delta_i^{corr}} \tag{5.5}$$

This approach has the advantage of compensating for laser instability which was a real and known problem, and for non-uniformity in the light transmission and beam splitting. However, it is not able to detect global effects that influence all TileCal PMTs such as PMT

ageing. In order to evaluate such effects, the results obtained by this method are compared with the results obtained with a statistical method, intended to be less sensitive to the beam light intensity instability and transmission since it is based on the statistical nature of photo-electron production and multiplication in the photomultiplier tubes [57].

In the statistical method, the PMT gain is determined from the measurement of the charge distribution and from the distribution of the light intensity during a laser run:

$$G \propto \frac{Var(q)}{< q >} - < q > \frac{Var(I)}{< I >^2} \tag{5.6}$$

where $< q >$ and $Var(q)$ are respectively the average and variance of the charge distribution measured by a PMT during a run. $< I >$ and $Var(I)$ are the average and variance of the light intensity distribution for that run and depend exclusively on the laser system. It can be shown that the $\frac{Var(I)}{<I>^2}$ factor can be calculated from the correlation between different PMT measurements, eliminating the dependence of the statistical method on the light box measurements.

$$\frac{Var(I)}{< I >^2} = \frac{Cov(q_i, q_j)}{< q_i >< q_j >} \tag{5.7}$$

where $< q_i >$ and $< q_j >$ are the average of the charge distributions measured by PMTs i and j respectively, and $Cov(q_i, q_j)$ the covariance of the distributions.

The outcomes of the statistical and relative methods were compared and considered compatible. This indicates that the relative procedure is correctly compensating for the beam light instability and transmission effects. Therefore, it is used to determine the laser constants for simplicity reasons and because it is less affected by statistical fluctuations.

The LG and HG laser runs are both used to determine the laser constants. However, since the LG run has more precision due to larger light intensity with respect to HG, only LG-based constants are used in the energy calibration. The HG runs serve a cross-check purpose, and channels are only corrected by the C_{Laser} factors if the gain deviation observed in HG and LG are compatible. Moreover, only channels that have significant gain deviations, i.e. clearly above the precision of the laser system of about 1.5% in the barrel and 2% in the extended barrel, are calibrated [56].

5.4 Laser-based method to retrieve channel quality

In order to pursue the physics goals of LHC, the ATLAS detector must be carefully monitored over time and its data quality must be ensured. The TileCal is no exception, and therefore the detector operating functions are evaluated during collision data taking and with dedicated calibration runs. The monitoring and data quality assessment of TileCal comprehends routine tasks that will be discussed in this Section. Moreover, a method to automatically identify readout channel malfunctioning using long-term laser data is reported here.

5.4.1 TileCal Monitoring and Data Quality

Data flow control and quality assessment is a regular procedure of the detector operation. It was also of most relevance during the detector re-commissioning in the LHC end of year shut down, where several components were consolidated.

The TileCal monitoring activities involve the continuous analysis of the detector control system (DCS) information, online data review and inspection of calibration data. DCS not only allows the remote control of the detector but also provides a readout of the most basic parameters that influence the detector behaviour, such as temperature, high voltage set to the PMTs or LVPS state.

The quality control of collision data obeys an established protocol of data quality. A representative sample of the data taken is fast processed and soon made available for inspection. Most of the process is automated by routines providing the data quality shifter with the necessary material to make a statement on the detector condition. During pp collisions, a run must be analysed immediately after this first fast reconstruction and in situations where the data is not considered good enough for physics analyses it is excluded. If just few channels present bad quality data, a detector conditions database is updated with this information in order to mask them allowing its removal from the data processing.

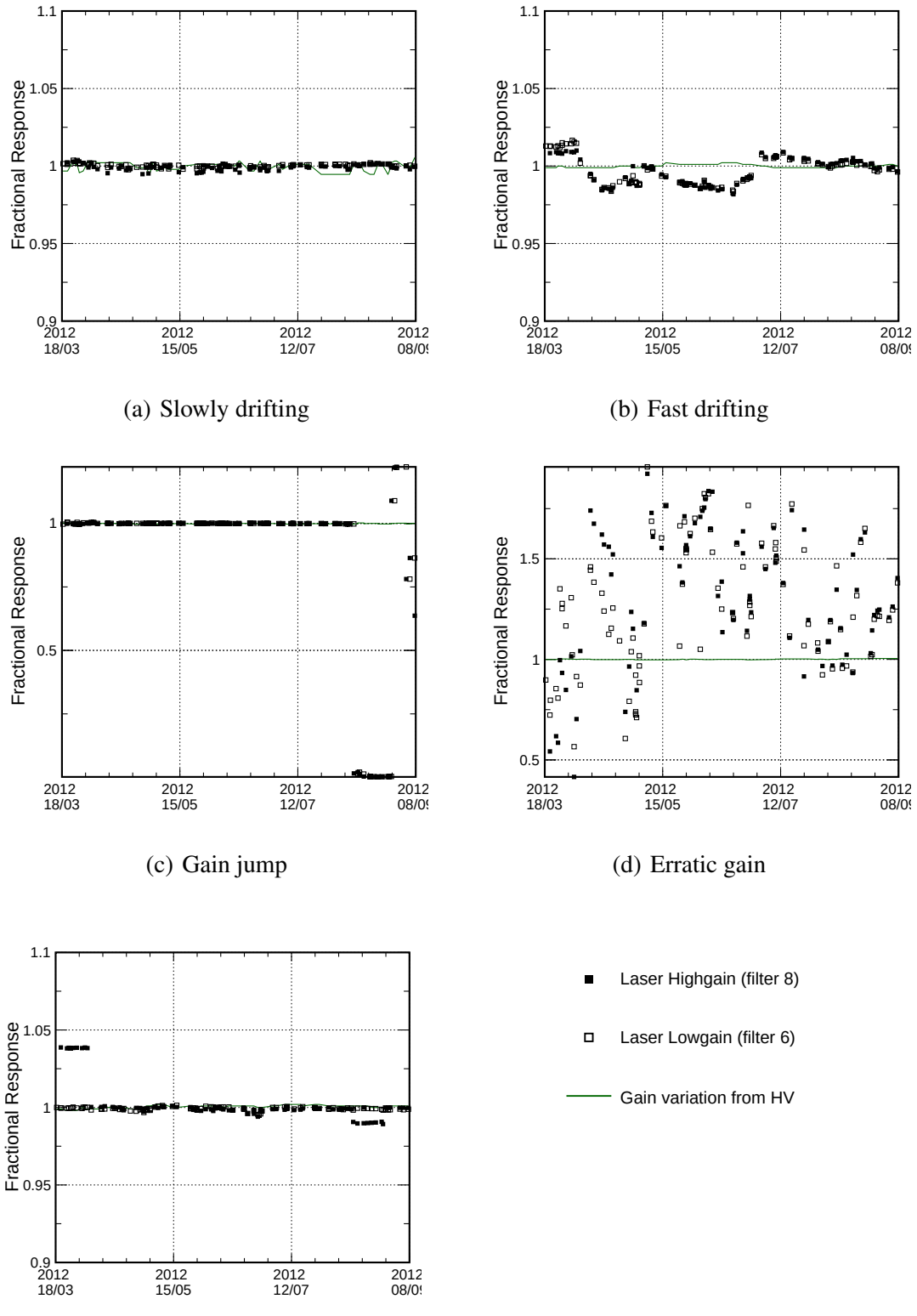
It is the processed run that is used for physics analyses and the DQ checks process must be done within a few hours after the run data taking. This procedure has a direct impact on the quality of the measurement of jets and $E_{\mathrm{T}}^{\mathrm{miss}}$. Since these objects rely on the calorimeter information, the usage of the energy measurement given by malfunctioning channels results in badly reconstructed jets and on erroneous determination of the $E_{\mathrm{T}}^{\mathrm{miss}}$.

Calibration runs are also periodically taken, providing additional information about the detector performance. These runs are particularly important to verify the response of the channels over time to a controlled energy deposit or light input. Problems that are not evident in single physics run, can in this way be identified.

5.4.2 Development of an Algorithm to identify bad channels

An automated method to detect an anomalous response of PMTs, as seen by the laser calibration system, was developed. By studying variables derived from laser calibration data, the aim is to provide quality discriminants for each channel, based on the evaluation over time of the PMT's response to the laser pulses.

During the LHC Run I, several patterns of photomultipliers misbehaviour have been recognised. Typically, the problematic channels response fall in one of the following categories: the gain exhibits a fast drift or a jump, the gain is erratic or the channel shows an incompatible response between high and low gain runs. About 60 channels were known to present one of the mentioned behaviours in 2012, according to a hand scan performed to the PMTs. A stable channel is one that does not exhibit either of the referred patterns. Figure 5.4 illustrates this classification, showing the typical channel response associated with each category.



(a) Slowly drifting

(b) Fast drifting

(c) Gain jump

(d) Erratic gain

(e) High and low incompatible

Figure 5.4: Example of the behaviour of the channels gain deviation: (a) Slowly drifting, (b) Fast drifting, (c) Gain jump, (d) Erratic gain and (e) High and low incompatible. The Fractional Response is defined as the PMT gain response, normalised to a previous laser run reference, corrected for global and fibre effects.

In order to automatize the detection of these malfunctions, a basic approach was implemented. The relative data treatment method described before in Section 5.3 is used as a basis. Channels are associated with a slow gain drift, fast gain drift, gain jump or erratic gain flag based on the derivative of the gain deviation with respect to time, defined as

$$\frac{\mathrm{d}}{\mathrm{dt}}\Delta = \frac{\Delta_{t_2} - \Delta_{t_1}}{t_2 - t_1} \tag{5.8}$$

where Δ_{t_1} and Δ_{t_2} are the gain deviations Δ^{corr} defined previously, at time t_1 and t_2, respectively. The high and low incompatible flag is attributed to channels for which the gain deviation is different for high and low gain runs, as will be explained ahead. The gain derivative is computed between two successive runs of the same gain to avoid disparities between high and low gain responses. Also, runs that do not fulfil the run quality criteria are excluded. Special E3 and E4 cells were not considered in this analysis since these are already known problematic cells, being highly exposed to radiation due to its position in the detector. Moreover, since the goal of the task is to provide information about channels which are systematically misbehaved, the statistical fluctuations that may occur in the channels responses are attenuated by computing the gain deviation derivative average and standard deviation (σ) and the average gain deviation difference between high and low gain runs as will be described. For each channel, the following variables are computed

- Average of the gain deviation derivative is calculated for groups of 10 runs starting on the first 10 runs:
$$< \frac{\mathrm{d}}{\mathrm{dt}}\Delta >_1 = \frac{1}{10}\sum_{i=1}^{10}\frac{\mathrm{d}}{\mathrm{dt}}\Delta_i \tag{5.9}$$
- σ of the gain deviation derivative is calculated for the first 10 runs:
$$\sigma(\frac{\mathrm{d}}{\mathrm{dt}}\Delta)_1 = \sqrt{\frac{1}{10}\sum_{i=1}^{10}(\frac{\mathrm{d}}{\mathrm{dt}}\Delta_i - < \frac{\mathrm{d}}{\mathrm{dt}}\Delta >_1)^2} \tag{5.10}$$
- Average of the High/Low deviation differences is calculated for the first 10 runs:
$$< \Delta^{HL} >_1 = \frac{1}{10}\sum_{i=1}^{10}|\Delta^H_{t_i} - \Delta^L_{t_i}| \tag{5.11}$$
- The same quantities are recalculated for the next 10 data points with a 5 data points offset, yielding:
$$< \frac{\mathrm{d}}{\mathrm{dt}}\Delta >_2 \text{ , } \sigma(\frac{\mathrm{d}}{\mathrm{dt}}\Delta)_2 \text{ and } < \Delta^{HL} >_2$$
- The process is repeated until the last data point is reached.

Figure 5.5 illustrates the technique. At the end, the following parameters are determined:

- Maximum average value of the high-low deviation difference as defined below. This variable is sensitive to differences in the channel response to the high and low gain runs,

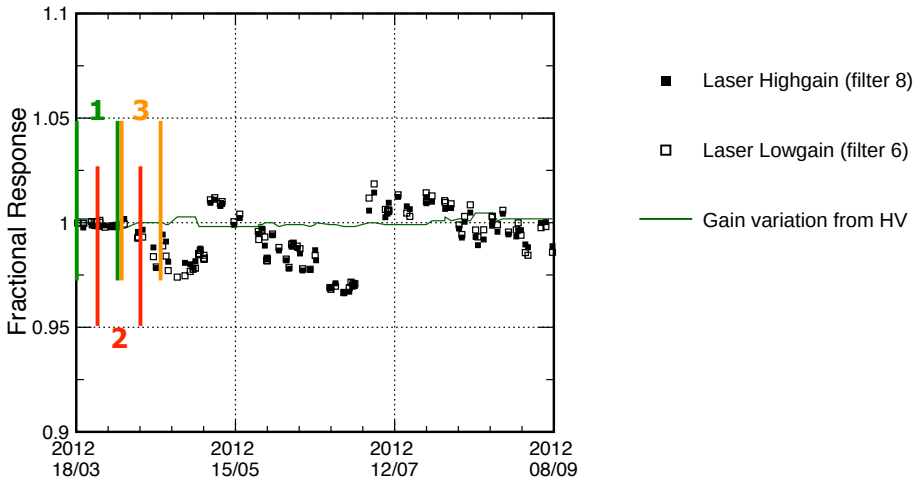


Figure 5.5: Smoothing technique used to compute the variables needed to identify problematic channels.

as shown in Figure 5.4(e);

$$\max\{<\Delta^{HL}>_1, <\Delta^{HL}>_2, ...\} \tag{5.12}$$

- Maximum average of the gain deviation derivative defined by Eq. 5.13. This variable is sensitive to the response trend along the time and can be useful to distinguish between slowly and fast drifting channels, as shown in Figures 5.4(a) and 5.4(b), respectively;

$$\max\{<\frac{\mathrm{d}}{\mathrm{dt}}\Delta>_1, <\frac{\mathrm{d}}{\mathrm{dt}}\Delta>_2, ...\} \tag{5.13}$$

- Maximum standard deviation of the gain deviation derivative, Eq. 5.14. This variable is sensitive to the channels erratic response as shown in Figures 5.4(d);

$$\max\{\sigma(\frac{\mathrm{d}}{\mathrm{dt}}\Delta)_1, \sigma(\frac{\mathrm{d}}{\mathrm{dt}}\Delta)_2, ...\} \tag{5.14}$$

- Maximum gain deviation derivative defined below. This variable is sensitive to jumps in the channels response as shown in Figure 5.4(c).

$$\max\{\frac{\mathrm{d}}{\mathrm{dt}}\Delta_i\}, \quad i = 1, .., N \text{ runs} \tag{5.15}$$

Since the PMT gain is proportional to V^7, where V is the applied high voltage (HV), the HV stability has a critical influence on the PMT gain stability. The TileCal HV system was designed to provide a nominal voltage of 690 V to each of the 9852 PMTs, coping with a stability requirement better than 0.5 mV for every channel. The applied voltage per channel is controlled and monitored by the DCS system allowing to quantify the PMT gain drift originated by HV drifts. Figure 5.6 shows an example of a channel in which a HV gain jump induced an

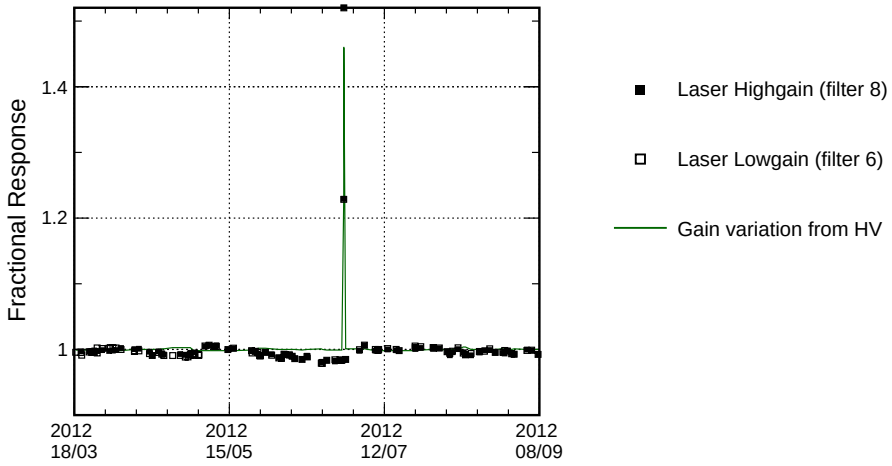
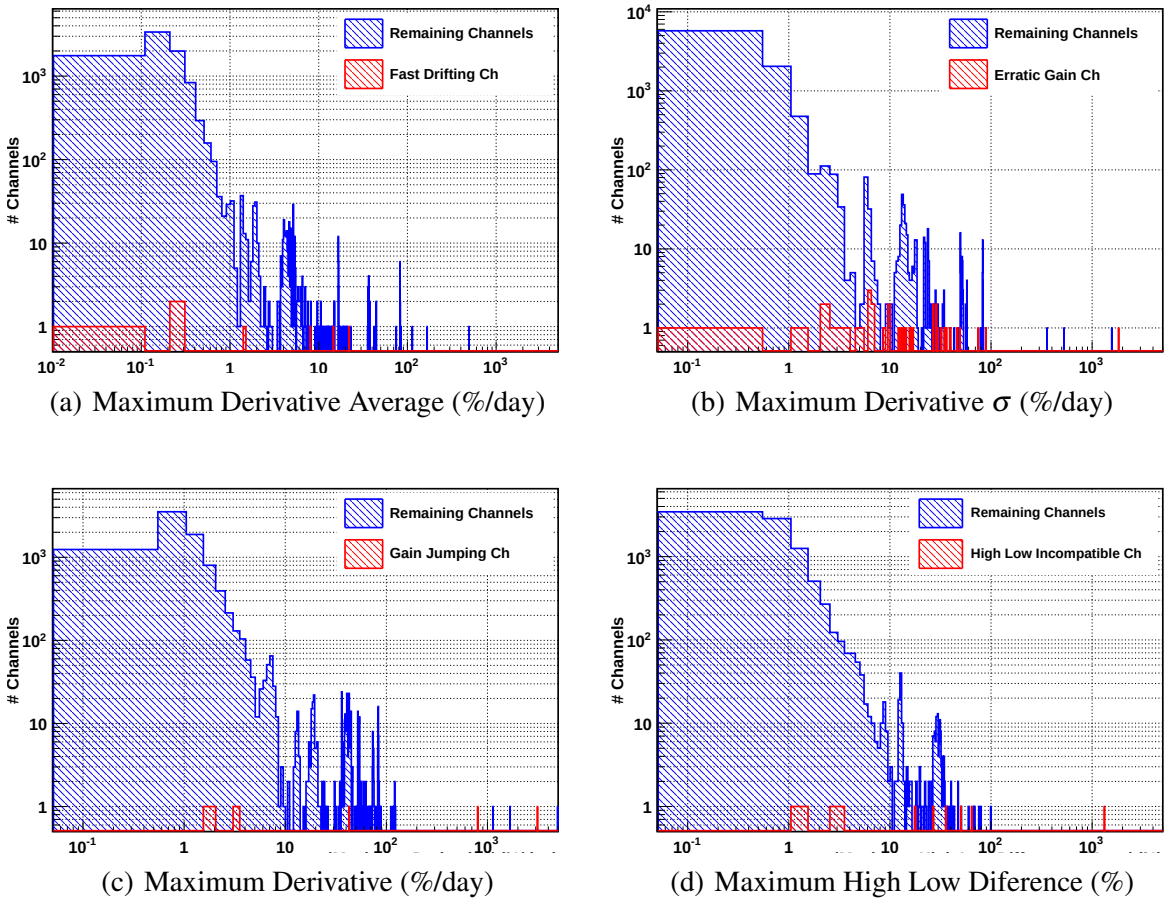


Figure 5.6: Example of a channel with a jump in the fractional response due to the HV set.



(a) Maximum Derivative Average (%/day)

(b) Maximum Derivative σ (%/day)

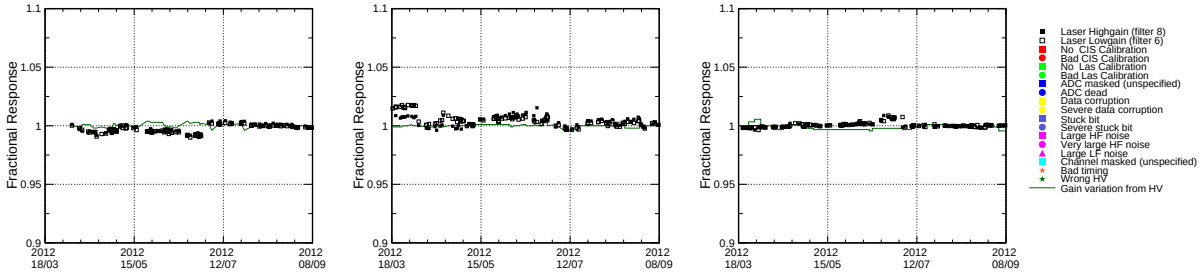
(c) Maximum Derivative (%/day)

(d) Maximum High Low Diference (%)

Figure 5.7: Distribution of the variables sensitive to channel behaviour for channels presenting a known problem and the remaining channels in TileCal, concerning the full 2012 laser calibration dataset.

Flag	Criterium
Slow Drifting Channels	Maximum Derivative Mean $< 3\%/day$
Fast Drifting Channels	Maximum Derivative Mean $\geq 3\%/day$
Erratic Gain Channels	Maximum Derivative $\sigma > 10\%/day$
Gain Jump Channels	Maximum Derivative $> 12\%/day$
High Low Incompatible Channels	Maximum High Low Difference Mean $> 5\%$

Table 5.1: Criteria for the attribution of the laser flags.



(a) EBA m22 ch41 (b) EBA m61 ch15 (c) LBA m64 ch14

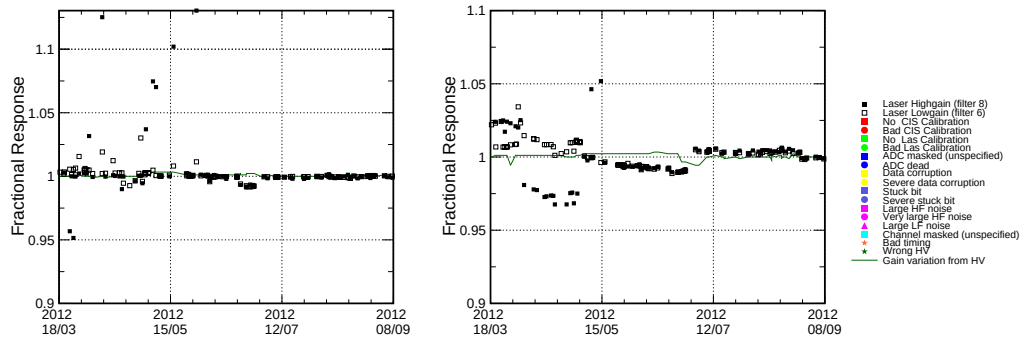
Figure 5.8: Fractional response for three TileCal channels considered unstable by the laser DQ activities and not flagged by the automated method.

equivalent gain jump as seen by the laser system. To avoid interpreting a large HV drift as an instability of the PMT gain, data points from laser runs with large gain deviations due to the HV set (above 10%) were primarily removed from the computation of the variables described above.

The method was developed based on data taken between March and September of 2012. The distributions of the variables defined above are presented in Figure 5.7. For the great majority of the PMTs, 8000, the maximum average of the gain drift observed in a 10 run period is below 0.4%/day, while the absolute maximum drift observed for the total period was 3%/day. The maximum standard deviation of the drift in a 10 run period is below 1%/day and the maximum difference in the response to low and high gain laser runs is lower than 1% for 80% of the PMTs.

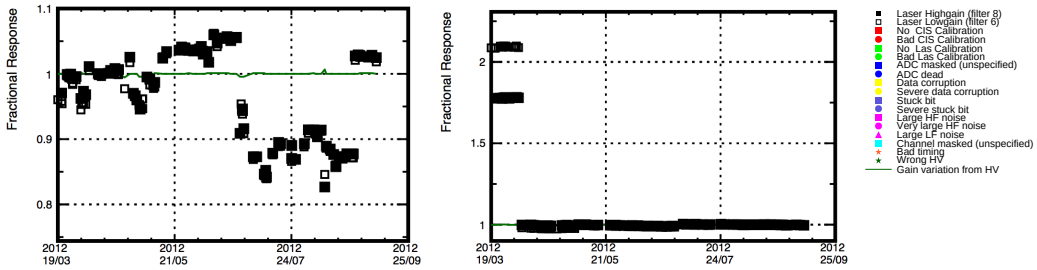
The shape of the distributions presents a strong slope towards higher values of the parameters under consideration, with a cluster of channels lying in the outlier region. This defined a set of criteria designed to identify the cluster of outlier channels. Table 5.1 summarises these criteria, that are then used to attribute a specific flag to a TileCal channel.

An indicative list of already known 62 problematic channels was already available. The intention of this analysis was to correctly associate the channels in this list with the corresponding flag, in order to validate the method. Known problematic channels belonging to that list are shown in red in the distributions of Figure 5.7. Some of them appear on the low region of the spectra meaning that either these variables are not sensitive to their specific problem or that they do not present a severe problem. Some examples of these channels are



(a) LBA m29 ch00 (b) LBA m56 ch07

Figure 5.9: Fractional response for two TileCal channels flagged by the automated method but not indicated as unstable by the laser DQ activities.



(a) LBC m63 ch41 (b) LBA m24 ch29

Figure 5.10: Fractional response for two TileCal channels associated with more than one problem: (a) erratic and fast drifting gain, and gain jump; and (b) gain jump and high/low gain difference.

presented in Figure 5.8. For these cases, the PMT gain instability is not very severe, with the gain variation falling below 3%.

On the contrary, many of the outlier channels were not indicated in the list provided by the laser DQ team. Figure 5.9 exemplifies these cases although many others were encountered. As shown, the gain is much more unstable over time for these two channels than for any of the three channels in the previous example.

Thus, this analysis and the criteria designed to identify bad channels, shown in Table 5.1, also needed to compromise in the fact that most of the channels in the list were not flagged by the method.

Many channels fall in several categories of problems. In some cases this is because they effectively present different problems. In other cases, this happens because the variables that determine the categorisation are correlated. No attempt was made to disentangle these cases. Figure 5.10 shows two examples of such situations.

	Flagged		Flagged by DCS/DQ	
	Number	Fraction (%)	Number	Fraction (%)
Total	578	5.9	274	2.8
LBA	178	6.2	13	0.4
LBC	280	9.7	156	5.4
EBA	15	0.7	2	0.1
EBC	105	5.1	103	5.2
LB A cells	209	8.1	83	3.2
LB BC cells	183	7.9	70	3.0
LB D cells	66	7.3	16	1.7
EB A cells	46	3.6	40	3.1
EB B cells	50	3.2	44	2.8
EB D cells	16	2.1	14	1.8
EB E cells	8	1.5	7	1.4

Table 5.2: Absolute number and fraction (%) of flagged channels by the laser calibration-based method in 2012. The results are splited by the TileCal partitions (LBA, LBC, EBA and EBC), barrel (LB) layers (A, BC and D cells) and extended barrel (EB) layers (A, B, D and E cells). The fraction of flagged channels is relative to the total number of channels in each category.

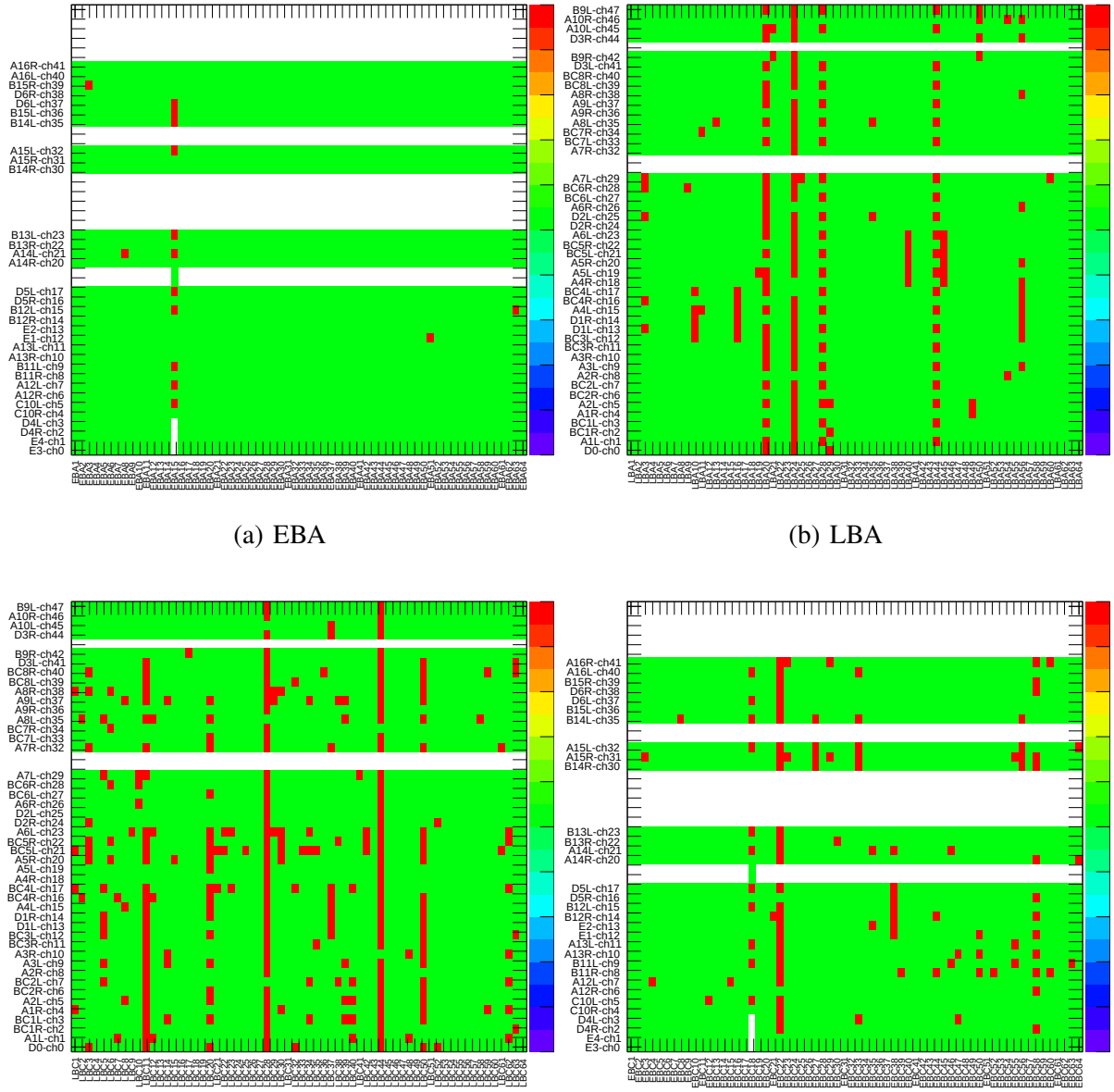
5.4.3 Results and Discussion

The analysis was applied to the entire data sets of 2011 and 2012 obtained with the TileCal laser system.

578 of the 9852 TileCal photomultiplier channels were considered to present an abnormal behaviour during at least a 10 run period in 2012 by applying the set of criteria summarised in Table 5.1. Table 5.2 shows the number and fraction of TileCal channels flagged by the laser-based method. DCS and DQ activities already identify some problems that can affect the PMT readout, listing them in a data base. These problems include bad CIS calibration manifesting affected ADCs, data corruption or stuck bits in the channel propagation of the digital signal through the readout chain. Bad timing of the PMT signal sampling is other issue that strongly affects the evaluation of the PMT gain deviation. Among the 578 flagged channels, 274 had already been tagged with these problems by the DCS and DQ activities. So, overall, 3.1% of the Tilecal channels, passing the detector control activities, were associated with at least one flag by the laser-based method for 2012.

In general, channels reading the barrel have more problems than the ones associated with the extended barrel. As the barrel is more exposed to radiation than the extended barrel (excluding E3 and E4 cells) this is an expected feature of the TileCal operation. The results do not present a clear relation between unstable channels and the TileCal layers, although channels reading A cells were expected to present more problems if exposure to radiation is what causes the PMT gain to fluctuate.

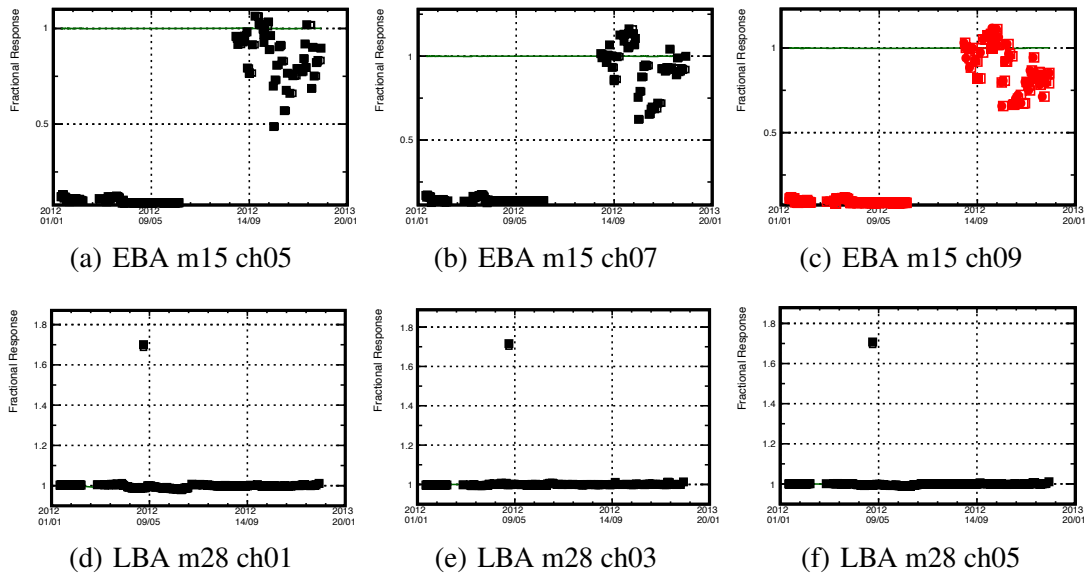
Figure 5.11 shows the the TileCal mapping of flagged channels per module for the four partitions of the detector. With the exception of some channels clusters lying across the same module, unstable channels are well spread across the detector volume. The clusters of bad



(a) EBA (b) LBA

(c) LBC (d) EBC

Figure 5.11: Mapping of the TileCal channels and modules for the (a) EBA, (b) LBA, (c) LBC and (d) EBC partitions. The *x*-axis contains the module number and the *y*-axis the channels number and corresponding TileCal cell name, as defined in Figure 5.1, with L and R standing for the left and right cell readout. Channels flagged with 2012 laser data, including the ones listed by DCS and DQ activities, are shown in red and uninstrumented channels in white. Green channels were considered stable or slowly drifting.



(a) EBA m15 ch05 (b) EBA m15 ch07 (c) EBA m15 ch09

(d) LBA m28 ch01 (e) LBA m28 ch03 (f) LBA m28 ch05

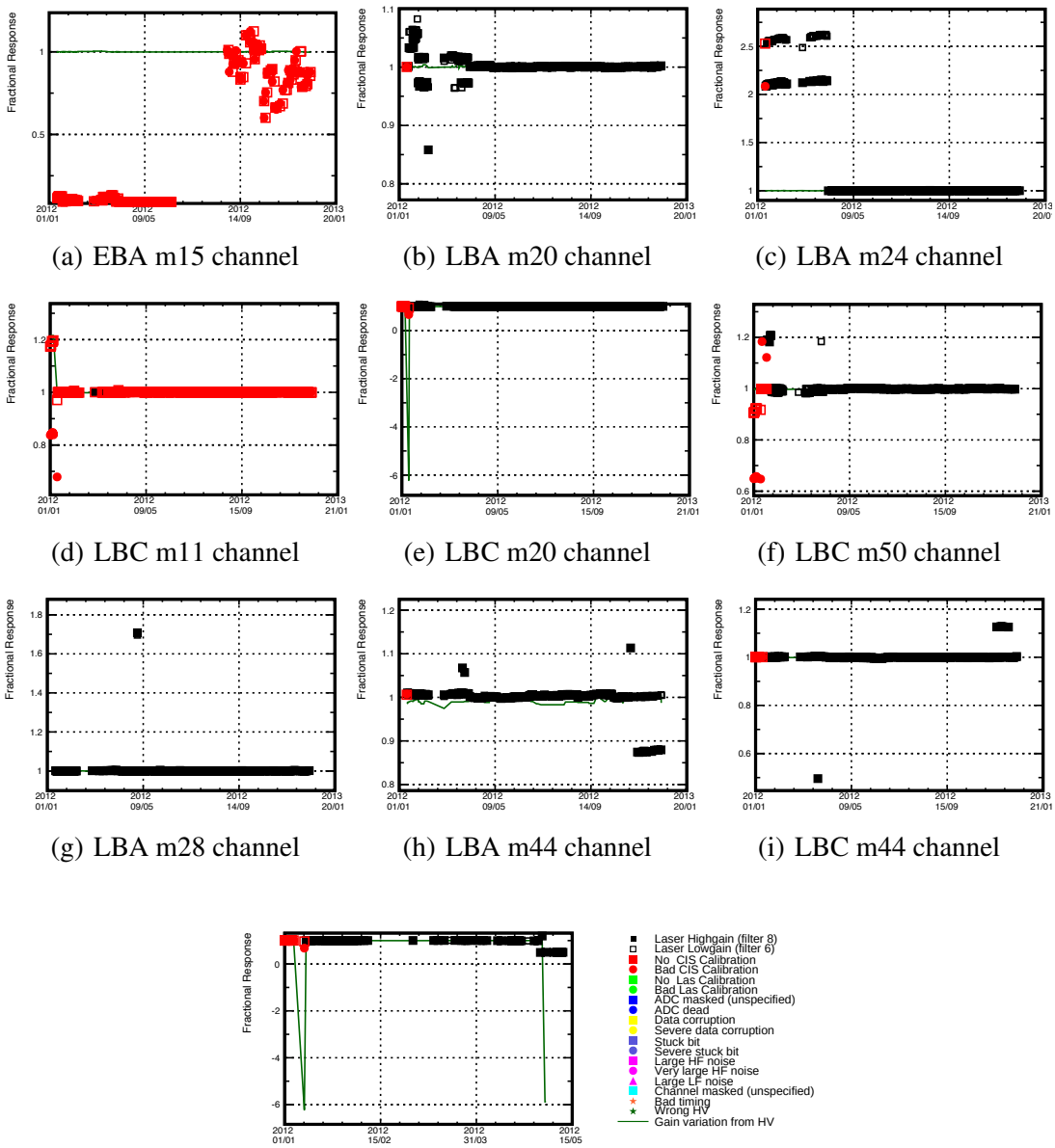
Figure 5.12: Fractional response for three channels in TileCal modules EBA 15 and LBA 28.

channels appear in the module 15 in EBA; the modules 20, 24, 28 and 44 in LBA; modules 11, 20, 28, 44 and 50 in LBC, and modules 22 and 58 in EBC. Further investigation revealed the following:

- First, the gain deviation as a function of time in 2012 had the same pattern for all the channels flagged in the same module. Figure 5.12 illustrates this observation for three flagged channels in EBA 15 and LBA 28 modules. The fact that this pattern exists is an indication of problems in the electronics of the drawer and not of the PMTs themselves. Figure 5.13 exemplifies the trend of the gain deviation with one channel per flagged module.

- The gain of the PMTs of module EBA 15 is clearly erratic since September 2012 although the low and high gain (LG and HG) responses are the same, see Figure 5.13(a). This suggests a problem of the readout and indeed the module was identified by DCS and DQ with the bad CIS calibration tag, stating problems in the ADCs. The first part of the year shows a slowly drifting gain with nominal value much lower than the reference. This can also be caused by the HV supply. Although not visible from the green line indicating the gain variation from HV set, it can happen that the DCS system was not able to monitor properly its value.

- Modules LBA 20 and 24; LBC 11, 20 and 50, and EBC 22 and 58, shown in Figures 5.13(b) to 5.13(f), recovered during the first 2012 trimester stop, where the consolidation and maintenance of the TileCal took place. The LHC physics runs re-started in April 2012 and all the channels recovered in time to take data. With the exception of LBA module 24, also unstable during 2011, the problems presented by these modules are observed only in the technical stop period, and are likely due to hardware adjustments and tests happening at the time.



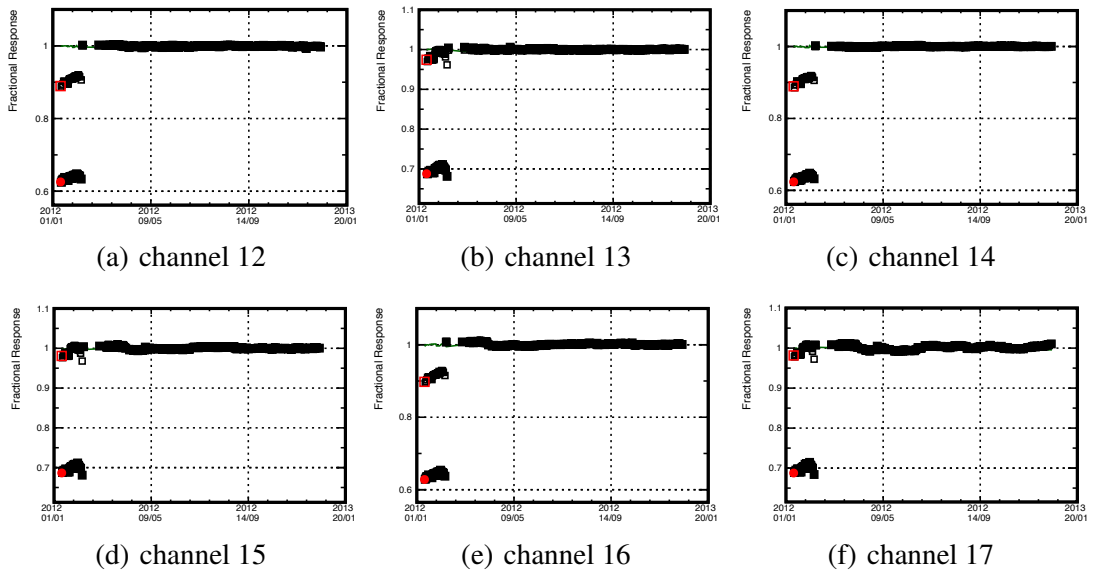
(a) EBA m15 channel (b) LBA m20 channel (c) LBA m24 channel

(d) LBC m11 channel (e) LBC m20 channel (f) LBC m50 channel

(g) LBA m28 channel (h) LBA m44 channel (i) LBC m44 channel

(j) LBC m28 channel

Figure 5.13: Example of the channel fractional response for TileCal modules where most of the channels were flagged by the automated method. The behaviour of the channels of EBC module 22 and 58 is similar to the one shown for LBC module 20.



(a) channel 12 (b) channel 13 (c) channel 14

(d) channel 15 (e) channel 16 (f) channel 17

Figure 5.14: Fractional response for flagged channels of digitiser 6 in LBA module 16.

- LBA 28, see Figure 5.13(g), had only two succesive runs with high gain deviation for the entire year, and only for odd channels. Most probably, this is a feature of the laser run itself and not of the PMTs. Although not detected by DCS, the cause may be related with the HV distribution system, common to all the odd (even) channels in a drawer, despite the HV value can be adjusted per PMT.

- LBA 44 and LBC 44, shown in Figures 5.13(h) and 5.13(i), have some runs out of the expected gain deviation value. In LBA this happens for odd channels only, suggesting local jumps in the HV supply. By the end of the year, the response to HG and LG differs by more than 10% for both modules, indicating a problem in the channel readout.

- LBC 28, in Figure 5.13(j), presents large gain deviations caused by jumps in the applied HV.

From figure 5.11, it can also be seen that some of the flagged channels correspond to a sequence of 5 to 6 channels along a module. In TileCal, a group of six PMTs is read by a single digitiser, so this method can be used to identify possible problems in the digitisers when the majority of the PMTs connected to it are flagged. This information was then handed over to the ADC monitoring system by the CIS. Eight problematic digitisers were found:

- Digitisers 3, reading channels 30 to 35, in modules EBC 27, 33 and 56

- Digitisers 5, reading channels 18 to 23, in modules LBA 40 and 45

- Digitisers 6, reading channels 12 to 17, in modules LBA 10, 16 and 56; LBC 37, and EBC 38

	Flagged		Flagged by DCS/DQ	
	Number	Fraction (%)	Number	Fraction (%)
Total	300	3.0	44	0.4
LBA	98	3.4	6	0.2
LBC	83	2.9	9	0.3
EBA	45	2.2	24	1.2
EBC	74	3.6	5	0.2
LB A cells	84	3.1	9	0.4
LB BC cells	73	3.2	4	0.2
LB D cells	24	2.7	2	2.2
EB A cells	43	3.4	6	0.5
EB B cells	40	2.6	13	0.8
EB D cells	26	3.4	7	0.9
EB E cells	10	1.9	3	0.6

Table 5.3: Absolute number and fraction (%) of flagged channels by the laser calibration-based method in 2011. The results are splited by the TileCal partitions (LBA, LBC, EBA and EBC), barrel (LB) layers (A, BC and D cells) and extended barrel (EB) layers (A, B, D and E cells). The fraction of flagged channels is relative to the total number of channels in each category.

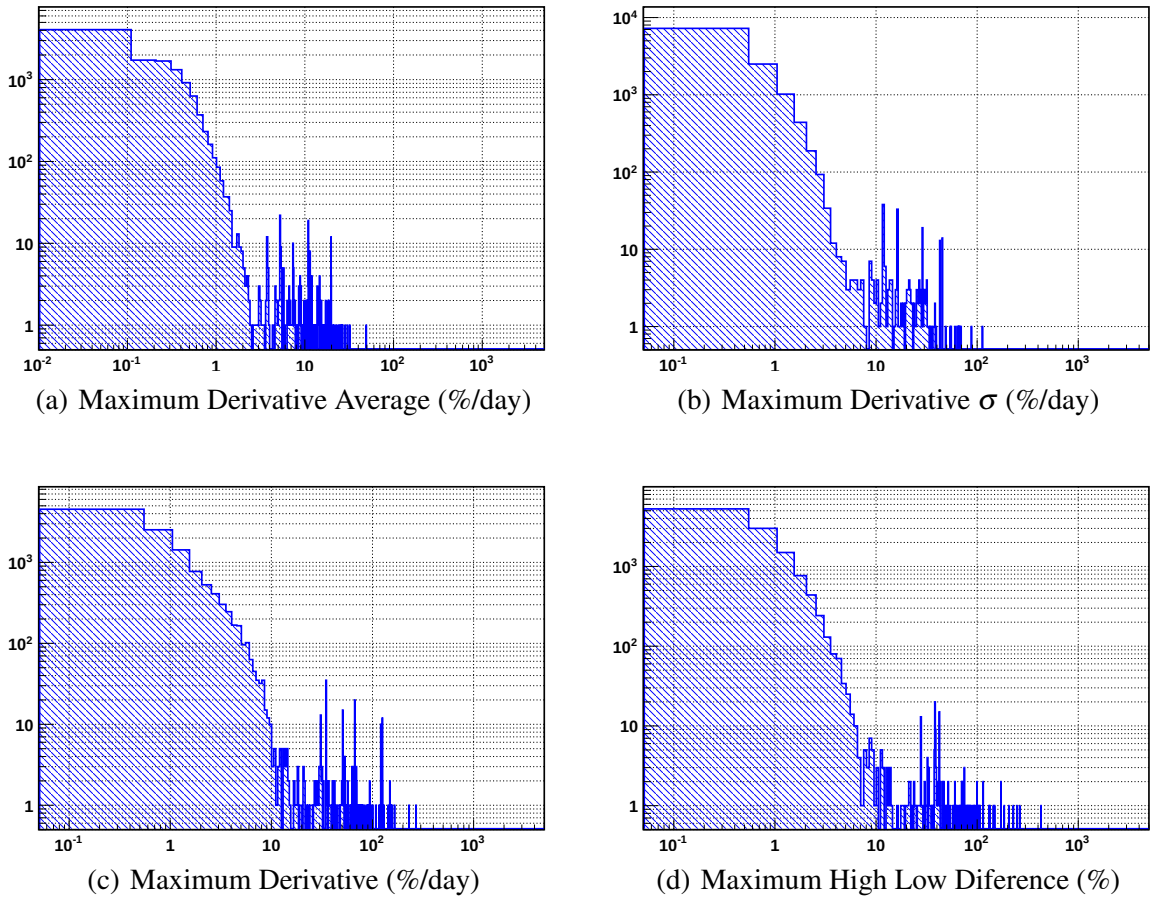
Further investigation also revealed that the gain deviation trend is similar within flagged channels read by the same digitiser and that the abnormal gain behaviour took place during maintenance period. All the channels were recovered before the collision runs. As an example, Figure 5.14 presents the fractional response of the different channels read by digitiser 6 in LBA module 16.

If flagged channels in the same module or digitiser are excluded, only 223 TileCal PMTs were flagged with some kind of gain instability, representing a percentage of 2.3% of all the PMTs. Figure 5.11 shows that these channels are spread across the detector and also that only three cells of the TileCal have both readout PMTs compromised: cells BC5 (channels 21 and 22), D1 (channels 13 and 14) and A3 (channels 9 and 10) in the modules LBC 3, 5 and 14, respectively. For the remaining cells, the readout redundancy of TileCal ensures that the cell energy can be well measured by the non-affected PMT.

The same analysis was applied to the 2011 laser calibration data. Figure 5.15 shows the distributions of the variables under analysis for this data set. The distributions are similar to the ones obtained with the 2012 data, with the great majority of channels lying in the low value region, a pronounced fall towards higher values and then a cluster of outlier channels located in the end of the spectra.

Table 5.3 summarises the results obtained for the 2011 data set. About 2.5% of the channels in the detector were identified with a gain instability problem or with very different responses in HG and LG runs. Again, the channels identified with problems do not tend to be specifically located in the detector. This is also visible from the detector mapping shown in Figure 5.16.

The procedure was repeated to investigate the nine modules of the detector with most of



(a) Maximum Derivative Average (%/day)

(b) Maximum Derivative σ (%/day)

(c) Maximum Derivative (%/day)

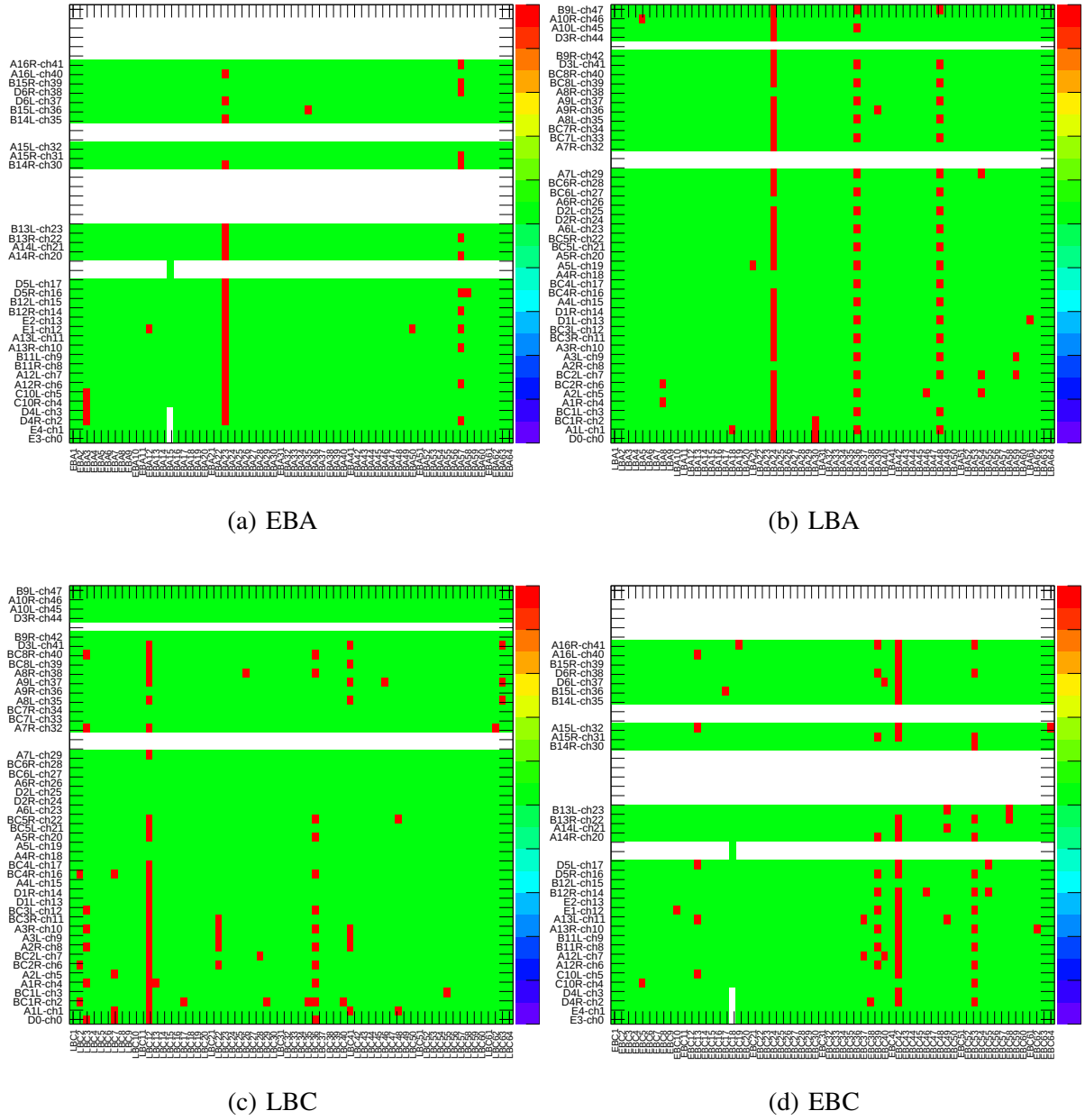
(d) Maximum High Low Diference (%)

Figure 5.15: Distribution of the variables sensitive to channel behaviour concerning the full 2011 laser calibration dataset.

the channels exhibiting an abnormal response in laser data.

- Two of them had just one isolated bad laser run while three modules had about four data points with large gain deviation affecting all channels in the same manner.
- Two modules were unstable during the beginning of the year technical maintenance and were recovered in time for the physics run.
- Even channels in EBC 53 have critical gain jumps, of about 30%, for a dozen runs during the entire 2011 revealing problems in the HV distribution system.
- Finally, LBA 24 has several gain jumps, different response to HG and LG laser runs and off periods. As seen before, this module is afterwards recovered for the 2012 run.

Moreover, two bad digitisers were located. If the channels corresponding to these modules and digitisers are discarded, a total number of 94 out of the 9852 PMTs of TileCal, i.e. 1% in percentage, are identified as presenting an abnormal response to the laser light.



(a) EBA (b) LBA

(c) LBC (d) EBC

Figure 5.16: Mapping of the TileCal channels and modules for the (a) EBA, (b) LBA, (c) LBC and (d) EBC partitions. The *x*-axis contains the module number and the *y*-axis the channels number and corresponding TileCal cell name, as defined in Figure 5.1, with L and R standing for the left and right cell readout. Channels flagged with 2011 laser data, including the ones listed by DCS and DQ activities, are shown in red and uninstrumented channels in white. Green channels were considered stable or slowly drifting.

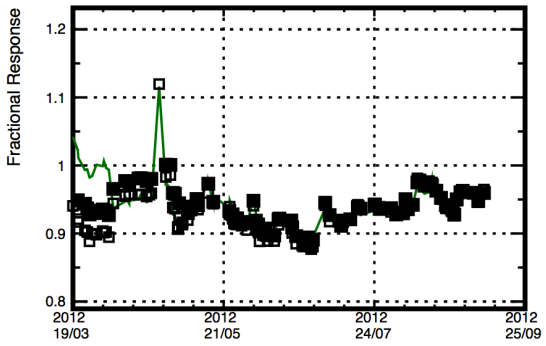


Figure 5.17: Example of the fractional response for a channel exhibiting a fast drifting gain due to the HV instability.

5.4.4 Prospects for future work

The laser-based method to identify bad channels was implemented in a common software framework dedicated to the analysis of the TileCal calibration data, that additionally has access to the detector conditions data bases and DCS monitoring information. The routine implemented is capable of detecting malfunctioning channels in the detector in an automated fashion. However, it does not dismiss the user from a subsequent carefull analysis, where interplaying with information from other TileCal systems and its operation is crucial.

This method can be further improved and below is a list of suggestions for future study and development:

- Since the HV supply critically affects the gain of the PMTs, it could be useful to decorrelate both effects by analysing the gain deviation normalised to the drift due to HV. Figure 5.17 shows an example of a channel exhibiting a fast drifting gain due to the HV instability that would not be considered a drifting channel if these effects were decorrelated. However, since currently there is no automated method to systematically monitor the HV power supply, this could result in loss of information.

- As discussed before, many of the flagged channels were recovered before the physics runs. It would be useful to check if at a certain point in time the channel recovers.

- Other common feature encountered was the isolated bad run effect, as Figure 5.13(g) exemplifies, where just one or few runs made entire modules to be flagged. As this is not likely a PMT problem, a category dedicated to trigger these runs could be included to allow investigation of the problem and check their influence on the ongoing physics runs.

- Study correlations between the different variables to disantangle cases where channels fall in more than one problem category and to better tune the method.

5.4.5 Conclusion

An automatic procedure for attributing a DQ encoded flag was developed in order to detect TileCal photomultiplier channels presenting an unstable response based on data from the laser calibration system. The aim was that this automatic procedure can serve as a starting point to more detailed investigations. The study shown here demonstrates that the method fullfils these objectives, validating the strategy. The method was implemented in the TileCal calibration data analysis software and key points for future improvement were diagnosed.

In 2012 (2011), 2.3% (1%) of the TileCal PMT channels with persistent problems in the photomultipliers were detected with this analysis. These exclude channels that recovered for the physics runs, that had only a few bad laser runs or that have problems most probably caused by the HV power supply or digital conversion and readout. In addition, the method can provide complementary information to the CIS monitoring system by identifying possibly malfunctioning digitisers.

Chapter 6

The Higgs boson search through $b\bar{b}$ decay and W associated production

This Chapter describes the search for the Standard Model Higgs boson produced in association with a W boson and decaying to $b-$quark pairs. An introductory overview of the analysis is given in Section 6.1, followed by the characterisation of the signal and background processes, a discussion about the pp collision generators used to simulate these processes and the data set analysed, in Section 6.2. Sections 6.3 and 6.4 describe the object and event selection employed, together with the procedure to validate the analysis tools.

The multivariate method Boosted Decision Tree, BDT, is used to enhance the signal significance. The BDT method and its usage in this analysis are discussed in Section 6.5. A study to optimise its performance is there presented as well.

6.1 Overview of the $WH \to \ell\nu b\bar{b}$ channel analysis

Given the very large branching ratio of the $H \to b\bar{b}$ decay (57.7%) for a $m_H = 125$ GeV, the measurement of the Higgs decay to $b-$quark pairs is fundamental to determine the Higgs boson decay width and couplings, and to confirm or reject the Standard Model hypothesis. However, the bb decay of the Higgs is one of the most challenging searches at the LHC. This was already shown in Figure 2.17 of Section 2.2, where the SM prediction of the production cross-section of different processes is shown as a function of the centre-of-mass energy of the collisions. The bb background cross-section at $\sqrt{s}$ =8 TeV pp collisions is seven orders of magnitude greater than the Higgs production cross-section. For this reason, with hadron colliders, the access to $H \to b\bar{b}$ is practically impossible in an inclusive search.

In fact, the $H \to b\bar{b}$ decay has not been observed yet, and the Higgs coupling to down-type quarks is still to be measured. But the associated production mode, where an off-shell W/Z boson radiates the Higgs, also called Higgsstrahlung in analogy to the bremsstrahlung effect, can provide further insight. By choosing events where the vector boson decays leptonically, $W \to \ell\nu$, $Z \to \ell\ell/\nu\nu$ as shows Figure 6.1, a lepton trigger that substantially reduces the

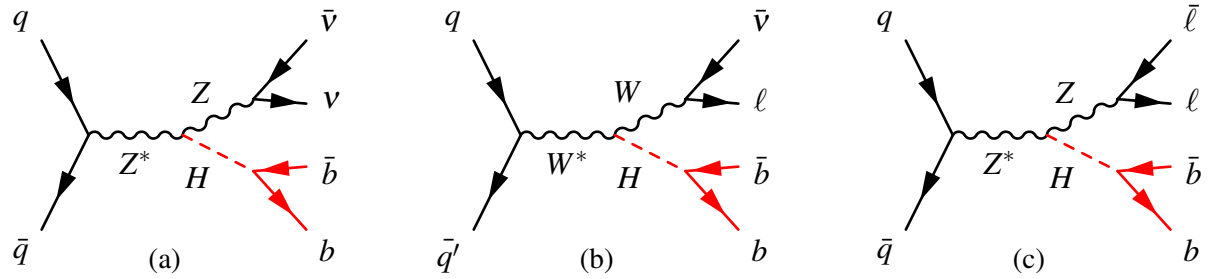


Figure 6.1: Dominant diagrams of the (a) $ZH \to \nu\nu b\bar{b}$, (b) $WH \to \ell\nu b\bar{b}$ and (c) $ZH \to \ell\bar{\ell}b\bar{b}$ signal processes at LO, where ℓ denotes (e, μ).

backgrounds while selecting the signal events, can be used. The background suppression largely compensates the rate of signal events lost by requiring this specific production mode, leading to a global increase in signal sensitivity.

Therefore, ATLAS is searching for the $H \to b\bar{b}$ decay in the associated production channel with a W or a Z boson, collectively called VH search [50]. In this context, this global analysis is conducted separately in three different channels depending on the number of charged leptons in the final state. The 0-lepton channel corresponds mainly to the signal process $ZH \to \nu\nu b\bar{b}$, the 1-lepton channel represents the search for $WH \to \ell\nu b\bar{b}$, and finally, the 2-lepton channel is associated with the $ZH \to \ell\bar{\ell}b\bar{b}$ process[1]. In all cases, ℓ represents an electron or a muon. Figure 6.1 shows the dominant diagrams of these channels. The three are combined in the statistical analysis described at Chapter 7.

The final state of the signal is characterised by 0, 1 or 2 charged leptons; two jets originated by $b-$quarks and large missing transverse energy - $E_{\mathrm{T}}^{\mathrm{miss}}$ - in the case of neutrino presence. The main backgrounds to the analysis are top pair production ($t\bar{t}$), W and Z+jets production, dibosons (WW, WZ and ZZ), single top and multijets.

The composition of the real data sample is predicted from Monte-Carlo (MC) simulation of the signal and backgrounds, with the exception of multijets. The simulation reproduces the conditions of the ATLAS Run I data, with $\sqrt{s} = 7$ and 8 TeV pp collisions. Section 6.1.2 discusses the details about the simulation of the backgrounds. Their normalisation is performed using constraints from real data in the global fit presented in Chapter 7. The MC does not reproduce accurately enough the multijet process, so a data-driven method described at 6.4.3 is used instead. The signal, described in Section 6.1.1, was generated for a Higgs mass hypothesis ranging from 100 to 150 GeV, although the analysis described in this thesis used only the m_H =125 GeV sample.

The event selection is optimised independently for each channel to maximise the sensitivity. The criteria were designed to maximise the signal to background ratio based on MC information, and will be presented in Section 6.4 for the 1-lepton case. For the same reason, some variable cuts are dependent on the vector boson transverse momentum, p_{T}^{V}. Common to

[1]This association is not totally accurate since misidentification or loss of leptons can lead to WH signal in the 0-lepton channel and $ZH \to \ell\bar{\ell}b\bar{b}$ events in the 1-lepton channel.

all selections is the requirement of 2 or 3 high transverse momentum jets. The two highest p_{T} jets have to be identified as originated by $b-$quarks. The b-tagging method uses three levels of cuts of varied tagging efficiency - 50, 70 and 80 % - corresponding to the tight (TT), medium (MM) and loose (LL) b-tagging categories. The analysis is further categorised into two regions of the p_{T}^{V} spectra: smaller or larger than 120 GeV.

The 0-lepton channel requires no charged leptons, one isolated high p_{T} electron or muon is required in the 1-lepton channel and two oppositely charged electrons or muons are required in the 2-lepton channel. In the case of the 1- and 2-lepton channels, at least one of the selected leptons must have fired the trigger while in the 0-lepton channel the $E_{\mathrm{T}}^{\mathrm{miss}}$ trigger is used. Jets, electrons and muons are reconstructed, identified and calibrated according to the techniques explained before in Chapter 4. Following calibration, the $E_{\mathrm{T}}^{\mathrm{miss}}$ is re-evaluated.

Slightly different selections are designed to obtain samples enriched in particular backgrounds allowing to control their modelling by MC and to constrain their normalisation with data. The signal sensitivity is further enhanced by the use of the Multivariate Analysis (MVA) method Boosted Decision Tree (BDT). After event selection, a BDT is trained to separate signal from background exploring fine correlations between variables, as described in Section 6.5. Several types of BDTs are constructed depending on the analysis categories: 0-, 1- or 2-lepton channels, 2 or 3 jets and p_{T}^{V} interval.

A maximum likelihood binned fit is performed simultaneously on the three channels. It combines their information to measure the expected and observed significance of the deviation from the background-only hypothesis, and the ratio of the measured signal yield to the SM expectation. The former is also referred to as signal significance while the latter as signal strength μ.

Several systematic uncertainties of both experimental and theoretical sources contribute to the full uncertainty of the result. These are incorporated in the fit as a set of nuisance parameters, and the impact of each contribution on the signal strength uncertainty $\Delta\mu$ is evaluated independently. The complete set of systematic uncertainties is discussed in 7.1.

A representative diagram of the analysis flow is exhibited in Figure 6.2. The three top blocks are the object and event selection, the multivariate analysis of the selected events and the statistical analysis of the outcome. Each of these steps will be detailed in the following in Sections 6.3, 6.4 and 6.5. The statistical analysis is presented in Chapter 7.

6.1.1 Signal event characterisation

The cross-section for the Higgs production in association with a W boson is $0.6966^{+3.7\%}_{-4.1\%}$ pb for $m_H = 125$ GeV and $\sqrt{s} = 8$ TeV centre-of-mass pp collisions at the LHC, evaluated from NLO perturbative QCD and including NNLO QCD and NLO electroweak corrections [15].

In this very rare process where an off-shell W emits a H, the main processes contributing to the W production from pp collisions are $u\bar{d} \to W^{*+}$ and $d\bar{u} \to W^{*-}$ since they benefit from one of the proton valence quarks uud. Given this, the W^+ production rate is roughly twice the

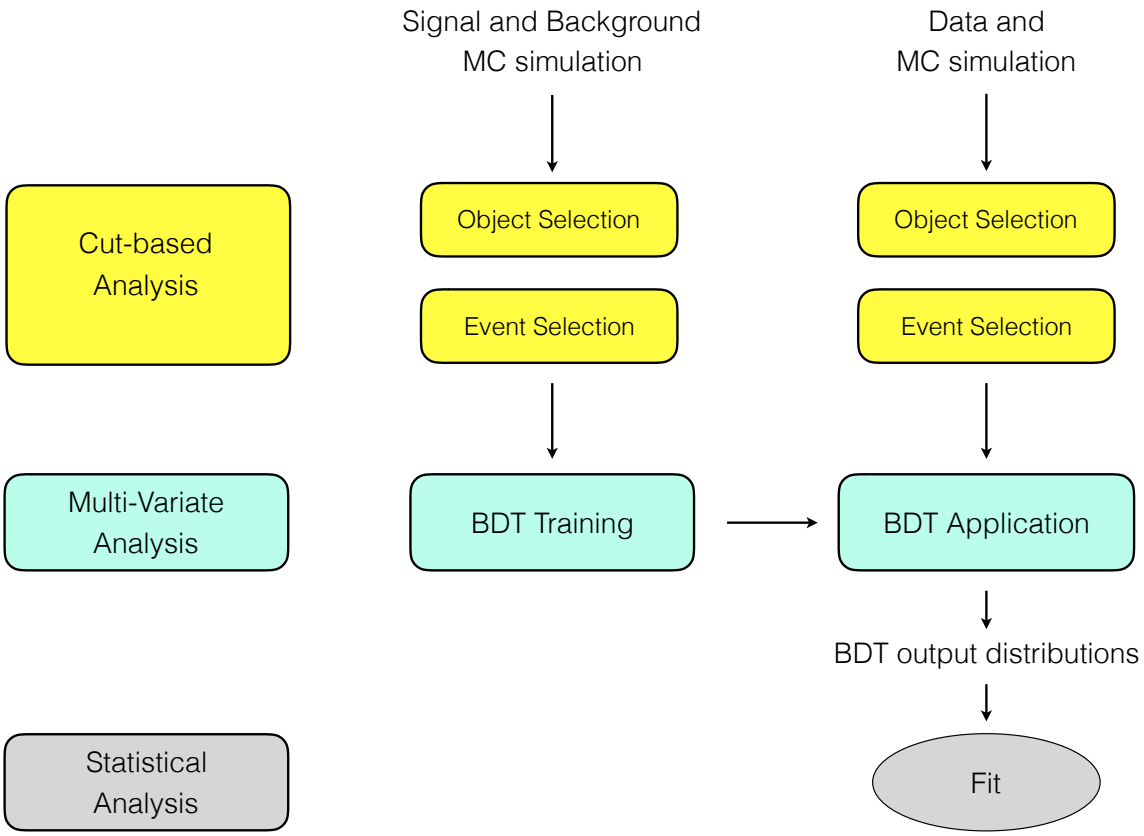


Figure 6.2: Representative diagram of the WH analysis flow.

W^-, resulting in a manifest asymmetry in the electric charge of the gauge bosons produced at pp colliders.

The intermediate state W radiates a H becoming on-shell. The final state W and Higgs bosons decay immediately after production. The W lifetime and the Higgs predicted lifetime are at the order of 10^{-10} ps [37]. For an on-shell W, $m_W = 80.385$ GeV, the most probable decay channels are $W \to q'\bar{q}$ and $W \to \ell\nu_\ell$, with respective branching ratios of 67.41% and 10.86 % [37]. The most important $W \to q'\bar{q}$ type decays are $W \to u\bar{d}/c\bar{s}$ since the decay to $t\bar{b}$ is kinematically forbidden given the top mass of 173.21 GeV [37], and final state configurations involving quark generation mixing are very suppressed as they involve non-diagonal elements of the CKM matrix. The $W \to \ell\nu_\ell$ branching ratio is independent of the lepton flavour and the W decays almost equally frequently to the $e\nu_e/\mu\nu_\mu/\tau\nu_\tau$ final states.

As shown in the $WH \to \ell\nu b\bar{b}$ LO diagram in Figure 6.1(b), the signal process comprehends the WH production and the $H \to b\bar{b}$ and $W \to \ell\nu_\ell$ subsequent decays. The total process cross-section is then given by $\sigma(pp \to WH) \times BR(W \to \ell\nu_\ell) \times BR(H \to b\bar{b})$ and corresponds approximately to 132 fb. For an integrated luminosity of 20.3 fb^{-1}of proton collisions at $\sqrt{s} = 8$ TeV an with $m_H = 125$ GeV, 2632 ± 105 signal events are thus expected.

The final state of the signal has two jets originated by b quarks, an electron or muon and missing energy associated to the weakly interacting neutrino. Due to the prompt decay of the

$m_H = 125$ GeV, $\sqrt{s} = 8$ TeV pp collisions		
Signal Process	$\sigma \times BR$ [fb]	Calculation order
$q'\bar{q} \to WH \to \ell\bar{\nu}_\ell b\bar{b}$	131.7	NNLO QCD NLO EW
$q\bar{q} \to ZH \to \ell\bar{\ell}b\bar{b}$	14.9	
$q\bar{q} \to ZH \to \nu\bar{\nu}b\bar{b}$	44.2	
$gg \to ZH \to \ell\bar{\ell}b\bar{b}$	1.3	
$gg \to ZH \to \nu\bar{\nu}b\bar{b}$	3.8	

Table 6.1: Production cross-section $\sigma \times$ branching fraction and calculation order in perturbative theory of the signal processes for $\sqrt{s} = 8$ TeV pp collisions considering $m_H = 125$ GeV [15].

Higgs and W, these final state objects all share the same primary vertex. The W candidate is reconstructed from its detected decay products: the electron or muon and the neutrino. Since the neutrino 4-momentum can not be fully measured neither can the W. The Higgs candidate is constructed from the two detected b-jets, and therefore the bb system invariant mass distribution should peak around Higgs mass.

The H and W momenta balance in the transverse plane and for this reason p_{T}^{W} equals p_{T}^{bb} for signal events, apart from detector resolution effects and initial and final state radiation. In its rest mass frame, the Higgs decays isotropically because of its scalar nature and the two decay products are emitted back to back conserving momentum. In the laboratory frame, the aperture between the two jets is related to the Higgs momentum, with larger ΔR between jets corresponding to low momentum Higgses. In fact, for large Higgs momentum, the double jet system can become unresolved, and only a unique wide jet is observable. This is not the dominant case in $\sqrt{s} = 8$ TeV pp collisions, and in this analysis the Higgs candidate is always reconstructed from two isolated jets.

The search with the 1-lepton channel is optimised to select $WH \to \ell\nu b\bar{b}$ events but the ZH signal shown in Figures 6.1(a) and 6.1(c) can also contribute despite their selection efficiency being very low, $\mathcal{O}(1\%)$. The final state characteristics of ZH are similar to the ones of the WH signal. $ZH \to \nu\nu b\bar{b}$ events contribute to the 1-lepton analysis when an electron or a muon, originating from a hadron decay inside a jet, is reconstructed as an isolated lepton or when QCD radiation is misidentified as an electron. On their turn, $ZH \to \ell\bar{\ell}b\bar{b}$ events can pass the 1-lepton selection if one of the leptons is not reconstructed, resulting in fake transverse missing energy.

The ZH production mode can be initiated by a $q\bar{q}$ or a gg pair. The respective cross-sections times branching fractions in pp collisions at $\sqrt{s} = 8$ TeV are shown in Table 6.1 for $m_H = 125$ GeV [15].

6.1.2 Backgrounds

The relevant backgrounds to the 1-lepton analysis are top pair production, single top, W or Z plus jets, dibosons and multijet production. These physical processes constitute backgrounds to the search since their final state is very similar to the signal one. They will dominate the

$\sqrt{s} = 8$ TeV pp collisions		
Background Process	$\sigma \times BR$ [fb]	Calculation order
Top quark		
$t\bar{t}$	252.89×10^3	NNLO QCD
$t-$channel	87.76×10^3	
$s-$channel	5.61×10^3	
$Wt-$channel	22.37×10^3	
Vector Boson + jets		
$W \to \ell\bar{\nu}$	12.07×10^6	NNLO QCD
$Z/\gamma* \to \ell\bar{\ell}$	1.24×10^6	
$Z/ \to \nu\bar{\nu}$	6.71×10^6	
Diboson		
WW	52.44×10^3	NLO QCD
WZ	9.241×10^3	
ZZ	3.171×10^3	

Table 6.2: Production cross-section $\sigma \times$ branching fraction and calculation order in perturbative theory of the background processes for $\sqrt{s} = 8$ TeV pp collisions [15]. The fb unit is chosen to represent the order of magnitude of the signal processes cross-section.

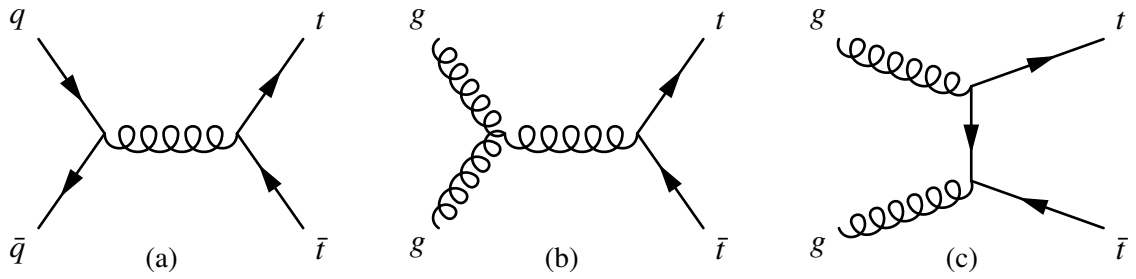


Figure 6.3: Dominant diagrams of the (a) $q\bar{q} \to t\bar{t}$ and (b/c) $gg \to t\bar{t}$ processes at LO.

selection outcome because their production cross-section is several orders of magnitude larger than the WH production, as Table 6.2 shows. For this reason, the WH search is dependent on the precise modelling and normalisation of backgrounds. This fact led to the creation of special control regions, with selection criteria designed to choose a phase space enriched in precise backgrounds. With this approach, it is possible to control their modelling by MC and constrain normalisations with data. Nevertheless, the uncertainty associated with background normalisation is still an important contribution to the analysis systematic uncertainties.

Figure 6.3 shows schematic diagrams of the $t\bar{t}$ production, a pure QCD process that can be initiated by a quark anti-quark pair or gluon-gluon. At $\sqrt{s} = 8$ TeV pp collisions, its production rate is 3 orders of magnitude greater than the signal. The top lifetime is very short, predicted to be only 5×10^{-25} s, so it decays immediately even before the quark hadronisation process starts. The decay involves almost always the weak charged current: $t \to Wq$, where q is a down-type quark. Therefore, its width is proportional to the $|V_{tq}|$ element of the CKM matrix making the $t \to Wb$ decay remarkably favoured: the $BR(t \to Wb)$ is nearly 96 % [37]. Thus, $t\bar{t}$ production will mostly result into $WWbb$. When both W bosons decay leptonically, the $t\bar{t}$ final state is

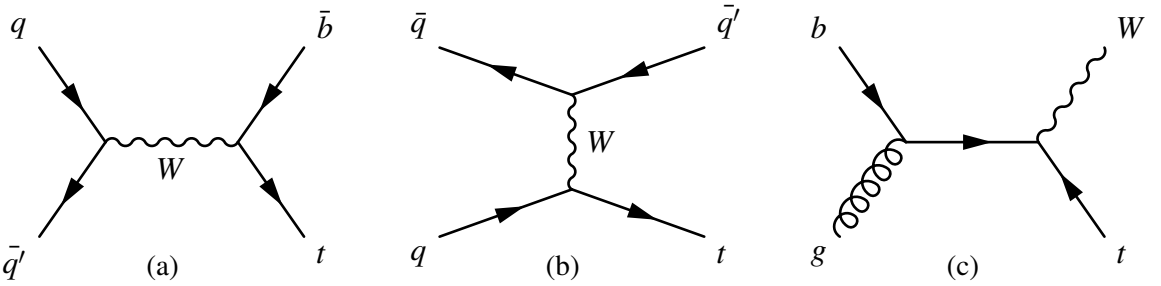


Figure 6.4: Dominant diagrams of the (a) $s-$channel, (b) $t-$channel and (c) $Wt-$channel single top processes at LO.

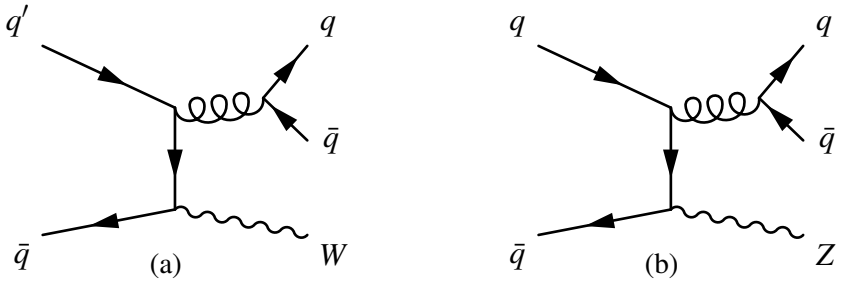


Figure 6.5: Dominant diagrams of the (a) $W + jets$, (b) $Z + jets$ background processes at LO.

composed of two charged leptons, large $E_{\mathrm{T}}^{\mathrm{miss}}$ and two $b-$jets that can fake the signal events if one of the leptons is mis-reconstructed. On the other hand, when one W decays leptonically and the other hadronically it results in one lepton, $E_{\mathrm{T}}^{\mathrm{miss}}$, two $b-$jets and two other jets. If one of these additional jets is lost, for instance through the very forward region of the detector or due to reconstruction inefficiency, the event ends up in the three jets signal region. This region is, in fact, very pure in $t\bar{t}$ and crucial to control and constrain the normalisation of this background.

The single top production can be separated in three main mechanisms as depicted in Figure 6.4: $t-$, $s-$ and $Wt-$channels, leading to different experimental signatures in the detector. The $t-$ and $s-$ channels nomenclature allude to the Mandelstam variables encoding the 4-momentum conservation in $2 \to 2$ scatterings, while the $Wt-$channel simply reflect the hard scatter products. As shows Table 6.2, the $t-$channel is the more relevant among the three options, followed by the $Wt-$channel. Concerning the final state, the $s-$channel is expected to have the highest selection acceptance: if the W decays into a charged lepton and a neutrino, it has the same final state as the signal. When the final state quark of the $t-$channel is a bottom quark, it also corresponds to the signal signature. In the $Wt-$channel diagram at LO, two W bosons come out of the reaction, and the final state configuration is very similar to $t\bar{t}$, except for having only one $b-$quark. At NLO, $Wt-$ production has an interference with top pair production at LO, that needs to be resolved at simulation-level, and is a potential source of uncertainty on the Wt modelling. Section 7.1.4 discusses this topic.

W or Z production accompanied by jets, collectively called V+jets, is a highly relevant background to the analysis. The relevance is in part due to the very large production cross-section, roughly six orders of magnitude greater than the signal, as shown in Table 6.2, but also due to the high resemblance with the signal processes when the vector gauge bosons decay to

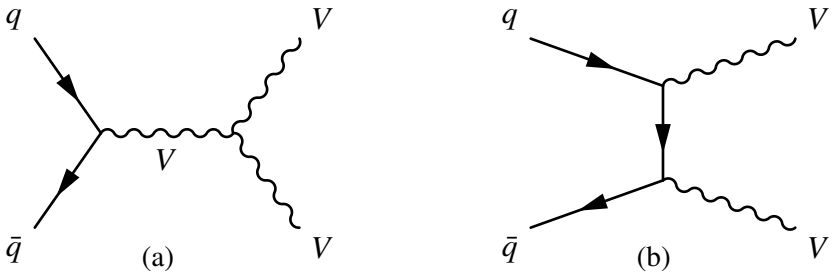


Figure 6.6: Dominant diagrams of the dibosons production at LO.

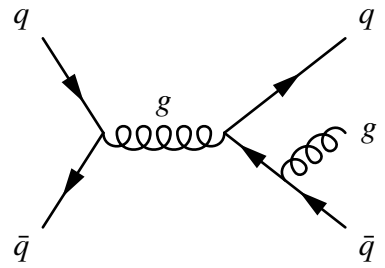


Figure 6.7: Example of a diagram of the multijet background process with three final state jets.

leptons. Figure 6.5 shows the dominant production modes of V+jets. W+jets impact more the 1-lepton channel analysis, while Z+jets are more significant for the other channels. $V+b$ jets have higher selection acceptance due to the quark flavour nature. Despite not having jets originated by $b-$quarks in the final state, the $V+c/$light jets contribution is non-negligible. On one hand, the mistagging rate of the $b-$tagging algorithm used in the analysis makes it permeable to light and $c-$jets. On the other hand, the lighter the final state quark, the higher its production rate, for more phase space is available. V+jets events can, however, be partially distinguished from the signal ones since the two b-jets from the signal result from a resonance decay and are, therefore, correlated.

Continuum production of W/Z boson pairs will also contribute to the 1-lepton channel background. Figure 6.6 illustrates the main processes giving rise to dibosons. In the WW case, when one of the W bosons decays hadronically and the other leptonically, the detector signature is similar to the signal one. The same happens in Z pair production, with one of the bosons decaying in two b quarks and the other decaying leptonically if one of these final leptons is not detected, leading to fake missing transverse energy. WZ has the same final state as the signal when the W decays to $\ell\nu$ and the Z hadronically.

All the backgrounds mentioned above are simulated to predict their contribution in the total data sample, with the exception of multijet production. Multiple jets result from pure QCD interactions, involving only the strong force, where quarks and gluons are the unique outgoing particles. An example of a mechanism underlying the production of this background is sketched in Figure 6.7. In this diagram, three jets would result in the final state, but many can arise from different scattering configurations.

Among the principal hard scattering processes occurring at the LHC pp collisions, this is probably the least known one and is therefore very difficult to model with event generators.

Involving only strong interacting particles, the perturbative QCD order implemented in the generator and parton shower models have a strong impact on the multijet simulation. NLO leads to very different descriptions of the jet multiplicity and flavour composition of the final state, for instance. But multijet production is the most probable process to happen in hadron collisions, and so, very relevant to different physics analyses. Even for the WH search where one charged lepton is required, and the multijet events are unlikely to fullfil all the selection criteria, their relevance is far from being neglected. For these reasons, data-driven methods, consisting of obtaining a sample of multijet events from real data, are usually favoured over MC simulation to evaluate the multijet contribution. Essentially, this is what is done in this Higgs search. For the 1-lepton channel, multijet events arise from jets misidentified as electrons and from light flavoured hadron decays to leptons, the latter contributing more to the 1-muon sub-channel. Section 6.4.3 contains a detailed description of the data-driven method used to extract the multijet background from data.

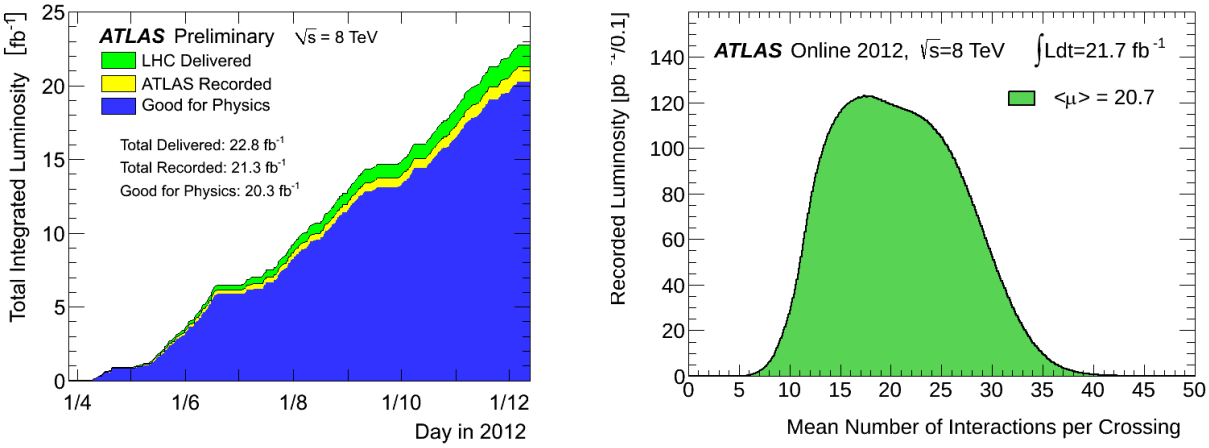
6.2 Data and Simulation samples

6.2.1 Data

The analysis reported in this thesis uses the data set taken by the ATLAS detector from April to December 2012. It corresponds to the integrated luminosity of 20.3 fb^{-1}of LHC proton collisions at $\sqrt{s} = 8$ TeV centre-of-mass energy. The ATLAS data taking efficiency can be evaluated from Figure 6.8(a) showing the integrated luminosity in function of time in 2012 as delivered by the LHC and recorded by ATLAS. The detector recorded 21.3 fb^{-1}of data out of the 22.8 fb^{-1}delivered by the LHC. However, not all the recorded data meet the quality requirements for physics analysis, and therefore only 20.3 fb^{-1}are used in the $WH \to \ell\nu b\bar{b}$ search.

During this period, the maximum instantaneous luminosity reached was 8×10^{33} cm^{-2}s^{-1}, and the LHC ran with proton bunches crossing every 50 ns. The high luminosity implied an unprecedented large amount of in-time collisions pile-up and constituted a major challenge to physics analyses. As Figure 6.8(b) shows, the average number of interactions per bunch crossing during the LHC 2012 run was 20.7 but reached the maximum value of 40. The 50 ns bunch spacing contributed to non-negligible effects of out-of-time pileup, a phenomenon related to the detection of particles that are products of interactions happening on adjacent bunch crossings, leading to event number misassignment and incorrect energy measurements in case of particle overlap.

The experimental data set analysed in the 1-lepton search described here is conducted only for data events passing electron or muon triggers, i.e., to the Muon and Egamma data streams as defined before in Chapter 3 Section 3.2.4. Particular care has to be taken in the case of event overlap between streams: only the Egamma stream event is used when that event is recorded also in the Muon data stream.



(a) (b)

Figure 6.8: (a) Integrated luminosity as a function of time in 2012 for pp collisions at $\sqrt{s} =$ 8 TeV centre-of-mass energy as delivered by the LHC and recorded by the ATLAS detector. (b) Luminosity recorded by ATLAS as a function of the number of interactions per bunch crossing at the LHC. Taken from [28].

6.2.2 Simulation

The ATLAS event simulation uses Monte-Carlo (MC) methods to generate events, simulate the interaction of particles with the detector, their energy deposition in the sensitive material, and the digitisation of the output electronic signals [58]. This provides the event observables as measured in real data and access to the MC truth values of physical quantities as simulated by the generator.

Several generators are available to simulate the pp collisions at the LHC. The simulation of the interaction of out-coming particles with the ATLAS detector is based in GEANT4 [59] and includes the full description of the detector. From that point on, a MC event is reconstructed as a real data event by the ATLAS trigger and offline reconstruction chain.

The MC method is a numerical algorithm that randomly generates variables according to their probability density functions. In the case of high energy particle collisions, these can represent for example the differential cross-sections for a process to take place, the parton distribution functions inside two protons before colliding or the angular separation of the products of a resonance decay. In all the cases, they constitute direct experimental measurements or are parametrised by models explaining observations.

Event generation

Typically, the generation of a collision event involves the following steps, and in every step of this chain the 4-momentum, spin, charge and colour are conserved [60, 61]:

- Hard process or inelastic collision simulation;
- Decay of short-lived resonances produced above;

- Initial and final state radiation emission by partons before and after the hard scatter, respectively;
- Underlying event simulation;
- Fragmentation of outgoing quarks and gluons where hadronisation takes place;
- Decay of unstable hadrons produced above;
- Simulation of multiple interactions in the event.

Hard process The hard process simulation is typically optimised to "$2 \to 1$" and "$2 \to 2$" processes, where two initial state particles produce one or two final state particles, respectively. For hadron collisions in the Standard Model context, the most relevant processes are the $qg \to qg$ hard QCD process, the top quark production $gg \to t\bar{t}$, the W/Z production $q\bar{q} \to Z$ and $q'\bar{q} \to W^{\pm}$, the Drell-Yan process $q\bar{q} \to \gamma^*/Z^*$ and the Higgs boson production $gg \to H$, $q'\bar{q} \to WH$ and $q\bar{q} \to ZH$.

It results from the convolution of the partonic cross-section, codified in matrix elements for each specific physical process, with the PDFs of the colliding protons. The partonic cross-section can be calculated in different orders in the QCD perturbation theory as discussed. The LO is the most commonly used, but generators with NLO matrix elements also exist.

Decays The decay is generated for the unstable products of the hard scatter, such as gauge bosons, the top quark or the Higgs boson, according to the branching ratios. The decay products are usually quarks and leptons or yet other resonances, for which the decay is in turn generated as, for instance, in the case of $H \to ZZ \to \ell\bar{\ell}\ell\bar{\ell}$. The spin nature of the particles is coherently propagated through the decay chain and the angular configuration of the final decay products is properly established.

Initial and final state radiation The initial and final state radiation is necessary to more realistically model the multijet structure of events involving quarks and describe the substructure of jets. To every initial and final state quark or gluon, a parton shower (PS) develops starting with branchings as $q \to qg$ or $g \to gg$. For hadron collisions, final state electrons also undergo the similar showering process: $e \to e\gamma$.

Essentially, the PS is a method that adds higher-order effects of the perturbation theory to the hard scatter, so special care has to be taken when matching the PS simulation with the matrix elelements of the hard event generator to avoid double counting of diagrams.

Underlying event The reactions between partons that do not participate on the hard scatter is often simulated as soft $2 \to 2$ scatterings of partons from each beam.

Fragmentation and decay of short-lived hadrons The fragmentation or hadronisation step is the least understood within the event generation chain. Here the coloured outgoing quarks and gluons from the initial and final parton shower hadronise into colourless hadrons as pions, for

instance. In the case of a spray of hadrons, this process gives rise to the jet structure. Unstable hadrons decay until only stable particles are left in the jet. Several approaches serve the purpose of generating the hadronisation, such as the string model and the cluster model described before in Chapter 2 Section 2.2.

Multiple interactions The number of multiple interactions per bunch crossing is very relevant in high luminosity colliders. For the LHC this phenomenon has particular impact on physics. The collisions pile up when several pp interactions happen in the same event. This is additionally generated and needs to be precisely tuned.

Generators Several generators exist that differ on the approaches of specific steps of the generation, as the hadronisation model, the calculation of the matrix elements or the parton shower algorithm. Some generators have a general purpose, based on their ability to simulate all the mentioned aspects of the event, but others have a specific purpose and are only intended to generate particular steps of the chain, and therefore need to be combined with others to fully simulate the event.

PYTHIA [60, 62] is one of the best examples of the former kind of generators. Widely used in HEP, it implements the hard scatter process in LO in QCD and uses the string model for hadronisation. SHERPA [61] and HERWIG [63] are also general purpose generators, LO in QCD, but offer the alternative cluster model for hadronisation.

On the other hand, POWHEG [64] has the specific purpose of generating the hard-scatter, and have NLO calculated matrix elements providing a better description of most of the SM processes. It needs to be interfaced with the PYTHIA or HERWIG parton shower and hadronisation models. Other examples of hard-scattering generators include ACERMC [65] or MC@NLO [66]. PHOTOS [67], for instance, generates QED corrections to decays of resonances.

For physics analysis, it often happens that only specific processes are interesting or relevant. This can be a particular decay of a particle or a specific flavour of the fermions produced at the hard scatter, for instance. To accommodate these demands, the generators use final state filters to reject the undesired events, enriching the sample with the wanted ones. However, these MC filters have associated inefficiencies that have to be accounted for when the samples are analysed.

Interaction with the detector

Following event generation, every long-lived particle produced is propagated through the detector. Their interaction with the detector material volumes and magnetic fields is simulated and so is the electrical output signal digitisation. The material volumes of ATLAS detector are fully described using GEANT4 [59]. For repeated structures, a single physical volume can be defined and reproduced in space to match the whole detector subsystem. This volume

parametrisation is implemented in the GEANT4 [59] description of the ATLAS detector for the calorimeters, but it can not, however, compromise the detail of the detector description, for that is needed to ensure the best modelling of physical observables. In fact, the detector simulation is detailed to the point of incorporating known real misalignments, specially for the muon chambers case.

In addition to this, more dynamic information about the detector can be included in the simulation by accessing databases. These databases store information related to dead channels in the detector, temperature measurements, or calibration constants for a specific run. The digitisation process is also simulated. It converts the electrical pulses resultant from energy deposits in the sensitive materials of the detector into digital bit streams. Real effects such as channels noise or gain fluctuations are also included to emulate the detector behaviour as faithfully as possible.

The interaction of particles with the detector materials is the most time consuming step of the whole event simulation chain, with about 80% of the total simulation time wasted to simulate particles traversing the calorimeter. It is in fact impossible to generate samples with enough statistics for all the processes needed by physics analyses with this full simulation scheme. To overcome the problem, ATLAS uses a fast simulation approach called ATLFASTII [58], where the simulation of the energy deposited by single particle showers in the calorimeters uses a high granularity energy parametrisation determined from full simulation of photons and pions. This speeds up the computation time by a factor of 10.

Corrections to simulation

Despite the effort to realistically simulate high-energy collision events, the procedure is complex and involves many steps and approximations. Therefore, data and prediction by MC are carefully compared in reference data samples to derive corrections to MC when needed.

The VH analysis uses different MC event generators, depending on the signal and background process. Multiple interactions are simulated separately with PYTHIA8 plugged with the MSTW2008LO [68] parton distribution functions (PDFs) describing the interacting protons, and the external A2 [69] tunes to the hadronisation and parton shower models of PYTHIA8. The pile-up events are then overlaid to the main simulation of pp collision events. However, the real data luminosity profile used to generate the multiple interactions was unknown at the time the pile-up samples were generated. This resulted in a different spectra of the mean number of interactions per buch crossing $< \mu >$ between data and MC, as Figure 6.9 shows. The MC events are re-weighted to correct this feature.

For similar reasons, the longitudinal displacement of the primary vertex from the origin of the coordinate system in MC also exhibits differences with respect to the real data spectrum. Event weights are attributed to MC to reproduce the real data.

Other corrections are also applied to MC to account for generator-level mis-modelling of variables relevant to the analysis as detailed in what follows.

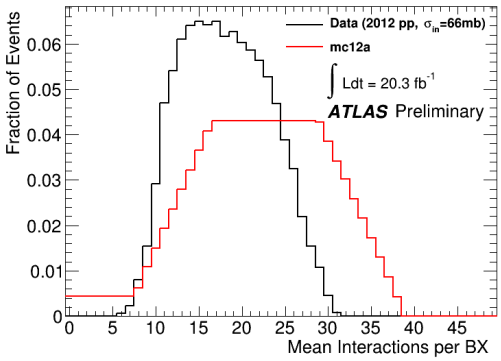


Figure 6.9: Mean number of interactions per bunch crossing $< \mu >$ weighted to the luminosity for MC and data. For real pp collisions, $< \mu >$ was determined from the luminosity assuming an inelastic cross-section of 66 mb. Taken from [70].

Signal Process	Generator	
$q'\bar{q} \to WH \to \ell\bar{\nu}_\ell b\bar{b}$ $q\bar{q} \to ZH \to \ell\bar{\ell}b\bar{b}/\nu\bar{\nu}b\bar{b}$	PYTHIA8	LO
$gg \to ZH \to \ell\bar{\ell}b\bar{b}/\nu\bar{\nu}b\bar{b}$	POWHEG+PYTHIA8	NLO

Table 6.3: Generators used in the signal event simulation.

Simulation of signal events

Signal events were generated for a Higgs boson mass of 125 GeV decaying to $b-$quark pairs. A list of the generators used is presented in Table 6.3. In both WH and ZH production modes, MC events have a leptonic decaying W/Z filter. Decays to τ are considered since they can end up in the muon or electron channel when decaying to lighter leptons through $\tau \to \ell\bar{\nu}_\ell\nu_\tau$ ($BR = 17.41$ %) [37]. The quark-initiated signal events were generated by PYTHIA8 [62] with the CTEQ6L1 [71] PDFs. AU2 [69] tunes to the parton shower, hadronisation and underlying event interactions models derived to describe better the ATLAS data were used. The PHOTOS [67] generator was used to generate QED final state radiation. The PYTHIA8 event generation is done at LO in perturbative QCD and QED, but the signal samples are normalised to data luminosity using the partonic cross-section calculated at NLO in QCD and including NNLO QCD and NLO EW corrections. The dependence of the LO and NLO cross-section on the p_{T} of the vector boson, p_{T}^V, differs. A differential NLO EW correction is applied to the qq-initiated samples as a function of p_{T}^V to account for this feature. The signal events initiated by gluon fusion are generated at NLO in QCD using POWHEG [64] with CT10 [72] PDFs interfaced with the PYTHIA8 [62] AU2-tuned showering and hadronisation models.

Background simulation

Table 6.4 summarises the list of background generators used in the $WH \to \ell\nu b\bar{b}$ analysis. Their detailed references and properties can be found in Appendix B.

Top pair and the $s-$ and $Wt-$channel single top events were generated at NLO in QCD

Background Process	Generator	
Top quark		
$t\bar{t}$		
$s-$channel	POWHEG+PYTHIA6	NLO
$Wt-$channel		
$t-$channel	ACERMC+PYTHIA6	LO
Vector Boson + jets		
$W \to \ell\bar{\nu}$		
$Z/\gamma* \to \ell\bar{\ell}$	SHERPA 1.4.1	LO
$Z/ \to \nu\bar{\nu}$		
Diboson		
WW		
WZ	POWHEG+PYTHIA8	NLO
ZZ		

Table 6.4: Generators used to simulate the background processes.

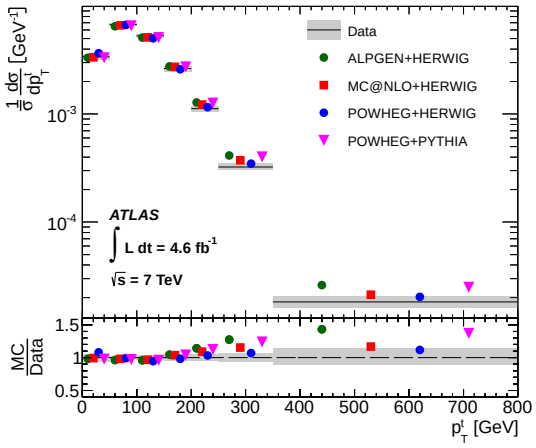
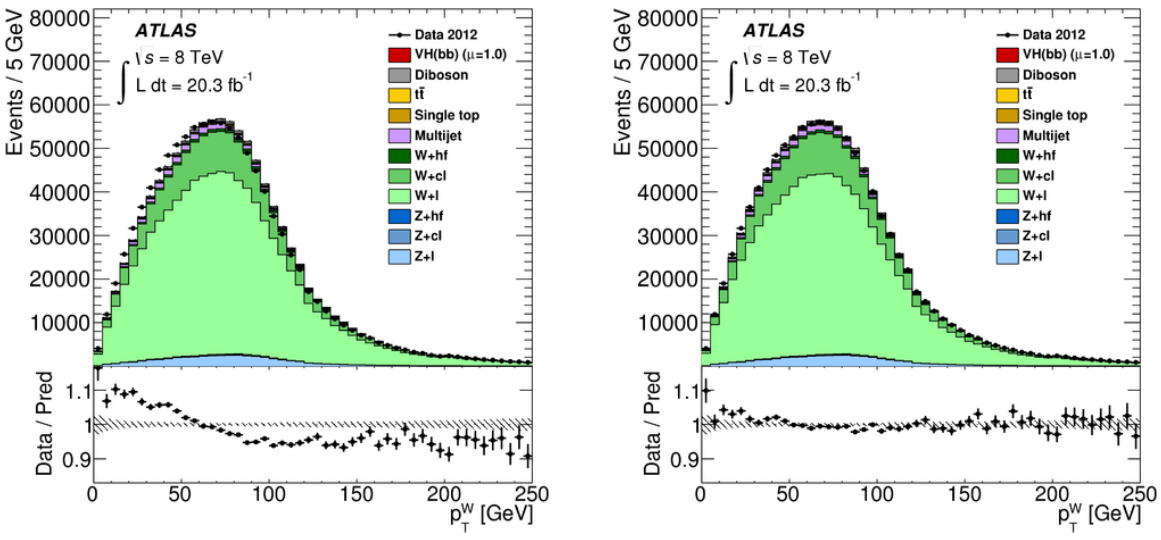


Figure 6.10: $t\bar{t}$ differential cross-section measured from data and for different generator predictions as a function of p_{T}^{t}. Taken from [73].

using the POWHEG [64] generator and CT10 [72] PDFs as input. The parton shower and hadronisation is obtained from the PYTHIA6 [60] implementation using CTEQ6L1 [71] PDFs and the PERUGIA2011C [69] tunes to underlying event and parton shower. The $t-$channel production mode of single top is simulated at LO in QCD with ACERMC [65] interfaced with the PERUGIA2011C-tuned PYTHIA6 [60]. The $t\bar{t}$ simulation has a filter imposing at least one leptonically decaying W with efficiency very close to unity. The NLO generation with POWHEG [64] results in a simulated top p_{T} spectrum for $t\bar{t}$ events that is different than the one observed in data. Top quarks are predicted to have larger momentum than observed in real data as shown in Figure 6.10 [73]. Since this impacts the p_{T}^{W} of reconstructed top pair events, which is a particularly relevant variable to the analysis, the mis-modelling is corrected by adding (subtracting) statistical weight to events with lower (larger) top p_{T}.

The W/Z+jets samples are generated with the SHERPA [61] generator, implemented in LO in perturbative QCD using the CT10 PDFs with massive b and c quarks representations.



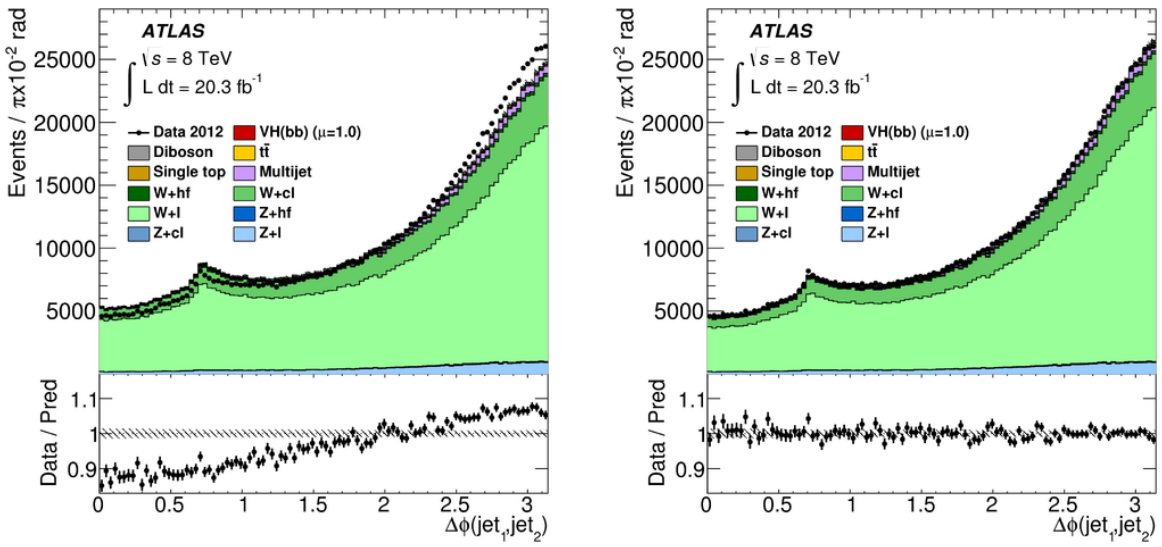
(a) (b)

Figure 6.11: p_{T}^{W} distribution for data and MC prediction in a W+jets control region with no $b-$tagged jets in the muon sub-channel (a) before and (b) after applying the $\Delta\phi(j_1, j_2)$ correction. The shadowed band in the ratio plot indicates the size of the statistical uncertainty. Taken from [50].

Samples enriched in b, c or light jets are generated independently using filters to select b, c or light flavoured hadrons, respectively. In order to increase the statistics in the high p_{T}^{V} spectra region, events are further filtered and separated according to the following p_{T}^{V} intervals: $\{[0,40[, [40,70[, [70,140[, [140,280[, [280,500[, [500,+\infty]\}$ GeV. The generation process is also split according to the flavour of the leptons from the W/Z decay.

Modelling corrections are applied to the W/Z+jets SHERPA samples to account for discrepancies observed between data and MC: SHERPA generates a much harder p_{T}^{V} spectrum than observed in real data, as shown in Figure 6.11(a) for the case of a W+jets-enriched sample obtained by requiring events with no $b-$tagged jets in the muon sub-channel. It was found that this feature was strongly correlated with a clear mis-modeling of the ϕ separation between the two leading jets, $\Delta\phi(j_1, j_2)$, of the W+jets samples, exhibited in Figure 6.12(a). A $\Delta\phi(j_1, j_2)$-based correction derived from the data/MC ratio is applied to the MC W/Z+jets sample in the form of event weights. The correction is only applied to W+light or cl flavoured jets, defined at MC truth level, because these are the dominant composition of the total W+jets samples, without a significant amount of bl and bb flavoured jets events to deduce a similar correction for these samples. Additionally, since the correction also improves the modelling of the W+jets samples in the remaining analysis regions, it is used across them and for all the VH channels. Its performance can be evaluated from Figures 6.11(b) and 6.12(b) showing that the correction not only fixes the $\Delta\phi(j_1, j_2)$ prediction but also the simulated p_{T}^{W} spectrum. The simulated Z+jets sample is re-weighted in a similar manner, but in this case, a p_{T}^{Z}-based correction was instead justified for events containing two b-jets.

The prediction of the dibosons background uses the NLO POWHEG generator with



(a) (b)

Figure 6.12: $\Delta\phi(j_1, j_2)$ distribution for data and MC prediction in a W+jets control region with no $b-$tagged jets in the muon sub-channel (a) before and (b) after applying the $\Delta\phi(j_1, j_2)$ correction. The shadowed band in the ratio plot indicates the size of the statistical uncertainty. Taken from [50].

CT10 PDFs. Showering and hadronisation techniques were implemented with PYTHIA8 [62], configured with the ATLAS underlying event tune AU2 [69].

In the VH combined analysis, all the processes except multijets are initially normalised to the data luminosity using the cross-section predicted by theory. This is essential to design the event selection in such a way that its acceptance to signal events is maximised at the same time that the background is minimised. Except for dibosons, for which the cross-section used corresponds to the NLO generator final estimate, all the other backgrounds are normalised to the NNLO predicted cross-section taken from [74], [75], [76], [77] and [78] for the $t\bar{t}$, single top $s-$, $t-$ and $Wt-$channels, and W/Z+jets, respectively. Finally, the normalisation of the main backgrounds, $t\bar{t}$ and W/Z+jets, are constrained through the binned likelihood fit of the signal and background prediction to the data, as described in Chapter 7.

6.3 Object Selection

The search for $WH \to \ell\nu b\bar{b}$ events, with $\ell = e, \mu$, requires one electron or muon, large $E_{\mathrm{T}}^{\mathrm{miss}}$ and at least two jets identified as $b-$jets by the $b-$tagging algorithm. The reconstruction and identification of these objects, along with the methods used for their calibration were already presented in Chapter 4. This section describes the criteria used to select the final state objects that will be used to reconstruct the Higgs signal. The criteria are motivated by quality and efficiency issues and additionally aim to clean the event from objects that do not come from the primary vertex, associated with the hard scatter vertex where the signal process originated.

The object selection is complex as it deals with numerous sequential cuts and different

types of objects. Its correct implementation in the analysis code must, therefore, be cross-checked to validate any outcoming result. To do so, the number of objects meeting each selection criterion for a reference sample of signal events was compared among all the groups participating in the VH analysis. The results converged only after a few iterations and will be presented in this section. Finally, the reconstruction of the Higgs and W candidates from a signal-like event is also addressed.

6.3.1 Electrons

The WH analysis requires one central isolated lepton in the event. In case the event has electrons, these have to be calibrated. The electron energy calibration consists of calibrating the energy scale of electrons in real data and smearing the energy resolution of the simulated electrons as discussed in Section 4.2. Additionally, to account for reconstruction and identification efficiency differences between data and MC, the MC events with electrons are corrected using the event scale factors shown in Figures 4.3 and 4.3 of Section 4.2.

The electron selection defines two sets of electrons: loose and signal electrons. Loose electrons are only used to veto events with more than one electron and are defined as follows:

- Must be reconstructed in the central region of the detector, approximately matching the ID coverage, by both the calorimeter and the tracker.
- $E_{\mathrm{T}} > 7$ GeV to meet the required cluster efficiency reconstruction of 97%.
- VeryLooseLH identification based on the likelihood method described in Chapter 4.
- Must have a loosely isolated track: the sum of the p_{T} of all tracks inside a cone of radius 0.2 around the electron track must be smaller than 10% of the electron p_{T}.

Signal electrons meet tighter selection criteria, presented below, and therefore have higher quality. These are used in the analysis to reconstruct the W signal candidate.

- $E_{\mathrm{T}} >$25 GeV is required in order to match the maximum efficiency of the electron trigger, shown in Figure 4.5 of Section 4.2.
- VeryTightLH identification condition.
- Track isolation criterion is tightened to 4%.
- Calorimeter isolation cut: the E_{T} inside a cone of radius 0.3 around the electron extrapolated track should be smaller than 4% of the electron E_{T}.

The isolation requirement has a different efficiency for data and MC. To account for this, the MC is corrected through an event re-weighting that does not change the event weight more than 3%.

This selection and the number of electrons that fulfil each criterion for the LIP analysis code are shown in Table 6.5. The different groups participating in the VH analysis validated

Selection	Number of selected electrons
None	3388073
Calorimeter and Tracker	1293957
VeryLooseLH	156859
$E_{\mathrm{T}} > 7$ GeV	1546277
$\lvert\eta\rvert < 2.47$	2293302
Track isolation	868224
Loose	100885
$E_{\mathrm{T}} > 25$ GeV	365261
VeryTightLH	84003
Track isolation	832329
Calorimeter isolation	72460
Signal WH	49042

Table 6.5: Electrons selection criteria and number of selected electrons after applying each selection condition obtained with the LIP analysis code. Cuts are applied independently except for the Loose and Signal selection that correspond to the logical AND of all their previous conditions. A signal sample with 300000 events was used as reference. The numbers were compared to the outcome of other groups codes and an absolute agreement was found for each condition.

the codes performing the electron selection by comparing the number of electrons passing each cut with a reference sample of signal events. A 100% agreement was found between the groups after a few iterations.

6.3.2 Muons

The muon sub-channel of the WH analysis requires one isolated signal muon of high momentum. The requirements on muon reconstruction quality are listed in Table 6.6 and presuppose that the simulated muon momentum is corrected for scale and smearing as described previously in Chapter 4. The criteria depend on the muon type presented before: combined (CB), segment-tagged (ST), stand-alone (SA), silicon-associated forward (FW) and calorimeter (Calo).

The analysis strategy is to use high-quality muons as signal muons to reconstruct the W candidate, and loose muons are used to veto events with multiple leptons. Loose muons are defined as follows:

- **ID hit cuts** Muons reconstructed within the acceptance of the ID (CB, ST and Calo), must have a minimum number of hits on the ID and a maximum number of dead sensors, as Table 6.6 details.
- **Impact parameter** Muons resulting from the W boson decay have an origin close to the main PV. Therefore, cuts on the transverse (d_0) and longitudinal (z_0) impact parameters are applied as indicated in Table 6.6 for the muon categories that have an ID track.
- η **limits** CB and ST muons are used within the whole muon spectrometer η coverage

	Selection	Number of selected muons
CB+ST muons	None	308776
	Number of pixel hits ≥ 1	283019
	Number of SCT hits ≥ 5	285751
	Number of Si holes ≤ 2	288450
	Number of TRT hits ≥ 6	284841
	$d_0 < 0.1$ mm	189286
	$z_0 < 10$ mm	226882
	$\|\eta\| < 2.7$	288450
	$p_T > 7$ GeV	152297
	Track isolation	183454
	Loose	97162
	$p_T > 25$ GeV	87240
	$\eta < 2.5$	296604
	Tight track isolation	203093
	Calorimeter isolation	133084
	Signal WH	64565
SA+FW muons	None	23856
	$2.5 < \|\eta\| < 2.7$	7926
	$p_T > 7$ GeV	4778
	Loose	2704
Calo muons	None	256436
	Number of pixel hits ≥ 1	322020
	Number of SCT hits ≥ 5	322020
	Number of Si holes ≤ 2	322020
	Number of TRT hits ≥ 6	284140
	$\|\eta\| < 0.1$	13264
	$p_T > 20$ GeV	103821
	$d_0 < 0.1$ mm	220896
	$z_0 < 10$ mm	198821
	Track isolation	244383
	Calo overlap removal	229096
	Loose	1653

Table 6.6: Muons selection criteria and number of selected muons after applying each selection condition obtained with the LIP analysis code. Cuts are applied independently except for the Loose and Signal selection that correspond to the logical AND of all their previous conditions. The number of pixel and SCT hits is subtracted from the number of dead sensors (Si holes) but this must not overcome 2. A signal sample with 300000 events was used as reference. The numbers were compared to the outcome of other groups codes and an absolute agreement was found for each condition.

while SA and FW muons are used in the region covered by the MS but not instrumented with the inner tracker ($2.5 < |\eta| < 2.7$). Calorimeter muons are used to recover muon reconstruction efficiency within $\eta < 0.1$.

- **p_{T} thresholds** Loose muons are required to have $p_{\mathrm{T}} > 7$ GeV since from this point on the reconstruction efficiency is already very close to 99%. Exceptionally, Calo muons must have $p_{\mathrm{T}} > 20$ GeV, because only with this threshold their reconstruction efficiency is greater than 94%.
- **Track isolation** Muons tracked by the ID, must have a loosely isolated track: the p_{T} sum of all tracks inside a cone of radius 0.2 centred at the muon track should be smaller than 10% of the muon p_{T}. This reduces background muons coming from jet hadrons decays.
- **Calo overlap removal** Calo muons that are also reconstructed by the CB method are discarded in favour of the CB muons to avoid double counting.

Signal muons are either CB or ST loose muons and fulfill additional requirements:

- $|\eta| <$**2.5** to match the region covered by the ID.
- $p_{\mathrm{T}} >$**25 GeV** corresponding to the un-prescaled muon trigger of lowest p_{T} threshold, that is used in the analysis.
- **Track isolation** is tightened to 4%.
- **Calorimeter isolation** signal muons must be also isolated in the calorimeter: the E_{T} deposited inside a cone of radius 0.3 around the muon combined track should be smaller than 4% of the muon p_{T}.

Every MC event containing a muon receives a weight accounting for the identification and reconstruction efficiency difference with respect to data as described before in Section 4.3. In addition, MC is corrected for efficiency differences with respect to data regarding the isolation cut, through event scale factors, that do not differ more than 5% from the unity.

The absolute number of muons fulfilling each selection criterion, shown in Table 6.6, is compared between the LIP analysis code outcome and other groups participating in the $H \to b\bar{b}$ analysis. An exact agreement within the various groups was obtained.

6.3.3 Jets

The reconstructed calorimeter jets used in the analysis are calibrated as described in Chapter 4 and selected according to a set of quality criteria, summarised in Table 6.7. These aim to reject background jets from cosmic showers, calorimeter noise or from interactions between the beam and residual gas molecules, the beam pipe or collimators, and background jets from pile-up interactions.

As for electrons and muons, two categories of jets, loose and signal, are defined according to the set of selection criteria and fake rejection efficiency. Loose jets obey looser quality

Selection	Number of selected jets
None	1935448
$\|\eta\| < 4.5$	1908125
$p_T > 20(30)$ GeV	933006
$\|JVF\| > 0.5$	1351066
Loose	829169
$\|\eta\| < 2.5$	1528019
Signal	766557

Table 6.7: Jets selection criteria and number of selected jets after applying each selection condition obtained with LIP analysis code. Cuts are applied independently except for the Loose and Signal selection that correspond to the logical AND of all their previous conditions. A signal sample with 300000 events was used as reference. The numbers were compared to the outcome of other groups codes and an absolute agreement was found for each condition.

requirements and are only used to veto events with extra jets. For signal jets the selection is tightened aiming at a large fake jet rejection. The latter are a subset of the loose category and are the candidate objects for the decay products of the Higgs boson.

The pseudorapidity of the jets is limited to $|\eta| < 2.5$ for signal jets and to $|\eta| < 4.5$ for the loose set. Signal jets are therefore central, contained in the region covered by the ID, designed for precision measurements. In particular, the access to the ID information is essential for the b-jet tagging, an issue of central importance for the $H \to b\bar{b}$ analysis.

The jet transverse momentum must be above 20 GeV for central jets and 30 GeV elsewhere. The threshold applied to central jets is related to the difficulty of having a reliable jet calibration at low energy. As seen in Chapter 4 Figure 4.16, calibration constants are not available below 20 GeV. In the forward region the threshold is increased because there is more activity due to minimum bias events coming from proton-proton scattering that produce many jets of softer p_{T}.

To reject jets that do not come from a hard scatter interaction, jets are required to have a jet vertex fraction (JVF) above 50%. The JVF is determined in an event basis for each jet. For a jet i and the main PV, the JVF is defined as the ratio of the scalar sum of the p_{T} of the m tracks spatially matched to the jet i that come from the main PV to the scalar sum of all l tracks matching the jet:

$$\mathrm{JVF}(\mathrm{jet}_i,\mathrm{PV}) = \frac{\sum_m p_{\mathrm{T}}(\mathrm{track}_m^{\mathrm{jet}_i}, \mathrm{PV})}{\sum_l p_{\mathrm{T}}(\mathrm{track}_l^{\mathrm{jet}_i})} \tag{6.1}$$

Figure 6.13 shows a schematic view of the JVF principle for two jets: JVF ranges from 0, when none of the tracks of the jet comes from the main primary vertex to 1 when all its matched tracks come from the main PV. The JVF calculation only considers tracks with transverse momentum above 500 MeV. The cut on JVF is only imposed to jets lying within $|\eta| < 2.4$, where tracks can be fully measured, and for jets with p_{T} <50 GeV since the cut goal is to reject pile-up jets that have a softer transverse momentum spectrum than jets coming from an hard

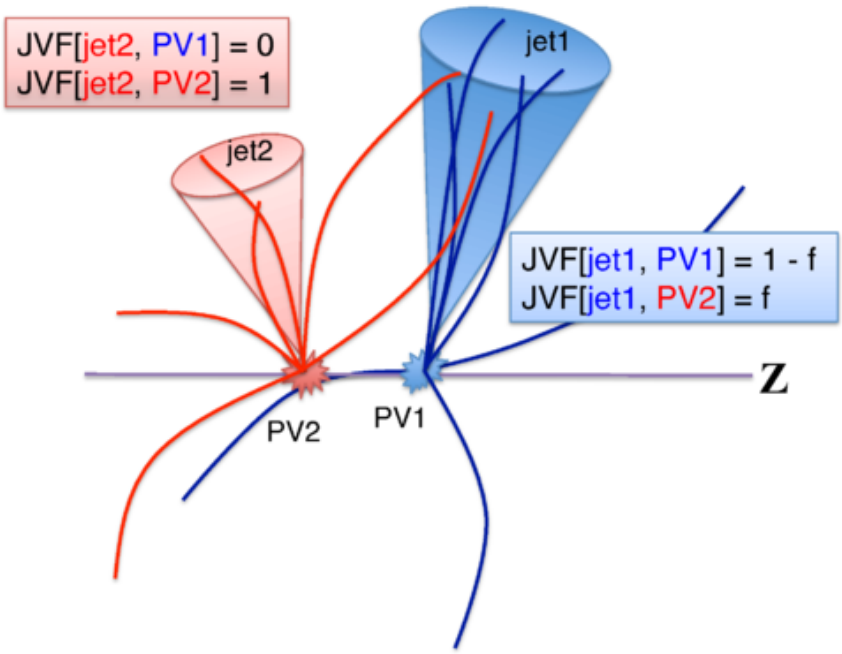


Figure 6.13: Schematic view of the JVF principle for two jets and two primary vertexes. Taken from [45].

Operating Point	MV1c cut	$b-$jet Eff. [%]	$c-$jet Rej.	light-jet Rej.	$\tau-$jet Rej.
Tight	>0.9237	49.99	26.22	1388.28	120.33
Medium	>0.7028	70.00	5.34	135.76	14.90
Loose	>0.4050	79.85	3.04	29.12	6.4

Table 6.8: Cut values and $b-$jet efficiency (Eff.) of the MV1c algorithm for the three operating points used in the $WH \to \ell\nu b\bar{b}$ analysis. Rej. stands for the rejection values for $c-$, light- and $\tau-$jets and corresponds to inverse of the mis-tag rate efficiency. Taken from [79].

scatter.

The number of jets fulfilling each selection criterion is compared between the LIP analysis code outcome and the other groups participating in the $H \to b\bar{b}$ analysis. An absolute agreement is found between the different groups for all the selection conditions.

6.3.4 $b-$Tagging

The MV1c $b-$tagging algorithm, described in Chapter4, is used to select $b-$jets in the $WH \to \ell\nu b\bar{b}$ analysis. For every event, the algorithm determines a value related to the probability for each jet to have origin on a $b-$quark.

Three distinct operating points defined by a cut value on this output value are used. The corresponding $b-$jet efficiency and $c-$ and light-flavour rejection factors are summarised in Table 6.8. They were labelled, in decreasing order of $b-$jet efficiency, as Loose, Medium and Tight $b-$tag. The larger the efficiency, the smaller the rejection factor and in this way, a Tight $b-$tagged sample of jets is purer in truth $b-$jets than a Loose $b-$tagged one. On the Loose working point, where the rejection factors are smaller for non $b-$jets, the MV1c discriminant

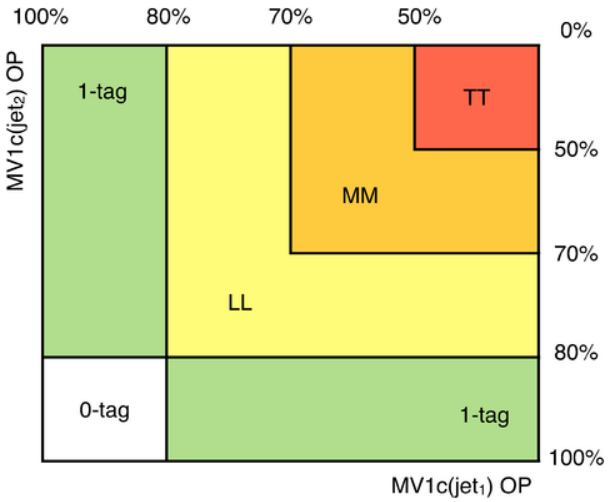


Figure 6.14: $b-$tagging categories of the two signal jets in the analysis, resultant of the MV1c algorithm efficiency points. Taken from [50].

provides a mis-tag rate of approximately 33% and 3.4% for $c-$ and light-jets, respectively. With the Tight $b-$tag, the corresponding rates decrease significantly to 3.8% and 0.07%.

The use of three operating points with independent calibrations is usually referred to as continuous $b-$tagging [80]. Strictly speaking, the calibration is not truly continuous since it is unique for each MV1c output interval, but still it constitutes a fair approximation for most of the applications.

The event selection requires two $b-$jets in the signal region and one or none in the control regions. The events are then classified according to the $b-$tag operating point of the two jets as Tight-Tight (TT), Medium-Medium (MM), Loose-Loose (LL), 1-tag and 0-tag as depicted in Figure 6.14. The categories are disjointed: the TT region corresponds to events with two Tight-tagged jets, the MM events have one Medium-tagged jet and a Medium or Tight $b-jet$ while for the LL category one of the jets passes the Loose $b-$tagging criterion and the other can either be a Loose, a Medium or a Tight $b-$jet. If only one jet is $b-$tagged, regardless of the MV1c working point, the event is classified as 1-tag and if none of the jets is identified as a $b-$jet the 0-tag category is assigned.

Truth $b-$Tagging

An effective $b-$tagging method has, however, some inconveniences. The high rejection power of MV1c for $c-$ and light-jets leads to a significant statistical loss for MC samples containing these jet flavours, as it is the case of W/Z plus $c-$ or light-jets. To recover MC statistics, while keeping the good modelling of all physical processes constituting background to the analysis, a procedure named as truth tagging is used in place of the one described so far, called direct tagging. Although truth tagging is specially motivated by W/Z + light- and $c-$jets, it is applied to all MC, including signal, for coherence motives.

The procedure is only applied to MC events containing jets for which no $b-$hadron is found inside the calorimeter jet cone:

$E_{\mathrm{T}}^{\mathrm{miss}}$ interval (GeV)	LIP	CERN	CPPM	Tsukuba
$[0,90[$	252619	0	0	0
$[90,120[$	25824	-0.01	-0.01	-0.01
$[120,160[$	12556	0	0	0
$[160,200[$	4472	0.11	0.11	0.11
$[200,\infty]$	3528	0	0	0

Table 6.9: Number of events per $E_{\mathrm{T}}^{\mathrm{miss}}$ interval as obtained with the LIP analysis code. The percentage deviation of the number of events per $E_{\mathrm{T}}^{\mathrm{miss}}$ interval as obtained by the other groups codes with respect to LIP, defined as $\Delta = (N_{group} - N_{LIP})/N_{LIP} \times 100\%$, is presented in the last three columns. A signal sample with 300000 events was used as reference.

- If a $b-$hadron is found, it is most likely that the jet is $b-$tagged by MV1c and the direct method is used.
- Otherwise, truth tagging is used. In this method, all jets are $b-$tagged but receive a weight that is proportional to the probability of misidentifying the jet as a b with the direct method. The event weight was determined such that the normalisation and shape of observables obtained using direct tagging are preserved.

6.3.5 Missing Transverse Energy

The $E_{\mathrm{T}}^{\mathrm{miss}}$ is recalculated to take into account the calibration of jets, electrons and muons, using a tool developed by the $E_{\mathrm{T}}^{\mathrm{miss}}$ performance group of ATLAS. The result was cross-checked between the different groups contributing to the WH analysis. Table 6.9 shows the number of events of a reference sample per each $E_{\mathrm{T}}^{\mathrm{miss}}$ interval as obtained by the LIP group analysis code and other groups participating in the analysis. The maximum difference observed is at the per mille level and was considered sufficient to carry on the analysis.

6.3.6 Overlap Removal

The object reconstruction and identification used in the ATLAS experiment makes no clear attempt to resolve detector signals assigned to multiple types of objects, leading to double counting. Examples of such situations are muons radiated from jets or jets also reconstructed as electrons. In the $WH \to \ell\nu b\bar{b}$ analysis, the object overlap removal is based on a set of ordered rules establishing the object hierarchy. These are applied to loose jets, electrons and muons as follows:

Jet-e Electrons can be misidentified as jets since the jet reconstruction and selection chain does not attempt to distinguish jets of quark or gluon hadronisation from electromagnetic shower developments caused by high energy electrons. On the contrary, the track isolation criteria imposed to loose electrons aims to reject jets mis-identified as electrons. To reject fake jets simultaneously reconstructed as electrons, jets within $\Delta R < 0.4$ to a loose electron are

Overlap Removal Rule	Number of selected objects
Loose jets after jet-e removal	735501
Loose jets after jet-μ removal	730537
Loose muons after μ-jet removal	98578
Loose electrons after e-jet removal	100885
Loose electrons after e-μ removal	100572
Calorimeter muons after e-μ removal	1506

Table 6.10: Overlap removal conditions and number of selected objects after applying each condition obtained with by the LIP analysis code. A signal sample with 300000 events was used as reference. The numbers were compared to the outcome of other groups codes and an absolute agreement was found for each condition.

removed.

Jet-μ Muons can also be misidentified as jets. Muons that radiate photons can give rise to calorimeter energy clusters and end up reconstructed as jets. Therefore, jets with at most 3 associated tracks of p_{T} >500 MeV, within a distance ΔR <0.4 to muons are removed. The 3 track threshold is used because if only few tracks are associated with the calorimeter energy deposit, the reconstructed object corresponds more likely to a muon. On the contrary, gluon or quark initiated jets tend to have larger track multiplicities.

μ-jet In the case of larger track multiplicity the jet prevails over the muon, as the object is more probably a real jet. So, muons within a distance ΔR <0.4 to jets with more than 3 matching tracks are removed.

e-μ Muons that produce delta rays in the calorimeter or that radiate photons subsequently undergoing pair production can end up reconstructed as electrons. If the MS information is available, i.e. the muon is not a calorimeter muon, electrons within ΔR <0.2 to muons are removed.

Calo μ-e Remaining calorimeter muons within ΔR <0.2 to electrons are removed since they are considered electrons.

Table 6.10 shows the different overlap removal rules and the number of objects after applying each rule as obtained by the LIP analysis code. The implementation of this procedure in the analysis code was validated amongst the different groups through output comparisons and an absolute agreement was found.

6.3.7 Reconstruction of the Higgs candidate

In the search for the $H \to b\bar{b}$ decay, the four-momentum of the Higgs candidate p^{H} corresponds to the resultant four-momentum of the $j_1 + j_2$ system: $p^{\mathrm{H}} = p^{j_1} + p^{j_2}$. The invariant

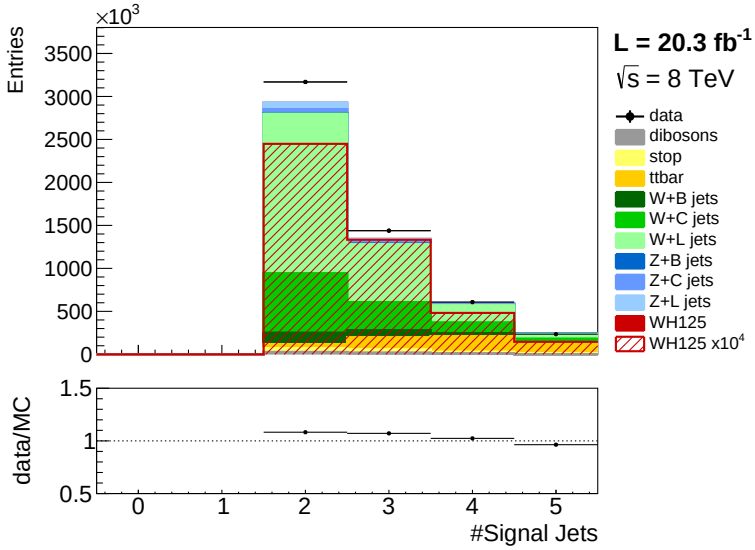


Figure 6.15: Distribution of the signal jet multiplicity for data and MC prediction after the forward jet veto. The multijet background expectation is not included.

mass of the Higgs candidate is given by:

$$m_H = M^{j_1 j_2} = \sqrt{(E^{j_1} + E^{j_2})^2 - ||\mathbf{p}^{j_1} + \mathbf{p}^{j_2}||^2} \tag{6.2}$$

For a 125 GeV Higgs mass, the $M^{j_1 j_2}$ spectrum for signal events is characterised by a peak indicative of the presence of the 125 GeV boson resonance, while for background events a wide continuously falling distribution is expected. For this reason, this observable is highly significant to the analysis. Since the amount of expected signal events associated with this search is very small with respect to the number of background events, the best mass resolution is desired.

Final state radiation can give rise to signal events with jet multiplicity greater than two. This corresponds mostly to gluon emission from the final $b-$quarks from the Higgs decay. Therefore, a third jet might appear in signal events. As shown in Figure 6.15, the effect is not negligible. To improve signal detection efficiency, a three-jet event category was created, where these type of signal events can be recovered and analysed. A non $b-$tagged jet is required since the third jet is light flavoured. The Higgs candidate, in this case, is correctly described by the three jet system: $p^{\mathrm{H}} = p^{j_1} + p^{j_2} + p^{j_3}$.

Corrections to the di-jet Invariant Mass

The jet energy resolution is the driving factor contributing to the mass resolution, but other effects contribute to its degradation. Initial and final state radiation increase the upper and lower tails of the $m_{b\bar{b}}$ spectrum. The initial state radiation is associated to quark or gluon emission from partons before the collision. If initial state jets fall within a reconstructed $b-$jet from the Higgs decay, the reconstructed energy of that signal jet will be larger than it truly

is and this type of events will populate the upper tail of the $m_{b\bar{b}}$ spectrum. On the contrary, not recovered final state radiation will lead to jet energy loss and smaller $m_{b\bar{b}}$. The hadronic cascade associated with the $b-$quark fragmentation includes muons with a 20% probability. So, muons escaping the reconstructed jet result also in energy loss and lower $m_{b\bar{b}}$ values. The same happens for neutrino emissions from hadronic cascades. To account for some of these effects, two corrections are applied to the two leading signal jets. Both are meant to improve the scale and resolution of the invariant mass of the $H \to bb$ system.

Muon-in-jet Signal jets that have a muon within $\Delta R < 0.4$ are corrected for the energy loss due to the escaping muon. The corrected jet 4-momentum is given by:

$$p^{jet}_{\text{corr}} = p^{jet} + p^{\mu} - p^{\mu}_{\text{calo}} \tag{6.3}$$

where the p^{μ}_{calo} term is associated with the muon energy loss in the calorimeter and already accounted for in the jet momentum measurement before the correction. The muon must be a combined muon and have a p_{T} larger than 4 GeV. If more than one muon is found in these conditions, the closest to the jet is taken to compute the correction.

Jet Reconstructed p_{T} After jet energy calibration, there are still residual biases in the jet energy measurement by the calorimeter, related specially with the jet energy resolution, affected, for instance, by neutrino emissions. In order to correct for its effects, a p_{T}-dependent correction is applied to signal $b-$jets leading to an improvement on the $m_{b\bar{b}}$ peak resolution for signal events. The correction was determined using the jet p_{T} spectrum for a sample of simulated signal events in the form $C(p^{\text{reco}}_{\text{T}}) = p^{\text{truth}}_{\text{T}}/p^{\text{reco}}_{\text{T}}$, where $p^{\text{reco}}_{\text{T}}$ and $p^{\text{truth}}_{\text{T}}$ are the measured and truth p_{T} of the jet respectively. Here, the truth p_{T} accounts for muons and neutrinos inside jets.

The muon-in-jet and reconstructed p_{T} corrections contribute to a 14% improvement of the $m_{b\bar{b}}$ resolution for signal events [50].

6.3.8 Reconstruction of the W candidate

In the case of the $W \to \ell\nu$ decay channel at the LHC, the complete description of the W can not be achieved since the neutrino longitudinal momentum is not directly measured. Therefore, only the transverse components of the W boson momentum are determined, and for that reason, its invariant mass is not measured experimentally. The transverse mass is used instead, as defined in Eq. 6.4, and for the case of the WH analysis, where only the electronic and muonic decay modes are considered, the lepton masses are neglected.

$$m^{W}_{\text{T}} = \sqrt{2E^{\text{miss}}_{\text{T}} E^{\ell}_{\text{T}} (1 - \cos\Delta\phi(\ell, E^{\text{miss}}_{\text{T}}))} \tag{6.4}$$

$\Delta\phi(\ell, E_{\mathrm{T}}^{\mathrm{miss}})$ is the azimuthal angle between the lepton and $E_{\mathrm{T}}^{\mathrm{miss}}$. m_{T}^{W} can range from 0 to m^{W} for real W events but can have random values for other processes that have the same experimental signature in the detector as the W leptonic decay. In this way, this observable can be used to select signal-like events and reject background containing fake Ws. As shall be seen later, it is one of the discriminant input variables to the $WH \to \ell\nu b\bar{b}$ multivariate analysis.

Along with m_{T}^{W}, the W transverse momentum, p_{T}^{W}, is a special observable to this analysis. In the plane transverse to the beamline, the Higgs boson momentum projection cancels the W p_{T} and consequently the latter is proportional to the Higgs boost. For this reason, many of the cuts in the analysis event selection are p_{T}^{W}-dependent, and the variable is used as well to separate the selected events into two categories corresponding to p_{T}^{W} below and above 120 GeV, as the event topology is different for the two regimes.

Neutrino Longitudinal Momentum

A full reconstruction of the leptonically decaying W depends uniquely on the neutrino longitudinal momentum, p_{z}^{ν}, that can not be measured. But assuming that the W is on-shell and neglecting its width, a simple kinematic condition based on the W mass can be used to estimate p_{z}^{ν} as given by Equation 6.5.

$$\begin{aligned} p_{\mathrm{z}}^{\nu} &= \frac{1}{E_{\ell} - p_{\mathrm{z}}^{\ell}}\left(a p_{\mathrm{z}}^{\ell} \pm E_{\ell}\sqrt{a - E_{\ell} p_{\mathrm{T}}^{\nu} + p_{\mathrm{z}}^{\ell} p_{\mathrm{T}}^{\nu}}\right) \\ a &= \frac{M_{W}^{2}}{2} + p_{\mathrm{T}}^{\nu} p_{\mathrm{T}}^{\ell} cos(\phi_{\ell} - \phi_{\nu}) \end{aligned} \tag{6.5}$$

where the ℓ and ν superscripts represent the charged lepton and the neutrino respectively. With this procedure, the W mass measurement is traded for its longitudinal information. However, this can open an new insight on the events that can serve the analysis.

The equation shows that one has to deal with possibly arising imaginary solutions, and ultimately choose between the $\pm$ sign. Therefore, the correct solution cannot be unambiguously determined. Imaginary solutions are a consequence of the finite resolution of the detector for real W events, and in the case of background, can additionally be due to an inexistent neutrino. So, the imaginary component of the solution is neglected for not having a physical meaning.

To choose between the $\pm$ solutions, four options were tested based on a generator-level study using simulated $WH \to \ell\nu b\bar{b}$ events and accessing the MC truth information of the W decay. For a 300000 event sample, using the plus or minus solution yields the correct solution 52% or 48% of the times, respectively, as summarised in Table 6.11. However, using the solution resulting in the minimal absolute difference between the W and Higgs bosons longitudinal boost, the correct result is reached about 68% of the times. Choosing the solution that minimises the $p_{\nu,z}$ absolute value yields approximately 57%.

The p_{z}^{ν} determination will be relevant for the study presented in Section 6.5.3 and so this subject is concluded there.

Solution	Correct
Plus sign	52.1 %
Minus sign	47.9 %
$p_{\nu,z}$: min $\lvert\beta_z^H - \beta_z^W\rvert$	67.7 %
$p_{\nu,z}$: min $\lvert p_z^\nu\rvert$	56.8 %

Table 6.11: Correct result rate using several $p_{\nu,z}$ solutions.

6.4 Event Selection

The object selection described so far only deals with the final state objects of the hadron collisions. It is needed to identify the analysis signal jets, electrons and muons, and to resolve the arising ambiguities between these objects on an event-by-event basis, without excluding any event from analysis. Then, the event is evaluated based on the multiplicity of these objects and on kinematics. If it has a topology compatible with a $WH \to \ell\nu b\bar{b}$ event, namely one isolated electron or muon, substantial $E_{\mathrm{T}}^{\mathrm{miss}}$ associated with the final state neutrino and at least two $b-$jets, the event is accepted as a signal event candidate. Furthermore, the event has to satisfy more general requirements that are analysis-independent, related with data or MC quality. This general selection and the selection criteria used to identify signal event candidates are described in this section.

6.4.1 General selection

The quality criteria summarised in Table 6.12 and described in the following, are applied to reject events with detector errors or that are badly reconstructed.

Good Runs List (GRL) During data taking, data quality is continuously monitored by a combination of automated software and analysis by shifters, checking detector functionality and performance. A GRL, listing the runs for which all the essential elements of the ATLAS detector were fully operational, is composed based on these checks. Data events not included in the GRL are excluded from physics analyses.

Vertex Rejects data or simulated events without a hard scatter vertex candidate, by imposing that the main primary vertex has at least three associated ID tracks. The main primary vertex is the one with the largest p_{T} quadratic sum of all the matching tracks.

$E_{\mathrm{T}}^{\mathrm{miss}}$ **cleaning** Identifies background jets resulting from cosmic rays interactions or detector effects in data or simulated events; the event is removed if it has at least one of these bad jets with calibrated $p_{\mathrm{T}} \geq 20$ GeV;

LAr error veto Remove data events with LAr error flag related with noise burst and data corruption;

TileCal error veto Remove data events with TileCal error flag related with noise burst;

Incomplete event veto Remove data events with incomplete events;

Selection	MC events	Data events
Initial	289999	157109
Good Runs List	289999	148520
Vertex	298831	147777
$E_{\mathrm{T}}^{\mathrm{miss}}$ cleaning	297981	147484
LAr error veto	297981	147283
TileCal error veto	297981	147283
Incomplete event veto	297981	147283
TileCal corrupted veto	297981	147283
Jet Cleaning	297981	147283

Table 6.12: Quality selection criteria applied to reference samples of data and simulated events and number of selected events after applying each selection condition obtained with the LIP analysis code. A signal and a data sample of the Muon stream were used as references. A signal sample with 300000 events was used as reference. The numbers were compared to the outcome of other groups codes and an absolute agreement was found for each condition.

TileCal corrupted veto Remove data events with TileCal corrupted information;

Jet Cleaning Remove data events with at least one jet reconstructed in known problematic calorimeter regions.

The implementation of this selection was validated within the different groups contributing to the analysis, and a perfect agreement was found for both simulated and collision data events.

6.4.2 Selection of $WH \to \ell\nu b\bar{b}$ events

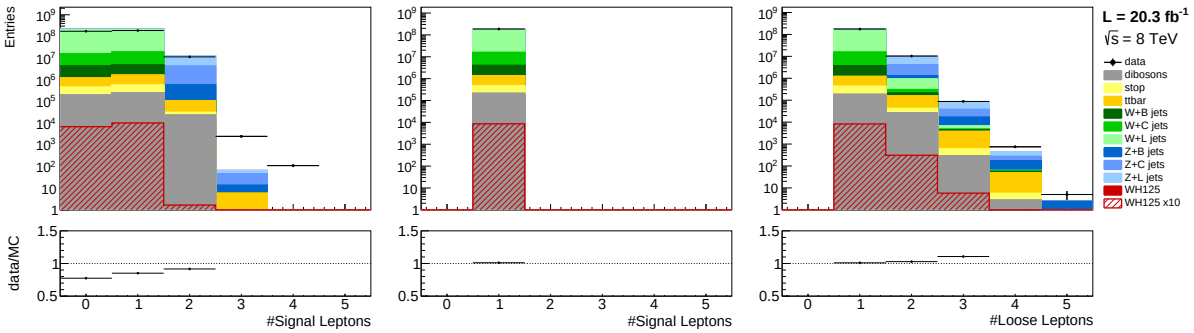
The event selection used in the search for the $WH \to \ell\nu b\bar{b}$ process and the efficiency of its criteria are summarised in Table 6.13. Some of the selection criteria are loosened with respect to the traditional cut-based analysis [50] and do not target the best signal significance but rather a good statistical description of all processes relevant to this search, either background or signal. Afterwards, it is the BDT method that explores each event topology to better separate signal from background events and refine the signal-to-background ratio.

The event must have exactly one signal lepton, electron or muon, matching the trigger lepton, i.e. reconstructed within a maximum distance of 0.15 (0.1) from the trigger electron (muon). Requiring 1 signal lepton has an approximate efficiency of 60% for signal events, as Figure 6.16(a) illustrates. The 40% loss reflects the inefficiencies in the electron and muon reconstruction algorithms, and losses due to the chosen fiducial region of $|\eta| < 2.5$. The trigger matching reduces the background and the signal by approximately 10% in average, as Table 6.13 shows.

Events with additional loose electrons or muons are vetoed, reducing about 5% of the total background and Z+jets to a half. The distributions of the signal and loose lepton multiplicities are shown in Figure 6.16 for pp collision data correspondent to the complete 8 TeV Egamma and Muon streams. The MC is normalised according to the predicted cross-section, as will be detailed at the end of this Section.

Selection	WH	$t\bar{t}$	top	W+jets	Z+jets	dibosons
>0 Loose Lepton	100 (1705)	100 (1.9×10^6)	100 (5.7×10^5)	100 (3.9×10^8)	100 (5.5×10^7)	100 (4.7×10^5)
1 signal lepton	59.6	56.2	54.1	49.8	36.6	51.7
Trigger matching	50.3 (858)	47.2 (8.9×10^5)	45.2 (2.5×10^5)	41.3 (1.6×10^8)	31.0 (1.7×10^7)	44.6 (2.1×10^5)
Loose lepton veto	48.1	40.8	42.2	40.6	14.3	38.7
>1 signal jets	36.4	39.8	31.9	2.6	1.5	22.3
$E_{\mathrm{T}}^{\mathrm{miss}}$ >20 GeV	35.9	39.4	31.6	2.6	1.4	22.1
M_{eff} >180 GeV	28.3	37.3	25.8	1.2	0.51	13.3
Forward jet veto	26.1	29.8	19.9	1.1	0.41	11.8
2 signal jets	14.3	2.6	7.2	0.67	0.23	5.7
2 b-tagged jets	9.4 (160)	2.0 (39649)	1.6 (8825)	0.0083 (33103)	0.0034 (1879)	0.16 (778)
$\Delta R(b_1,b_2)$ >0.7	9.3	2.0	1.6	0.008	0.0032	0.16
$p_{\mathrm{T}}^{b_1}$ >45 GeV	9.2 (157)	2.0 (38324)	1.5 (8742)	0.0071 (28442)	0.0031 (1706)	0.15 (724)
3 signal jets	7.8	7.4	6.7	0.27	0.11	3.8
2 b-tagged jets	2.3 (39)	1.4 (25980)	0.68 (3854)	0.0021 (8383)	0.0011 (599)	0.050 (233)
$\Delta R(b_1,b_2)$ >0.7	2.3	1.4	0.67	0.0020	0.0010	0.048
$p_{\mathrm{T}}^{b_1}$ >45 GeV	2.2 (38)	1.3 (25381)	0.65 (3730)	0.0019 (7501)	0.0010 (554)	0.047 (220)

Table 6.13: Cumulative efficiency (in percentage) of each event selection criterion for the various simulated processes under analysis. The absolute value of expected events at the beggining and at the end of the selection is shown within curved brackets for a few key selection criteria. The number of expected events is normalized proportionaly to each process cross-section to the integrated luminosity of 20.3fb^{-1}of the $\sqrt{s} = 8$ TeV data set.



(a) (b) (c)

Figure 6.16: Distribution of the (a) number of signal leptons before the 1 signal lepton cut and (b) after trigger matching. (c) Distribution of the number of loose leptons before the loose lepton veto. The events have at least one loose lepton. Data and prediction are shown (the multijet background expectation is not included).

For the plot in Figure 6.16(a), the MC does not have any trigger applied leading to an overestimate of data by MC. After performing the trigger matching, the feature almost vanishes as expected. The multijet background is not included in the expectation side for any distribution. This explains the residual disagreement between the yields of prediction and observation.

The triggers used to select the events are the lower threshold un-prescaled single-lepton triggers available for the 8 TeV run and are presented in Table 6.14. The electron trigger has E_{T} thresholds of 24 GeV and 60 GeV and an efficiency above 90% for electrons with $E_{\mathrm{T}} > 30$ GeV, as shown in Figure 4.5 of Section 4.2. The muon trigger, with p_{T} thresholds of 24 GeV and 36 GeV, reaches a stable efficiency of 85%(70%) above muon p_{T} =24 GeV for $|\eta| > 1.05(<1.05)$, see Figure 4.11. These are single lepton triggers, meaning that at least one lepton in the event gathers the necessary conditions to fire the corresponding trigger item. For both electrons and muons, the lowest energy threshold trigger requires track isolation to fight the pile-up conditions of the LHC:

Electron track isolation: the sum of the p_{T} of all tracks inside a cone of radius 0.2 around the electron track has to be smaller than 10% of the electron p_{T}.

Muon track isolation: the p_{T} sum of all tracks inside a cone of radius 0.2 centred at the muon track must be smaller than 12% of the muon p_{T}.

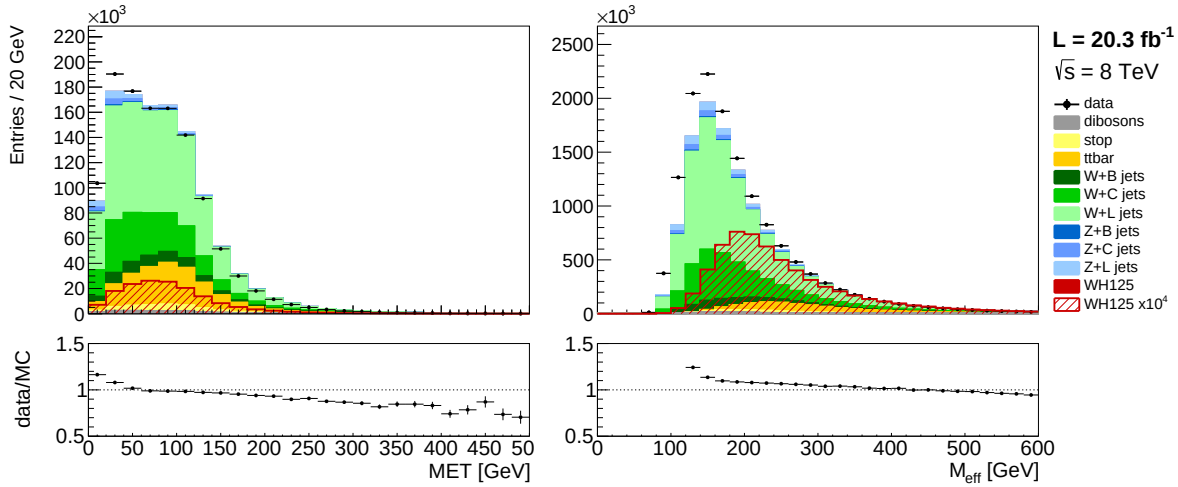
The `e60_medium1` and `mu36_tight` are considered to compensate the efficiency losses caused by the isolation requirement.

MC events are corrected for differences in trigger efficiency between data and simulation, according to the scaling factors presented in Figures 4.5 and 4.11. Matching the signal lepton with the triggering object ensures that this correction is well defined.

The presence of a final state neutrino motivates a lower limit on $E_{\mathrm{T}}^{\mathrm{miss}}$. This requirement is tight in the case of the cut-based version of the analysis to reject backgrounds with fake $E_{\mathrm{T}}^{\mathrm{miss}}$,

Trigger item	Threshold	Isolation	Identification
`e24vhi_medium1`	$E_{\mathrm{T}} > 24$ GeV	track isolation	medium e
`e60_medium1`	$E_{\mathrm{T}} > 60$ GeV	none	medium e
`mu24i_tight`	$p_{\mathrm{T}} > 24$ GeV	track isolation	tight μ
`mu36_tight`	$p_{\mathrm{T}} > 36$ GeV	none	tight μ

Table 6.14: Electron and muon trigger items used in the $WH \to \ell\nu b\bar{b}$ event selection.



(a) (b)

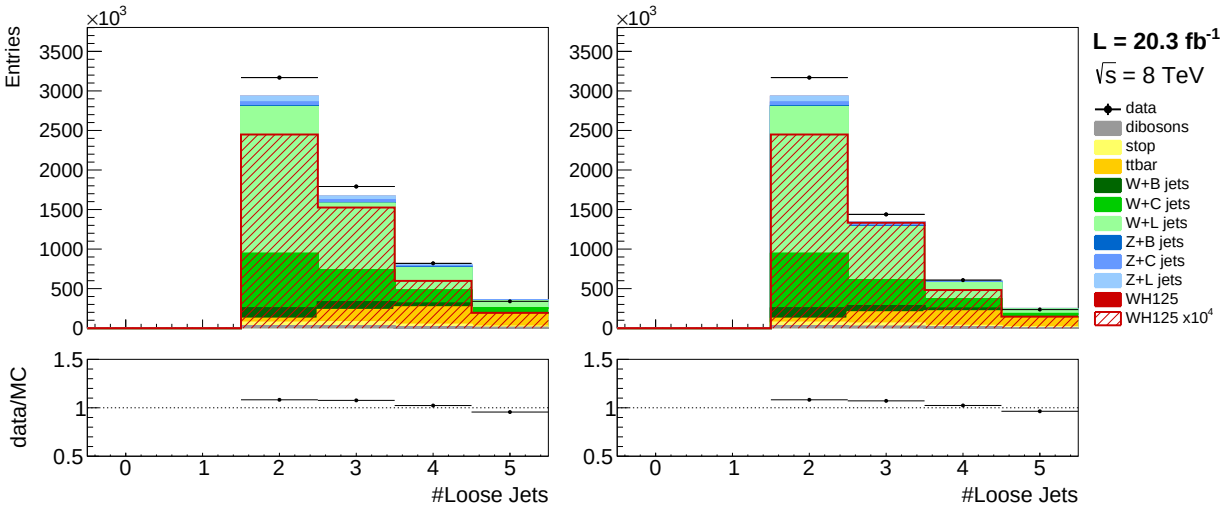
Figure 6.17: (a) $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution for events with $p_{\mathrm{T}}^{W} > 120$ GeV before the $E_{\mathrm{T}}^{\mathrm{miss}}$ cut and (b) M_{eff} distribution before the M_{eff} cut for data and prediction, except for the multijet background.

but since this variable is used as input in the WH MVA, the cut was loosened to only 20 GeV for $p_{\mathrm{T}}^{W} > 120$ GeV. In this way, the cut efficiency is nearly 100% for signal events, as Figure 6.17(a) shows, and the BDT is left with the task to better explore and judge the event. Z+jets are not much suppressed by this cut because the lepton veto already discarded most of these events. The cut is however specially relevant to fight the multijet background that has fake $E_{\mathrm{T}}^{\mathrm{miss}}$, as evidences the underestimated prediction at the low $E_{\mathrm{T}}^{\mathrm{miss}}$ region. The plots reveal a clear mis-modelling in the tail above 150 GeV, mostly composed of the W+jets background. This feature disappears after the full event selection.

The event is required to have an effective mass above 180 GeV for p_{T}^{W} <120 GeV. The effective mass M_{eff} is defined as the transverse momentum scalar sum of the four final state objects:

$$M_{\mathrm{eff}} = p_{\mathrm{T}}^{j_1} + p_{\mathrm{T}}^{j_2} + p_{\mathrm{T}}^{\ell} + E_{\mathrm{T}}^{\mathrm{miss}} \tag{6.6}$$

where j_1, j_2 and ℓ represent respectively the leading and sub-leading jets and the electron or muon in the event. Figure 6.17(b) shows the impact of the M_{eff} requirement. Setting $M_{\mathrm{eff}} >$ 180 GeV eliminates about 60% of the remaining W+jets background. The remaining cuts of the event selection fix the mis-modelling observed for M_{eff} >500 GeV.



(a) (b)

Figure 6.18: Loose jet multiplicity (a) before and (b) after the forward jet veto for data and MC prediction. The multijet background expectation is not included.

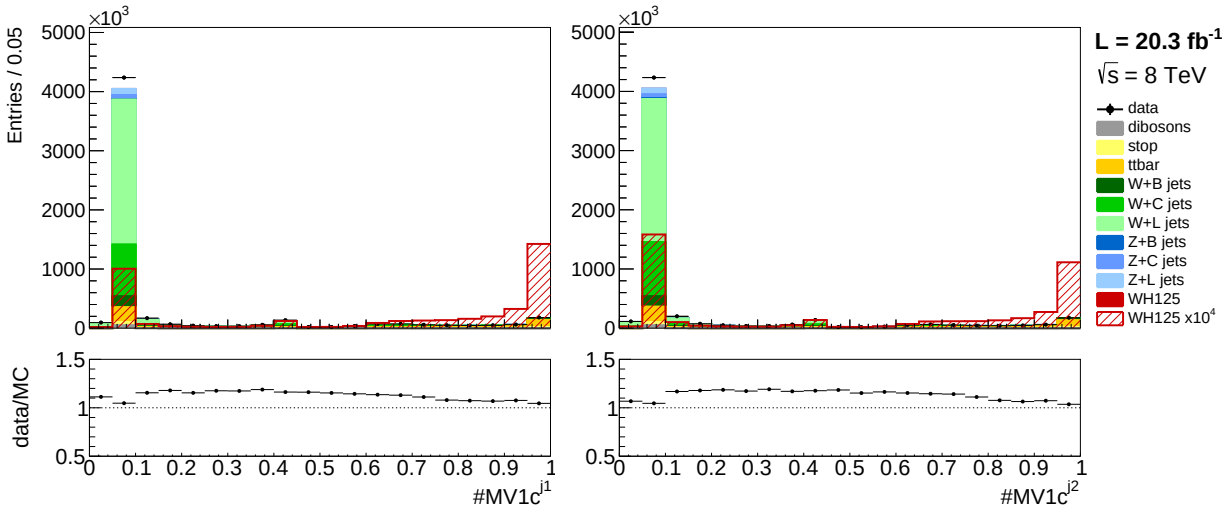
The event must have at least two signal jets and no forward jets. This last cut was designed to suppress top background events, reducing approximately 8 and 6% the $t\bar{t}$ and single top samples, respectively, as Figure 6.18 and Table 6.13 show.

The two p_{T} leading jets must be identified as $b-$jets by the MV1c $b-$tagging algorithm at the efficiency working point corresponding at least to the loose $b-$tagging of MV1c>0.4050. The $b-$tagging procedure rejects most of the $W+c-$ and light-jets as shown in Figure 6.19. The data and MC disagreement that appear on the plots is reduced after full selection. But in anyway, it is not problematic given the fact that the normalisation of the main backgrounds will be determined from data during the WH statistical analysis.

The Higgs boson candidate is formed by the two signal $b-$jets. The p_{T} of the leading jet must be larger than 45 GeV to reject mostly $t\bar{t}$ and $W+$jets events, as seen in Figure 6.20(b). Finally, the radial distance between the two $b-$jets must be greater than 0.7 when p_{T}^{W} <200 GeV. For this p_{T}^{W} range, the Higgs boson is not very boosted, and therefore the jets produced by its decay are typically widely separated. It can be seen from Figure 6.20(a) how the $\Delta R(b_1,b_2)$ distribution for WH is practically not affected by this requirement while a fair amount of $W/Z+$jets events is eliminated.

Validation of the Event Selection tools

The event selection code was validated by comparing the selection outcomes of all groups participating in the WH search. The process was iterative and allowed to find and solve existing problems, until a generally good agreement was established. Since the MVA and cut-based selections do not differ much, only the cut-based selection was cross-checked through a cut flow comparison. Three event samples were used as references: WH simulation, Egamma and



(a) (b)

Figure 6.19: Distribution of the MV1c $b-$tagging weight of the (left) $p_{\mathrm{T}}-$leading jet and (right) $p_{\mathrm{T}}-$subleading jet for events with two signal jets before $b-$tagging for data and MC prediction, except for the multijet background.

Muon streams data events. Although the MVA selection was not directly compared at a cut flow basis, the yields of the different samples were cross-checked for the different analysis categories. Table 6.15 shows the cut-based analysis criteria used in the $WH \to \ell\nu b\bar{b}$ search and the number of events passing each criterion as obtained by the LIP analysis code. The relative deviations, given in %, of the other group codes output with respect to the numbers obtained by the LIP code are also shown. As can be seen, a maximum relative deviation of 0.08% was found.

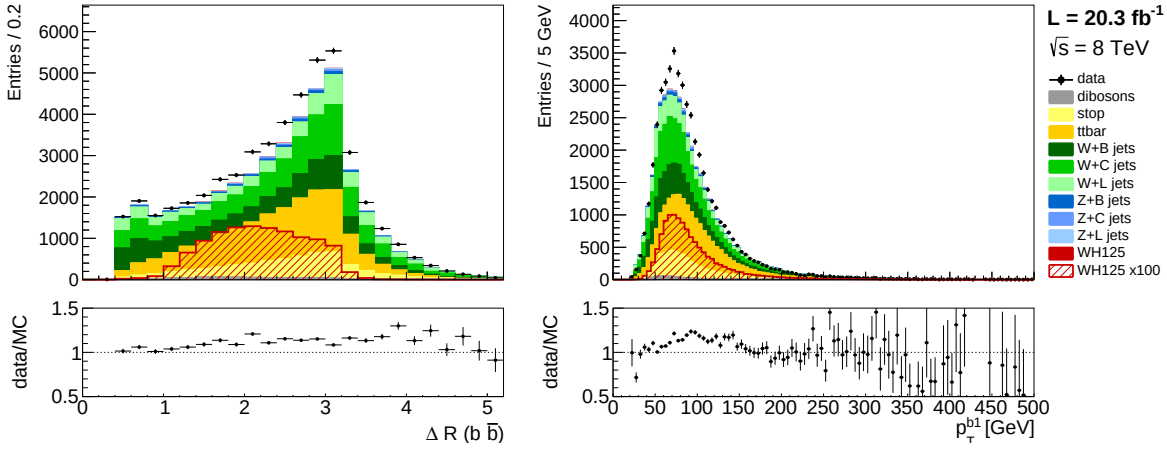
Normalisation of the Simulated Samples

The expected number of events must be determined for simulation, and the final samples normalised accordingly. For a given process of cross-section $\sigma(s)$, the expected number of events N, also referred to as yield, for an integrated luminosity L is simply given by $N = \sigma(s) \times L$. Thus, when dealing with an event selection of efficiency ε, N is determined by the following expression:

$$N = \sigma(s) \times L \times \varepsilon \tag{6.7}$$

where ε greatly depends on the topology of the process under consideration.

In the $WH \to \ell\nu b\bar{b}$ analysis, all simulated samples are normalised in this manner. The cross-sections of the processes assume the Standard Model hypothesis for both signal and backgrounds and were summarised in Tables 6.1 and 6.2, respectively. However, since the uncertainty on σ is not negligible for the majority of the processes, the normalisation of the main backgrounds will still be fitted to the data observation during the statistical analysis. Thus, for



(a) (b)

Figure 6.20: (a) $\Delta R(b_1, b_2)$ distribution before the $\Delta R(b_1, b_2)$ cut for events with p_{T}^{W} <200 GeV. (b) p_{T} distribution of the $p_{\mathrm{T}}-$leading signal jet before the leading jet p_{T} cut. Data and MC prediction, except for the multijet background, are shown.

these samples, the normalisation procedure described above sets the starting basis of a process where the observation is used to get the best out of the simulation.

While this procedure holds for MC, the multijet background normalisation obeys a different method, addressed in Section 6.4.3.

Event Categories

The events selected by the analysis are separated into multiple categories to enhance the signal significance and to establish background-enriched regions. The latter are denominated control regions and are very useful to constrain dominant backgrounds normalisation from data during the final fit of the analysis, as mentioned before. A total number of 16 categories are defined in the WH search. These are based on:

- jet multiplicity: 2 or 3 jets;
- two p_{T}^{W} intervals: below or above 120 GeV;
- three $b-$tagging categories: LL, MM or TT as defined previously;
- a control region is obtained by requiring only one $b-$tagged jet.

Figure 6.21 shows the signal and backgrounds yields, respectively S and B, and the total number of events expected for each category. The MC samples of W or Z plus jets are split according to the flavour of the jets into bb, bc, bl, cc, cl and light. The jet is assigned the flavour of the closest hadron using truth MC information. As expected, W plus light and $c-$jets, without true $b-$jets, populate more the 1 tag region. The $t\bar{t}$ background dominates the 3 jet category due to the large jet multiplicity that characterises this process. The 1 tag and 3 jet regions can be used to extract the normalisation of these backgrounds from real pp collisions

Selection	LIP	CERN	CPPM	Tsukuba
1 Signal lepton	112816	0	0	0
Trigger matching	94590	0.03	0	0.03
Loose lepton veto	91204	0	0	0
≥ 2 Signal jets	68695	0	0	0.03
$E_{\mathrm{T}}^{\mathrm{miss}}$ cut	67412	0	0	0.03
H_T >180 GeV	53768	0	0	0.04
m_{T}^W <120 GeV	49554	0	0	0.04
Forward jet veto	45517	0	0	0.04
Exactly 2 Signal jets	25004	0	0	0.02
Exactly 2 b-tagged jets	12231	0	0	0.04
$\Delta R(b_1, b_2)$ lower limit	12191	0	0	0.04
$p_{\mathrm{T}}^{b_1}$ >45 GeV	12021	0	0	0.04
$\Delta R(b_1, b_2)$ upper limit	11589	0	0	0.04
Exactly 3 Signal jets	13601	0.01	0.01	0.05
Exactly 2 b-tagged jets	3870	0.03	0.03	0.08
$\Delta R(b_1, b_2)$ lower limit	3842	0.03	0.03	0.08
$p_{\mathrm{T}}^{b_1}$ >45 GeV	3782	0.03	0.03	0.08
$\Delta R(b_1, b_2)$ upper limit	3682	0.03	0.03	0.08

Table 6.15: Event selection criteria used to select $WH \to \ell\nu b\bar{b}$ events. The number of events passing each cut obtained with the LIP analysis code implementation is shown for a reference samples of simulated WH signal events. The percentage deviation of the number of selected events as obtained by the other groups codes with respect to LIP, defined as $\Delta = (N_{group} - N_{LIP})/N_{LIP} \times 100\%$, is presented in the last three columns. A signal sample with 300000 events was used as reference.

data. Typically, multiple jets events have fake $E_{\mathrm{T}}^{\mathrm{miss}}$ and therefore their yield and proportion relative to the remaining backgrounds decrease with p_{T}^W.

The signal cross-section is very small compared to the background processes resulting in a low expectation of the signal amount. In Figure 6.21, the signal VH label corresponds both to the WH and ZH production, although the ZH contribution to the final sample of events fulfilling the 1-lepton selection can be considered negligible. The relative proportion of signal increases with $b-$tagging purity and p_{T}^W and so, by having the events separated in this manner the signal significance is enhanced. Figure 6.22 shows it precisely. Here, the signal significance $S/\sqrt{B}$ is plot as a function of the analysis categories. One can clearly see its dependence on $b-$tagging and p_{T}^W. The most sensitive region corresponds to the 2 jets, $b-$tagged as TT, and $p_{\mathrm{T}}^W > 120$ GeV but even in this case, $S/\sqrt{B}$ is only 0.4 roughly.

6.4.3 Multijet Background estimate

The multijet background arises from jets misidentified as leptons, and is the only background whose prediction is not obtained through simulation. An example of a diagram contributing to multijet production with pp collisions is depicted in Figure 6.7. It exclusively involves the strong interaction and hence is often denominated QCD background. Despite the

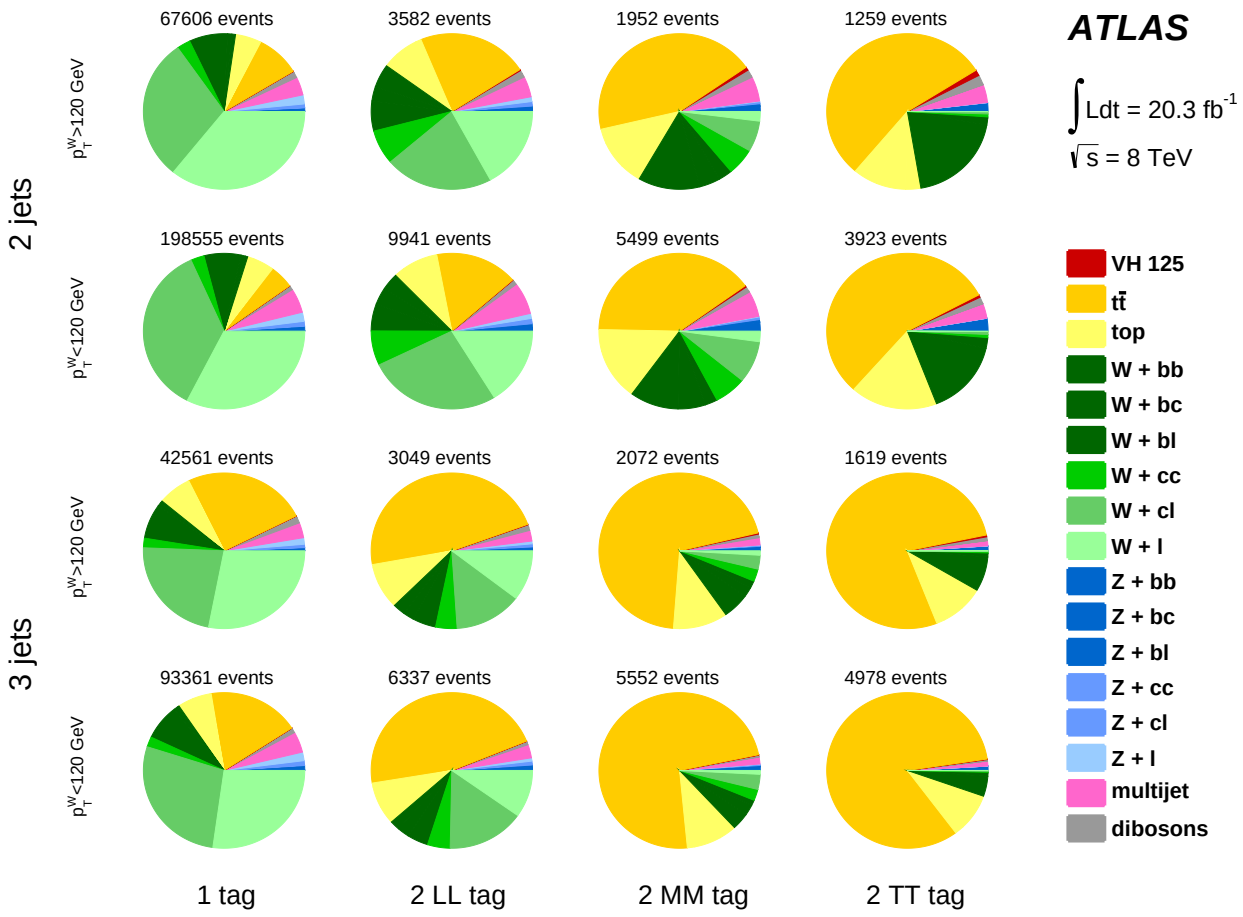


Figure 6.21: Relative proportions of signal and background events for the different analysis categories.

probability for these fake leptons to arise is very small, the QCD cross-section is very large and multijets result on an important contribution to the $WH \to \ell\nu b\bar{b}$ background, as already seen in Figure 6.21.

MC simulation is not accurate enough to describe fairly all the aspects of the QCD background. Besides, a statistically significant sample of these events is very hard to obtain from simulation due to the high rejection imposed by the analysis selection design. So, this background is obtained in a more reliable manner from experimental data.

Selection of Multijet-like events

Heavy flavour hadrons decaying semi-leptonically and jets misidentified as leptons result in electrons or muons in the detector signature of the multijet process. These fake leptons are typically not isolated, and this feature is used to obtain a multijet-enriched sample from real data events. The multijet sample is selected from data by using medium identified leptons and all the remaining event selection criteria, except from the isolation conditions used to define signal leptons. The multijet-like electrons or muons are required to be non-isolated in the tracker and loosely isolated in the calorimeter:

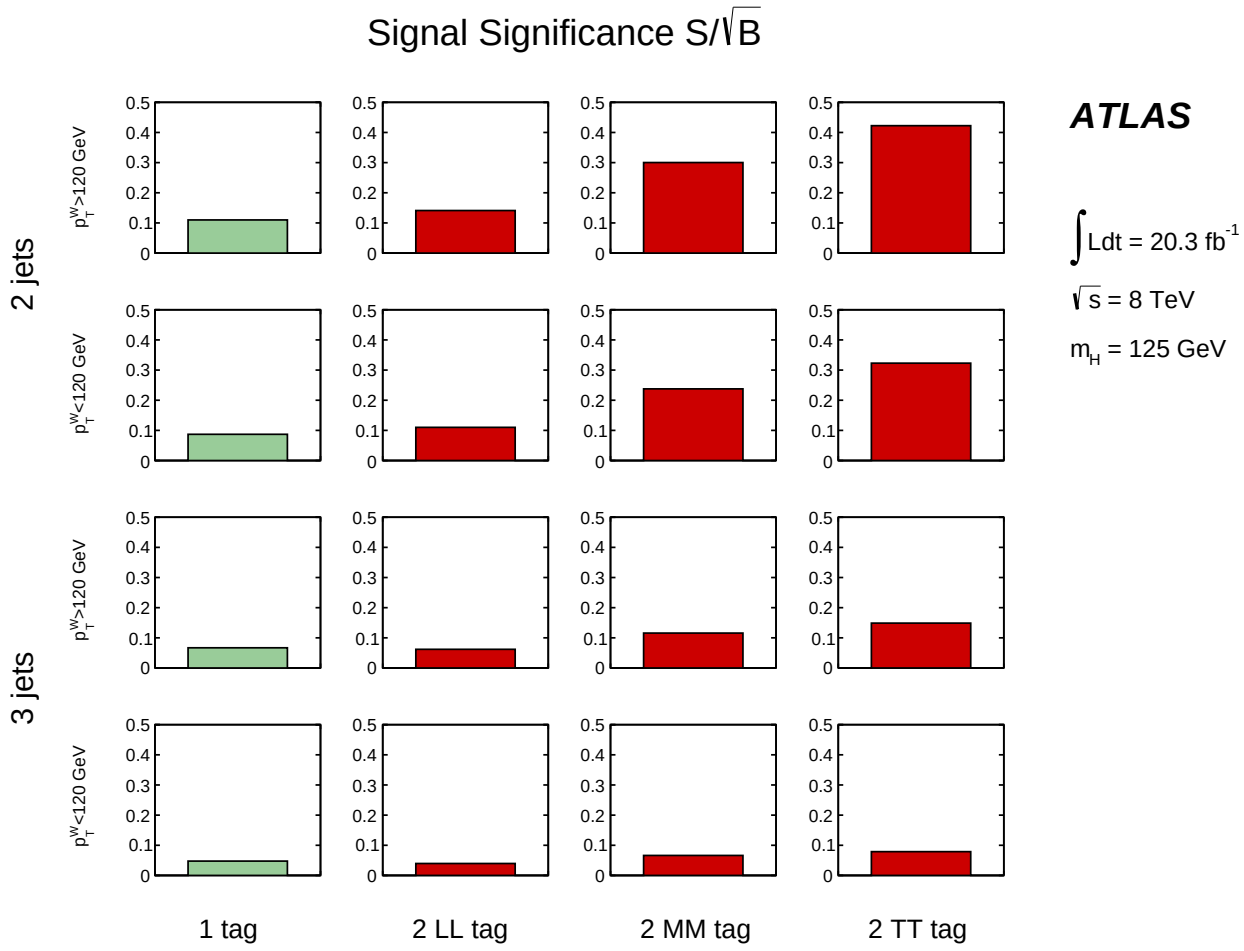


Figure 6.22: Signal significance $S/\sqrt{B}$ for the control regions (green) and signal regions (red).

- the sum of the p_{T} of all tracks inside a cone of radius 0.2 around the lepton track relative to the lepton p_{T} should be within the interval $[5\%, 12\%]$ for electrons and $[7\%, 50\%]$ for muons;
- the E_{T} deposited inside a cone of radius 0.3 around the lepton should be smaller than 7% of the lepton E_{T}.

The data sample selected provides a template for the multijet background shape for each distribution. The selection was designed to guarantee a very pure sample of multijet events, but other processes can still contaminate it. For this reason, events from all the other simulated backgrounds that pass the multijet selection conditions are subtracted from the multijet sample obtained from data.

Normalisation of the Multijet Background

The resulting sample is then normalised by extracting a scaling factor from a maximum likelihood fit using the $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution. In this fit, the multijet template plus all simulated processes passing the signal selection are adjusted to data in the signal region, by floating their normalisations. The resulting scaling factors are here denoted by α_{QCD} and α_{MC}, respectively,

	2 jets				3 jets			
	1 $b-$tag		2 $b-$tags		1 $b-$tag		2 $b-$tags	
	e	μ	e	μ	e	μ	e	μ
α_{QCD}	0.942	1.42	1.17	1.88	0.983	1.34	1.05	0.981
α_{MC}	1.04	1.02	1.09	1.09	0.972	0.955	1.05	1.03

Table 6.16: Multijet and Monte Carlo normalization scale factors, α_{QCD} and α_{MC}, obtained from the maximum likelihood fit to data.

where the α_{MC} scaling parameter is common to all the MC processes.

The usage of the $E_{\text{T}}^{\text{miss}}$ distribution was motivated by the fact that multijets have no real $E_{\text{T}}^{\text{miss}}$ associated to an isolated neutrino. Instead, fake $E_{\text{T}}^{\text{miss}}$ originates from large fluctuations of the calorimeter energy response, jet energy measurement or entire jets falling out of the detector acceptance. Therefore, $E_{\text{T}}^{\text{miss}}$ tends to be small for multijet events and the low region of this distribution concentrates most of the multijet population, that in this way can be better constrained from data.

The fit is done separately in the analysis regions, 2 or 3 jets and 2 b-tags, and for muons and electrons to better adjust the multijet estimate to the particular analysis phase space, but inclusively in the p_{T}^{W} and $b-$tag categories. The multijet contribution is expected to be more important in the electron channel than in the muon channel since jets can more easily fake electrons than muons.

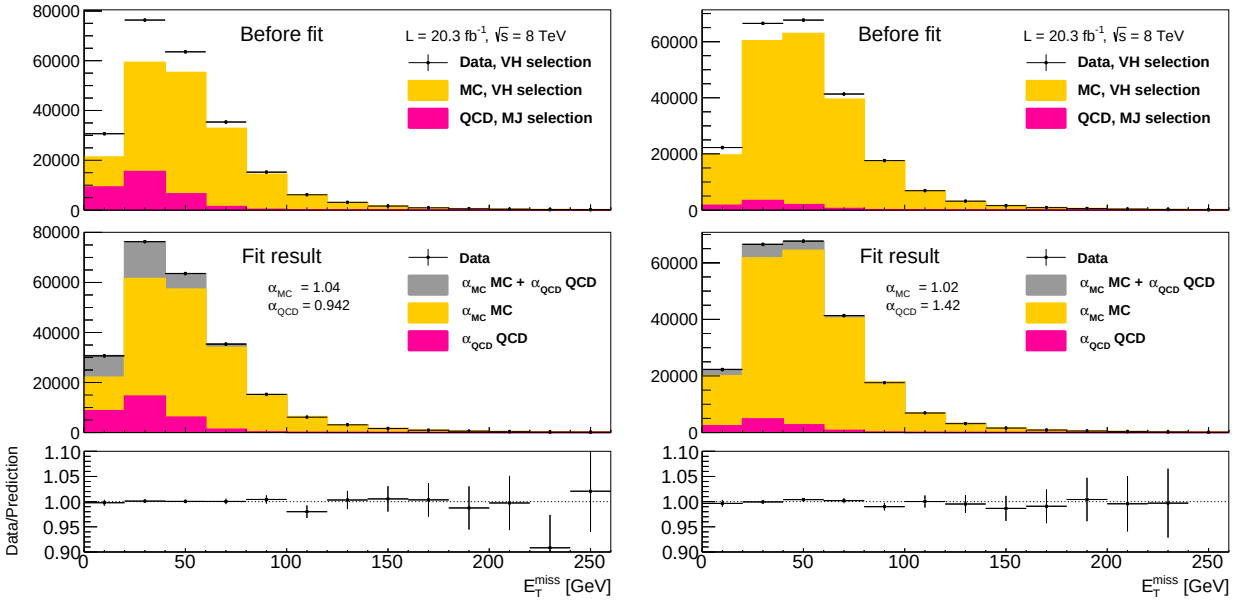
The fit consists of maximising a likelihood function using Poisson statistics, considering the statistical uncertainties on data and prediction [81]. Figures 6.23 and 6.24 show the outcome of the procedure. Data and prediction agree within statistical uncertainties after including the scaled multijet sample. Nevertheless, few surviving discrepancies are covered by the total systematic uncertainty as will be shown in Section 6.5.2, and can also be recovered through the final analysis fit, where the normalisation of the main backgrounds is finally adjusted. These were also the reasons why it was chosen not to force α_{MC} to one during the multijet fit.

Table 6.16 shows the scaling factors determined for both the multijet background and simulation. All the α_{MC} factors are consistent with one within 9%. On the other hand, the α_{QCD} factors range from 0.9 to 1.9 to normalise the template to the data observation. Data and prediction are systematically compared in all the analysis event categories and for different observables to validate this determination. The results are shown ahead in Section 6.4.4.

6.4.4 Distributions of key observables and intermediate Results

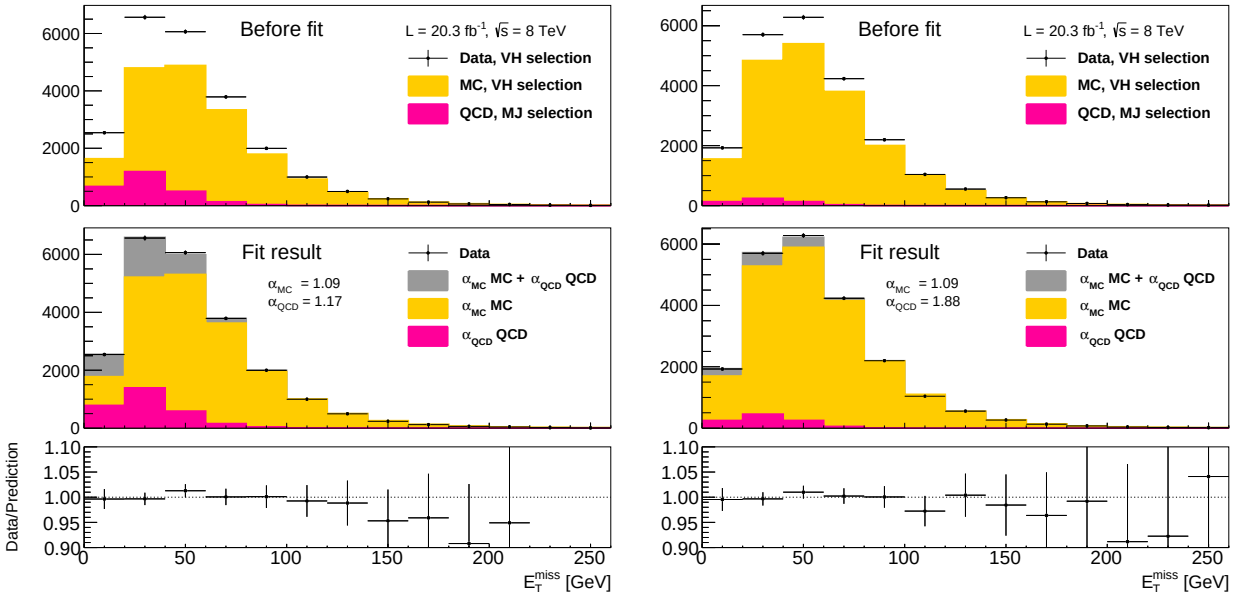
The m_{T}^{W} distribution is shown for the various event categories of the analysis in Figures 6.25 and 6.26. The events are further split into the electron and muon channels for the two b-tagged categories. The investigation of the data and prediction agreement is used as the method to validate the multijet sample estimate and the main simulated backgrounds modelling in general.

Overall, data and prediction agree within statistical uncertainty. The only exception is the



(a) 1 tag, 2 jets, $e-$channel

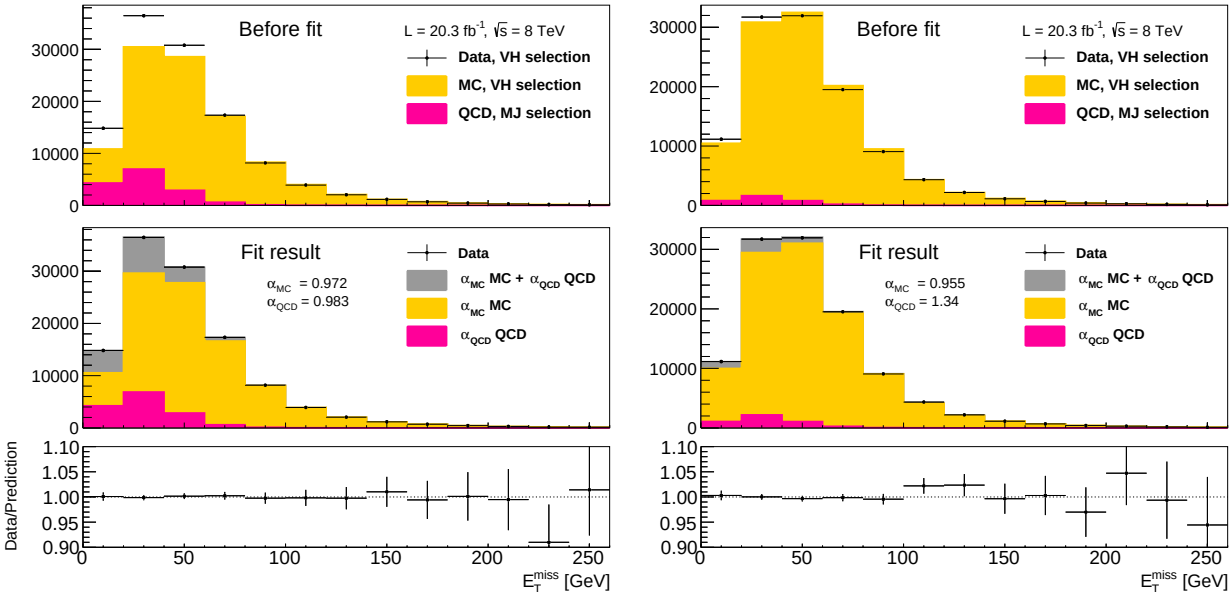
(b) 1 tag, 2 jets, $\mu-$channel



(c) 2 tag, 2 jets, $e-$channel

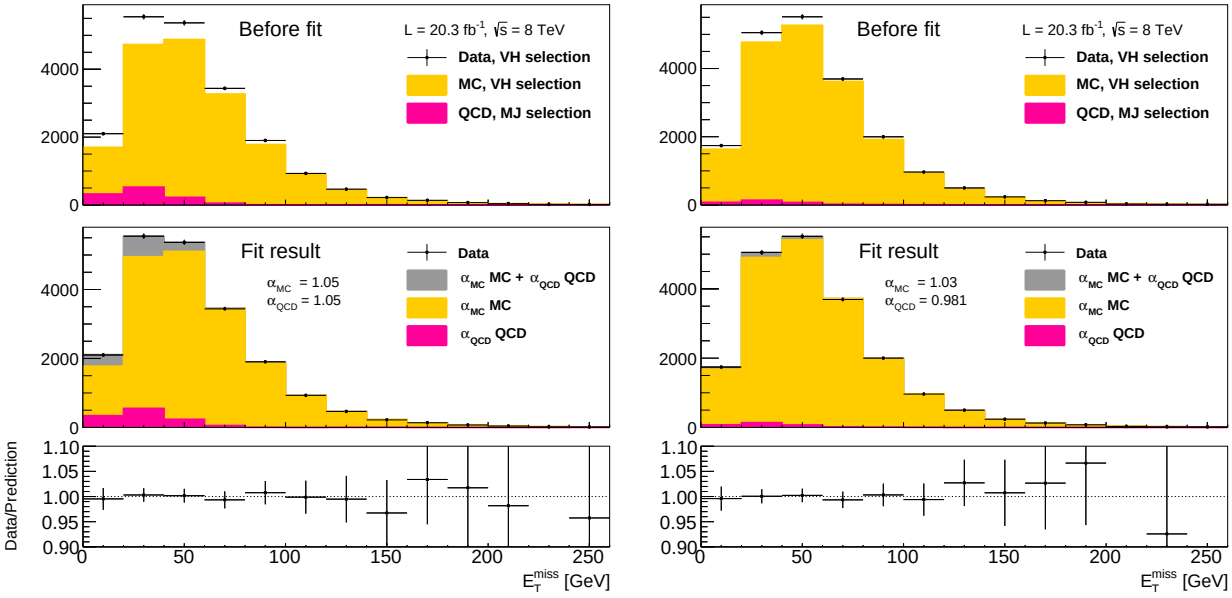
(d) 2 tag, 2 jets, $\mu-$channel

Figure 6.23: Fit procedure to determine the multijet template normalisation scale for events with 2 jets. Data, Monte Carlo and multijet template before (up) and after (centre) the fit. The data to prediction ratio after fit is shown in the bottom pannel.



(a) 1 tag, 3 jets, $e-$channel

(b) 1 tag, 3 jets, $\mu-$channel



(c) 2 tag, 3 jets, $e-$channel

(d) 2 tag, 3 jets, $\mu-$channel

Figure 6.24: Fit procedure to determine the multijet template normalisation scale for events with 3 jets. Data, Monte Carlo and multijet template before (up) and after (centre) the fit. The data to prediction ratio after fit is shown in the bottom panel.

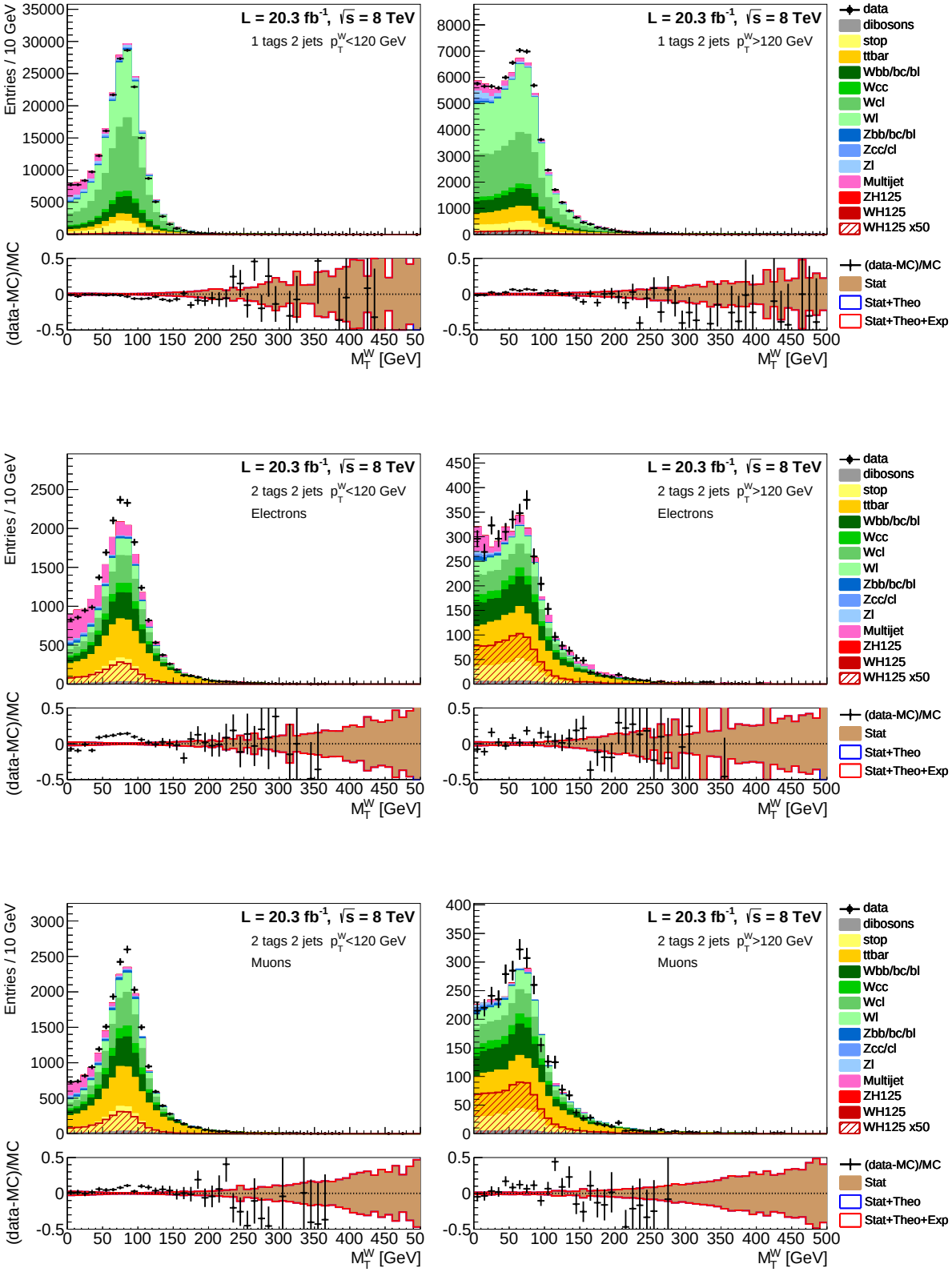


Figure 6.25: m_{T}^{W} distribution for data and prediction for events with 2 signal jets. The first row corresponds to the 1 tag category and the second and third rows correspond to the 2 tag category, separated in the electron and muon channel, respectively. The left and the right columns have the low and high p_{T}^{W} bins.

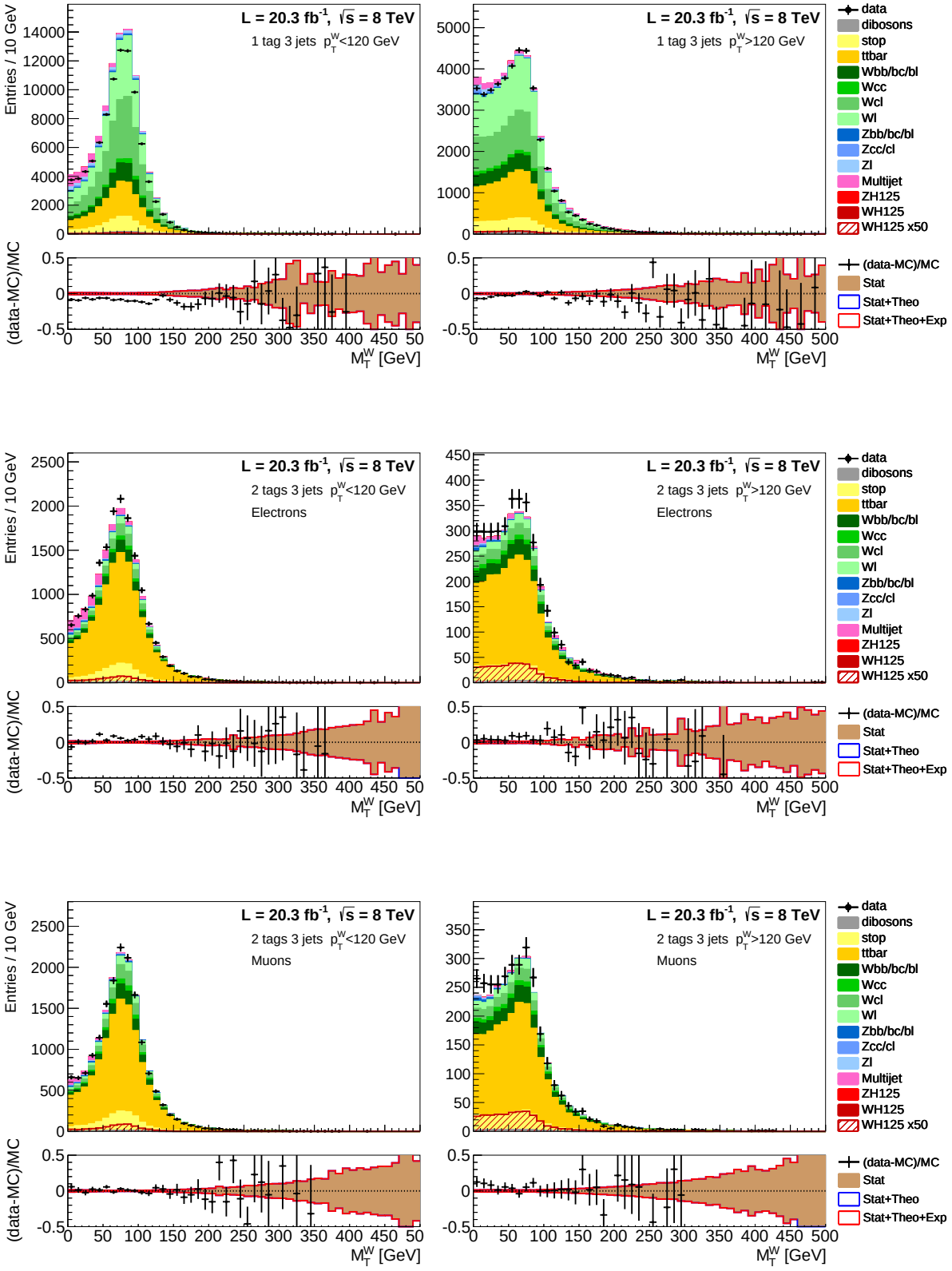


Figure 6.26: m_{T}^{W} distribution for data and prediction for events with 3 signal jets. The first row corresponds to the 1 tag category and the second and third rows correspond to the 2 tag category, separated in the electron and muon channel, respectively. The left and the right columns have the low and high p_{T}^{W} bins.

electron channel in the $p_{\mathrm{T}}^{W} < 120$ GeV region. Here, the multijet sample shape does not fit the data spectrum of m_{T}^{W}. This issue is known from the VH analysis of the Run I data and was reported before [50]. The discrepancy was never understood within the VH analysis group, and this category of events was removed from further analysis and will not be considered in the multivariate and statistical analysis that follow. For the other bins, the multijet sample shape appears adequate and the arising disagreement between data and MC can still be recovered by constraining the other background yields with data. Besides, they do not exceed 10%, which is the typical size of the systematic uncertainty.

The number of expected signal and background events after the analysis selection is shown in Table 6.17 and 6.18 for all the analysis categories, together with the observed data. Here, the electron channel considers only events with $p_{\mathrm{T}}^{W} > 120$ GeV in the 2 $b-$jets category. Data and prediction agree within 5%, the unique extreme deviation corresponding to 11.4% for events with two jets with tight $b-$tagging for $p_{\mathrm{T}}^{W} < 120$ GeV.

	2 jets							
	$p_{\mathrm{T}}^{W} < 120$GeV				$p_{\mathrm{T}}^{W} > 120$GeV			
	1 tag	LL	MM	TT	1 tag	LL	MM	TT
WH	37.2	10.3	16.1	18.1	27.6	8.2	12.5	14.2
qqZH	2.1	0.5	0.8	0.8	0.6	0.1	0.2	0.2
ggZH	0.09	0.01	0.02	0.02	0.06	0.01	0.01	0.02
$t\bar{t}$	9991.5	1627.9	2037.3	1937.1	5871.2	780.2	805.3	655.5
top	11281.4	910.2	782.4	606.1	3457.5	307.2	237.4	165.1
W+jets	163924	6164.5	1796.6	664.1	51185.2	2107.1	609.5	266.1
Z+jets	8058.3	360.9	156.7	93.1	2274.0	106.6	39.4	22.1
dibosons	1929.9	101.7	62.2	49.9	922.3	52.7	30.1	24.9
multijet	10353.4	650.0	271.9	105.8	2496.1	150.0	93.3	44.4
prediction	205578	9826.5	5124.4	3475.3	66234.8	3512.3	1828.2	1192.9
data	198555	9941	5499	3923	67606	3582	1952	1259
data stat uncertainty	446	100	74	63	260	60	44	35

Table 6.17: Expected and observed number of events for each analysis category for events with 2 signal jets. The number of expected events is normalized proportionaly to each process cross-section to the integrated luminosity of 20.3fb^{-1}of the $\sqrt{s} = 8$ TeV data set.

	3 jets							
	$p_{\mathrm{T}}^{W} < 120$GeV				$p_{\mathrm{T}}^{W} > 120$GeV			
	1 tag	LL	MM	TT	1 tag	LL	MM	TT
WH	14.0	2.9	4.5	5.1	13.5	3.2	5.0	5.6
qqZH	1.1	0.2	0.3	0.3	0.4	0.1	0.13	0.1
ggZH	0.1	0.01	0.02	0.02	0.07	0.01	0.01	0.02
$t\bar{t}$	18938.8	3072.3	3933.6	3964.8	10901.3	1380.8	1409.1	1176.3
top	7085.5	590.8	555.1	450.0	2969.3	278.7	219.7	160.3
W+jets	67494.5	2544.3	703.5	257.0	26828.3	1097.4	307.3	128.0
Z+jets	3914.3	172.0	71.2	43.5	1188.0	58.5	24.2	15.4
dibosons	1080.8	48.7	21.4	13.6	668.3	34.5	14.9	11.7
multijet	4429.2	183.0	76.0	49.6	1380.5	64.8	30.0	16.3
prediction	102958.3	6614.5	5365.9	4784.1	43949.9	2918.2	2010.5	1514.0
data	93361	6337	5552	4978	42561	3049	2072	1619
data stat uncertainty	305	80	23	70	206	55	46	40

Table 6.18: Expected and observed number of events for each analysis category for events with 3 signal jets. The number of expected events is normalized proportionaly to each process cross-section to the integrated luminosity of 20.3fb^{-1}of the $\sqrt{s} = 8$ TeV data set.

6.5 Multivariate Analysis

The search for $WH \to \ell\nu b\bar{b}$ events finds its major adversary in the small signal-to-background ratio. Hence, the ultimate challenge of this analysis is to obtain high signal detection efficiency while setting the background rejection to a maximum. But the usage of single observables rarely provide the discriminating power needed to fit this demand, specially when dealing with background processes that have topologies very similar to the signal. As an alternative to single observable discriminants, multivariate methods provide a way to combine several observables, and by taking advantage of a multi-dimensional perspective of the event, form a more powerful discriminant at the end. Such techniques are used broadly in HEP analysis nowadays, particularly for those involving searches for rare processes, where the traditional cut-based analyses meet their limitations.

The MVA method Boosted Decision Tree (BDT) is a machine learning classifier with features that satisfy the type of demand of the $WH \to \ell\nu b\bar{b}$ analysis. This method encloses a multi-dimensional cut technique to separate the signal and background phase space. With a BDT, the correlations between the input variables are explored to form an output able to much better separate the signal and background event classes. By doing so, the expected signal sensitivity of the WH search benefitted from a gain of $\sim 30\%$ relative to the cut-based analysis. This section describes this MVA technique, its configuration and how it is used in the context of the WH analysis. Furthermore, a study leading to the improvement of the BDT performance is presented.

6.5.1 Boosted Decision Trees

A decision tree (DT) [82] is a binary classifier formed by automatically finding the optimal splitting between two classes of events based on a sequence of single variable cuts. Each event is described by a set of discriminant observables and, at the starting root node, the full input sample is constituted by signal and background events in a one-to-one proportion. The method will determine and apply the cut that optimally separates the two event categories by scanning each input variable and, for each of them, determining the best separating cut. The criterion to evaluate the best separation can be configured but is generally based on the purity of the samples before and after the cut has been applied, where the purity p of a sample is defined as

$$p = \frac{N}{S+B} \tag{6.8}$$

where N is the number of events of a given class and S and B are respectively the number of signal and background events of the sample.

The resulting two split samples are then divided again based on the same or other of the input variables, employing the same method. The splitting will continue until a stopping criterion is fulfilled or the specified maximum tree depth has been reached. The process is referred to as the tree growing and is illustrated in Figure 6.27, where a schematic view of

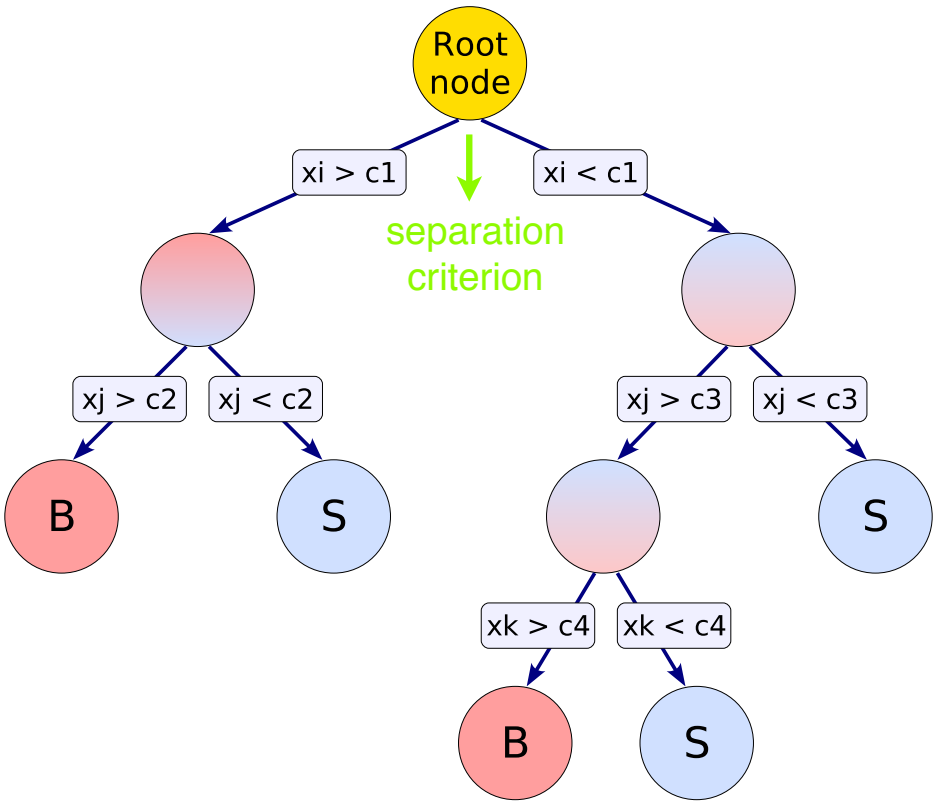


Figure 6.27: Schematic diagram of a decision tree of maximum depth 3. At the root node, the full sample is constituted by signal and background events in one-to-one proportion. At each successive node, a sequence of binary splits is applied to the sample until a stop criterion is fulfilled or the tree maximum depth has been reached. x_i represent the discriminant variables whereas c_j represent the correspondent variable cut at each node. In the final nodes, the events are classified in background (B) or signal (S) depending on the purity of the node sample. Adapted from [82].

a decision tree of maximum depth 3 is shown. The events in a final node, or "leaf", are classified according to the leaf purity into signal or background. The decision tree classification is encoded in an output weight that assigns -1 to events in the background-like leaves and +1 to events in signal-like leaves.

Adaptive Boost

The liability of the decision tree is its high sensitivity to statistical fluctuations of the training sample, which determines the whole tree structure. The boosting technique, illustrated by Figure 6.28, compensates this flaw by generating a "forest" of DTs instead of just one [82]. A few boosting algorithms exist. In the adaptive boost case, each new tree growth is based on the same initial sample, with the events of that sample being assigned a new weight α_i based on the classification from the previous ith tree defined as

$$\alpha_i = \left(\frac{1-\varepsilon}{\varepsilon}\right)^{\beta} \tag{6.9}$$

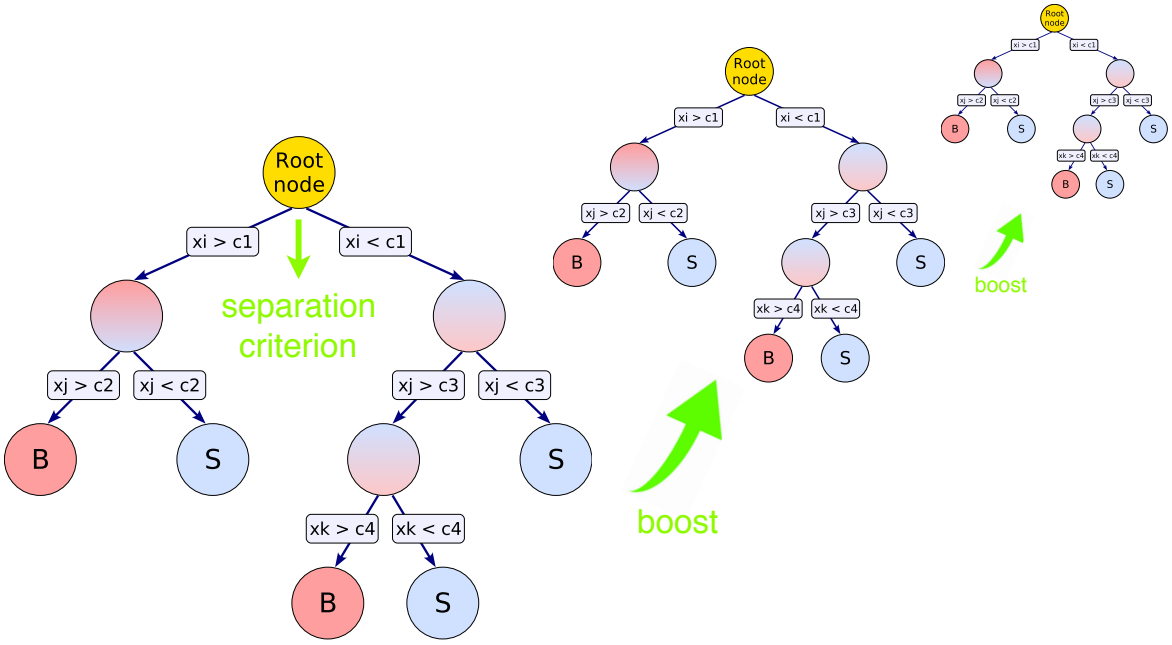


Figure 6.28: Schematic diagram of a boosted decision tree. Events that were wrongly classified by a decision tree are boosted during the next decision tree growth. Adapted from [82].

where β is a configurable real value of the adaptive boost algorithm, and ε is the fraction of mis-classified events of the ith decision tree. For $\varepsilon < 0.5$, which is the standard case of a classifier method, mis-classified events receive a higher statistical weight for the next tree growing step. Then, the training sample is reweighted to recover the signal and background 1:1 proportion. In this way, the sample composition at the root node of each new tree is composed of the same events but with a different statistical distribution, affecting the structure of each tree.

In a BDT, the event classification results from the boost weighted average of the classifications attributed by each tree in the forest and is given by

$$\text{BDT weight} = \frac{1}{N}\sum_{i}^{N} \ln(\alpha_i) h_i \tag{6.10}$$

where N is the number of trees in the forest and h_i is the individual output weight classification of the ith tree. So, with the boosting technique, events that are less trivial to classify are given a special attention, and the influence of statistical fluctuations of the input sample on the actual BDT classification is mitigated.

The BDT growth is known as training phase. Here, the MVA method is trained to recognise signal events and distinguish those from the background based on the knowledge that the MC provides. The result is a customised BDT that can be applied to real data at a later phase.

Option	Value
Separation criterium for node splitting	Gini Index
Number of steps during node cut optimization	100
Minimum number of events required in leaf node	100
Maximum depth of the tree	4
Number of trees in the forest	200
Boosting type for the trees	Adaptive
β parameter of the Ada Boost algorithm	0.15

Table 6.19: Boosted decision tree parameters used in the $WH \to \ell\nu b\bar{b}$ analysis.

BDT performance

The BDT technique became popular in recent experimental HEP applications and is now one of the most used machine learning methods. When compared with other techniques such as Neural Networks (NN) or Support Vector Machines (SVM) of similar performance, the BDT outperforms due to its simplicity. For this reason, it is faster at both the training and classification phases than NNs and SVMs. Besides, a decision tree has a straightforward interpretation enabling an intelligible view of the method, which is not always the case of competitor methods. Additionally, the BDT method demands little tuning effort regarding configuration parameters yielding a good performance without much optimisation. Another clear advantage of the BDT is its robustness with respect to non-discriminant variables. By method construction, these variables are ignored since the best separation variable is searched for at each step. In the NN and SVM cases, this can be a drawback since they can deceive the classification.

6.5.2 The $WH \to \ell\nu b\bar{b}$ BDT

The WH BDT implementation was done using the TMVA (Toolkit for Multi-Variate Data Analysis) code package version 4.1.2 [82]. With this tool, many of the BDT parameters can be configured to better adjust to the analysis needs. The number of trees in the forest, the splitting criteria or maximum depth of the tree clearly affect the BDT structure.

Method Configuration

The parameters used to train the BDTs of the $WH \to \ell\nu b\bar{b}$ analysis are summarised in Table 6.19. This setup resulted from the optimisation of each of the parameters as a function of the signal-to-background ratio. The optimisation was studied by colleagues participating in the VH analysis and reported internally in ATLAS [83].

A large number of trees in the forest, for instance, has in principle a positive impact on the signal-to-background ratio. It means that the trained BDT is more robust with respect to the classification of non-trivial events, but it comes at a cost of computational time. On the other hand, at some point, there is little benefit from adding more trees and the computing time

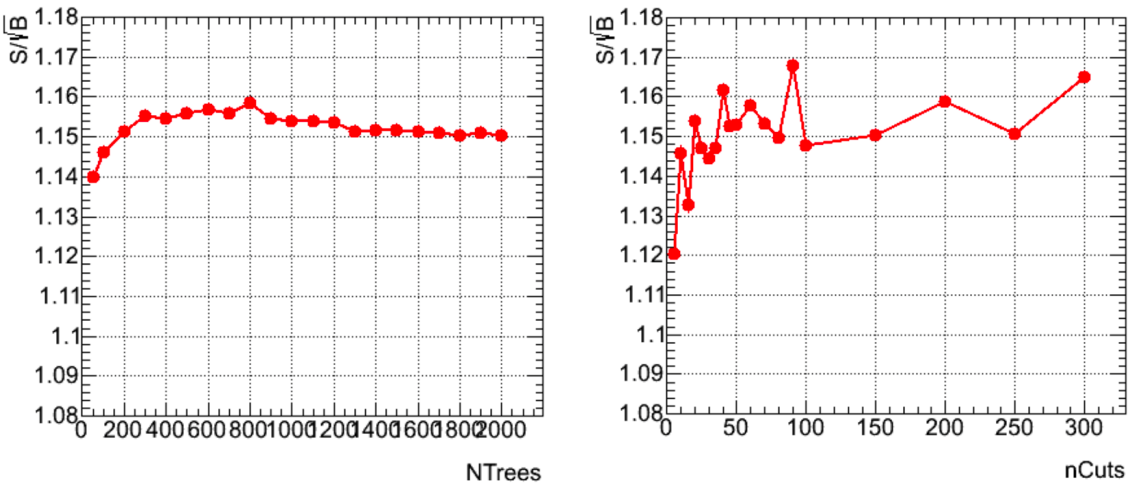


Figure 6.29: Signal sensitivity $S/\sqrt{B}$ as a function of (left) number of trees in the forest and (right) number of steps during node cut optimization. Taken from an ATLAS internal report [83]

argument prevails. Thus, the choice for this parameter results from the compromise between these two competing effects as Figure 6.29 shows.

The same holds for the number of steps during the optimisation of each node cut. This parameter defines the granularity with which the input variables spectra are scanned during cut optimisation at each DT node, with the spectra being equally divided by this quantity. In theory, the larger the value the better, as it corresponds to a finer optimisation, but the gain in sensitivity ends by saturating due to the fact that in practice unlimited statistics are not available, as Figure 6.29 proves. The remaining setup was determined in a similar way.

Maximum number of events required in the leaf node is the minimum number of events in a node required for the splitting to occur. This number prevents the tree from growing from nodes that are statistically insignificant and is set to 100 for this analysis. This is one of the stopping criteria mentioned before.

Gini Index is used to quantify the signal and background separation gain at the moment of cut optimisation in the tree nodes. The Gini Index is defined as $p(1-p)$, where p represents the sample purity. Based on this, a given cut on a given variable is optimal if it maximises the difference between the Gini Index of the parent node and the sum of the Gini Indices of the daughter nodes, each weighted by their relative sample fractions. This basically means increasing the background and signal separation.

Boost Type is the adaptive boost as defined previously with β parameter of Eq. 6.5.1 equal to 0.15.

Variable	Definition
$m_{b\bar{b}}$	Invariant mass of the $b-$jet pair
$\Delta R(b_1, b_2)$	Jets separation in the (η, ϕ) plane
$\Delta\phi(W, b_1 b_2)$	ϕ separation between W and b jet pair
$\Delta\phi_{min}(\ell, b_i)$	Minimum ϕ separation between the lepton and each of the $b-$jets
$p_{\mathrm{T}}^{b_1}$	Transverse momentum of the leading jet
$p_{\mathrm{T}}^{b_2}$	Transverse momentum of the sub-leading jet
MV1c(b_1)	Flavour weight of the leading jet
MV1c(b_2)	Flavour weight of the sub-leading jet
p_{T}^{W}	W transverse momentum
m_{T}^{W}	W transverse mass
$E_{\mathrm{T}}^{\mathrm{miss}}$	Missing transverse energy
$p_{\mathrm{T}}^{j_3}$	Transverse momentum of the third jet for events in the 3 jet category
$m_{b\bar{b}j}$	Invariant mass of the tri-jet system

Table 6.20: Set of discriminant variables used in the boosted decision tree method. $p_{\mathrm{T}}^{j_3}$ and $m_{b\bar{b}j}$ are additionally used in the 3 jets 2 b-tags category.

Input Variables and Modelling

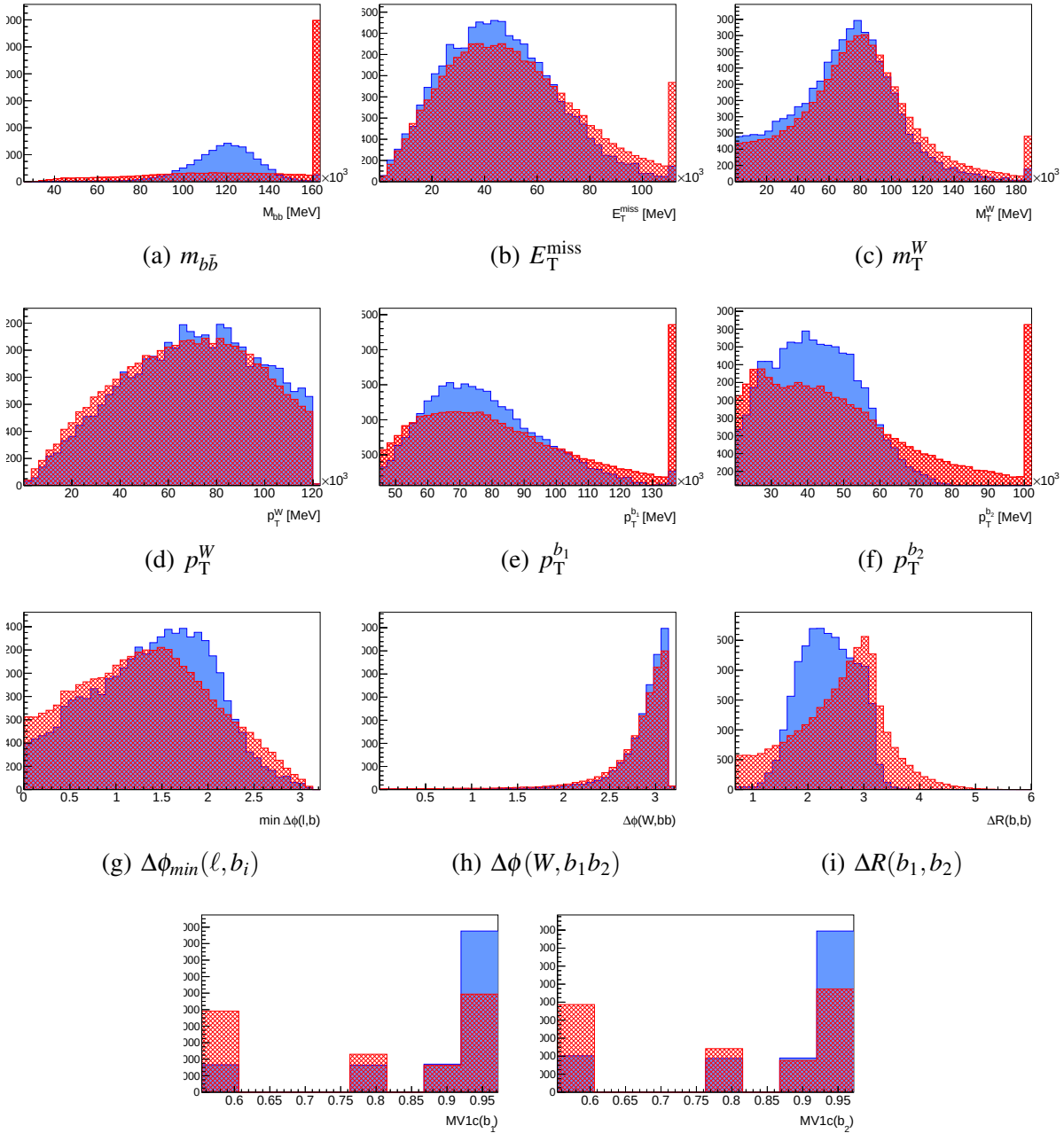
The input discriminant variables used as a baseline and their definition are summarised in the Table 6.20. These were chosen based on their impact on the signal and background separation power of the BDT output discriminant.

During the optimisation of the input variable set, two key features are desired in order to take full advantage of the BDT intrinsic machinery:

Variable discrimination In principle, the better the discriminant power of the input variables, the better the BDT output.

Correlation with the other variables The core of the BDT method precisely explores the correlation between different variables to classify the events. However, one adds little information to the method when using variables that are fully correlated with others already present. The most suitable situation is having variables that are differently correlated for signal and background event types, for this allows the BDT to access specific regions of the phase space where the discrimination is augmented.

The distributions of the input variables are shown in Figure 6.30 for events with two $b-$jets in the category of lower p_{T}^{W}. For the majority of the variables, the signal and background shapes are very similar, manifesting the difficulty in discerning between background and signal events in the WH search. The exceptions are $m_{b\bar{b}}$, $\Delta R(b_1, b_2)$ and $p_{\mathrm{T}}^{b_2}$. For signal events, $m_{b\bar{b}}$ corresponds to the resonance mass peak while for most of the backgrounds its spectrum is continuous. $p_{\mathrm{T}}^{b_2}$ is in average smaller for signal events. Since $p_{\mathrm{T}}^{b_1}$ is chosen as the leading p_{T}



(a) $m_{b\bar{b}}$ (b) $E_{\mathrm{T}}^{\mathrm{miss}}$ (c) m_{T}^{W}

(d) p_{T}^{W} (e) $p_{\mathrm{T}}^{b_1}$ (f) $p_{\mathrm{T}}^{b_2}$

(g) $\Delta\phi_{min}(\ell, b_i)$ (h) $\Delta\phi(W, b_1 b_2)$ (i) $\Delta R(b_1, b_2)$

(j) MV1c(b_1) (k) MV1c(b_2)

Figure 6.30: Distributions of the BDT input variables for signal (blue) and background (dashed red) simulated events with 2 $b-$tagged jets and $p_{\mathrm{T}}^{W} < 120$ GeV. The signal distribution is normalised to the background integral. The last bin of some distributions contemplates the sum of upper values entries.

jet, and given the signal Higgs mass restriction, $p_{\mathrm{T}}^{b_2}$ is limited to lower values of the jet p_{T} spectra. The same restriction does not exist for the majority of the backgrounds. A similar reasoning holds for $\Delta R(b_1, b_2)$. Since for signal events the $b-$jets come from the Higgs decay, the separation between jets depends on the boost of the parent. For events with $p_{\mathrm{T}}^{W} < 120$ GeV the jets are supposed to be widely separated in signal events but correlated.

To prevent granularity losses during the WH BDT cut optimisation step, the upper tails of the distributions were merged in one cut-off value for some cinematic observables, as seen in Figure 6.30. This is only used when the merged region is mostly populated by background events, and the cut-off value must assure that the signal distribution is practically non-affected.

The extent of correlation between the input variables is measured by the correlation coefficient that translates the degree of linear dependence between two variables in a normalised manner, such that +1 (-1) means that two variables are totally correlated (anti-correlated) and 0 indicates no correlation at all. This coefficient is shown in Figure 6.31 separately for the signal and main backgrounds. In the signal case, the variables present a poor degree of correlation in general, indicating that there is no redundancy in the signal description received by the BDT. The same does not happen for the background samples. For instance, $m_{b\bar{b}}$ has low correlation with others for the signal while for backgrounds is highly correlated with $\Delta R(b_1, b_2)$ and $p_{\mathrm{T}}^{b_2}$. The p_{T} of the signal jets are too differently correlated for signal and background: while the jets from the signal Higgs boson must share the available energy of the decaying parent turning $p_{\mathrm{T}}^{b_1}$ and $p_{\mathrm{T}}^{b_2}$ anti-correlated, in background events the most common is for the two jets to split equally the available energy.

When using a multivariate method that uses and combines different observables in an automatic manner, one needs to carefully check that the simulation correctly models the observables. This is an important step to ensure the correct modelling of the multivariate method output. Since the BDT employs a sequence of variable cuts, any mismodelling translates into a different cut efficiency for data and simulation, and therefore results in an output discriminator prediction that data does not follow. For this reason, the continuum spectrum of the MV1c flavour weight is transformed into a four-value spectrum according to the available calibrated efficiency points of the $b-$tagging algorithm, as Figure 6.30 shows. Moreover, the agreement of data and MC has been assessed for all the BDT input variables. Figure 6.32 shows the distributions of the input discriminants of the WH BDT for the sample of 2 jets, both $b-$tagged, and p_{T}^{W} <120 GeV. Both the statistical and systematic uncertainties, as described in Section 7.1, are exhibited in the plots. Data and MC agree within uncertainties for all the input variables. Appendix C has the same distributions for the other analysis regions.

Training

The training of the WH BDTs is done separately for the following analysis categories:

- jet multiplicity: 2 or 3 jets;
- two p_{T}^{W} intervals: below or above 120 GeV.

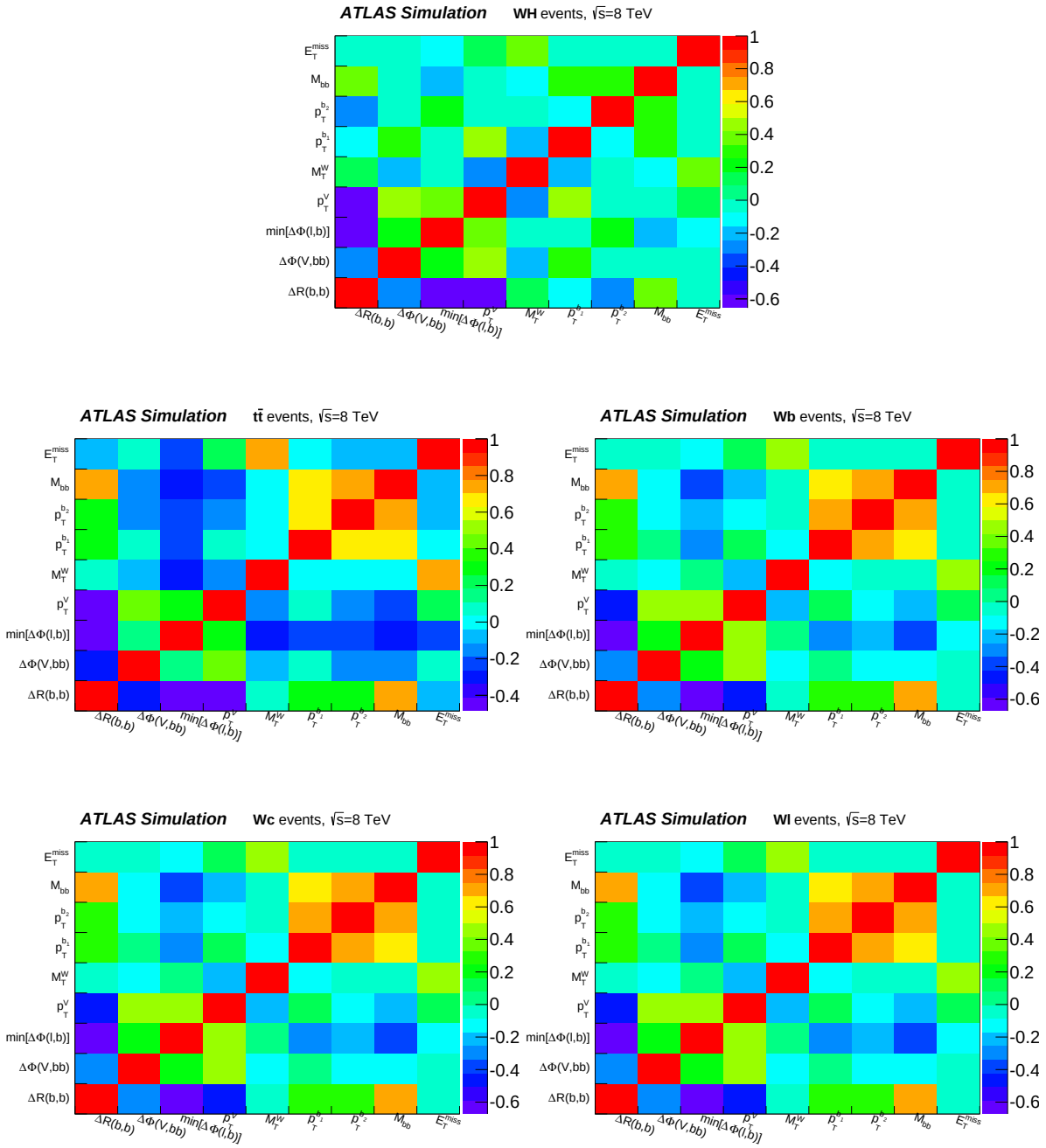


Figure 6.31: Correlation matrices of the BDT input variables for simulated WH, $t\bar{t}$ and $W+b,c-$ and light jets events with 2 $b-$tagged jets and $p_{\mathrm{T}}^{W} < 120$ GeV.

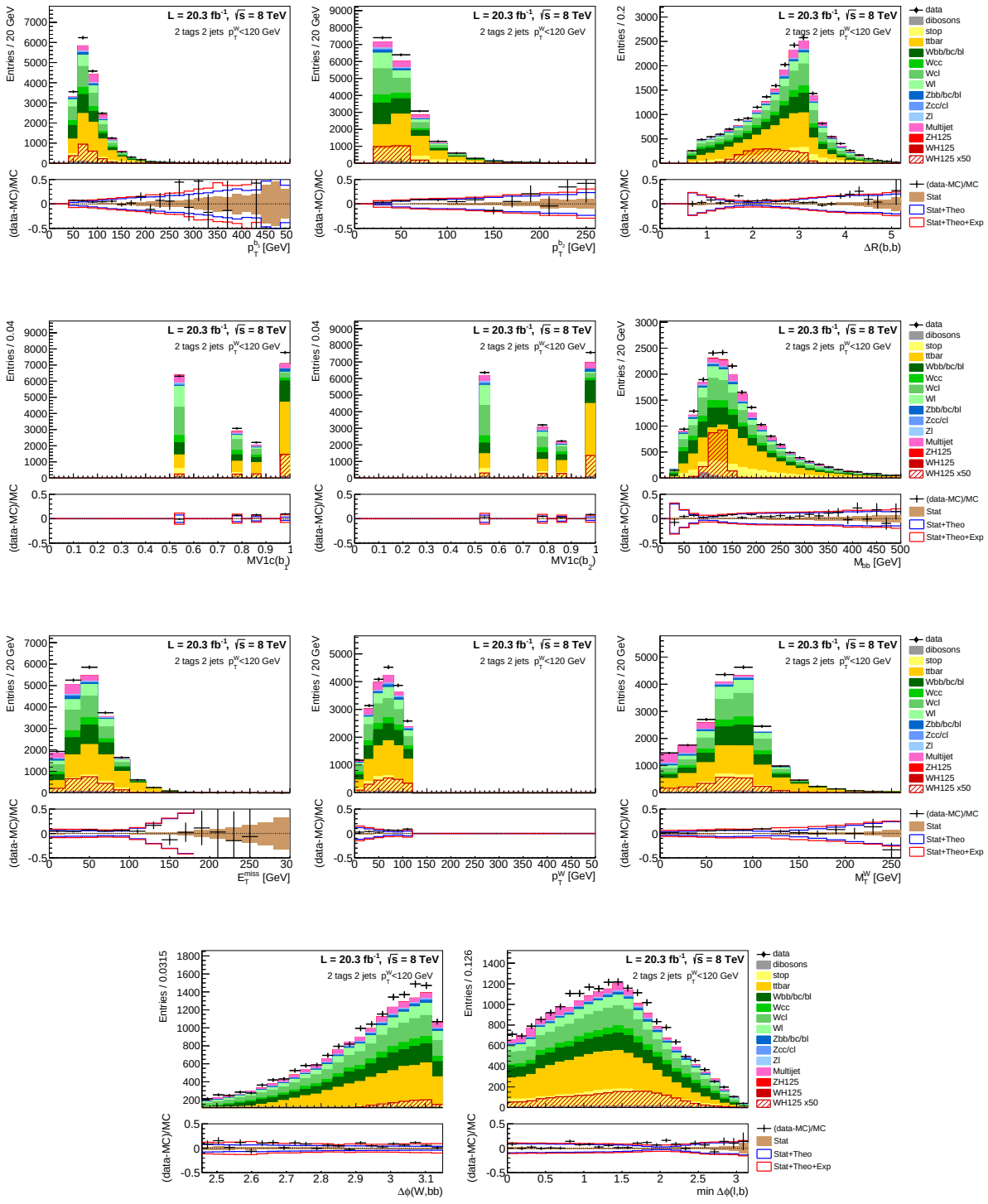


Figure 6.32: Distribution of the BDT training variables for data and prediction for events with two signal jets, both $b-$tagged, and $p_{\mathrm{T}}^{W} < 120$ GeV. The uncertainty bands are exhibited in the bottom plot as detailed in Section 7.1.

		Signal	Background
2 jets	$p_{\mathrm{T}}^{W} < 120\sim$GeV	29532/29470	1262972/1263164
	$p_{\mathrm{T}}^{W} > 120\sim$GeV	11101/11033	712335/712672
3 jets	$p_{\mathrm{T}}^{W} < 120\sim$GeV	9547/9723	746440/747026
	$p_{\mathrm{T}}^{W} > 120\sim$GeV	4814/4802	479321/479775

Table 6.21: Number of signal and background events used in the BDT training of the even/odd $k-$fold samples.

This allows building classifiers that adjust better to each category and obtain an overall better performance. A common risk inherent to the usage of multivariate methods is the overtraining. This usually arises when the training is done over statistically limited samples and the resultant BDT is too adapted to resolve the training sample. In order to control and overcome this feature, the following $k-$fold strategy is adopted during the WH BDTs training.

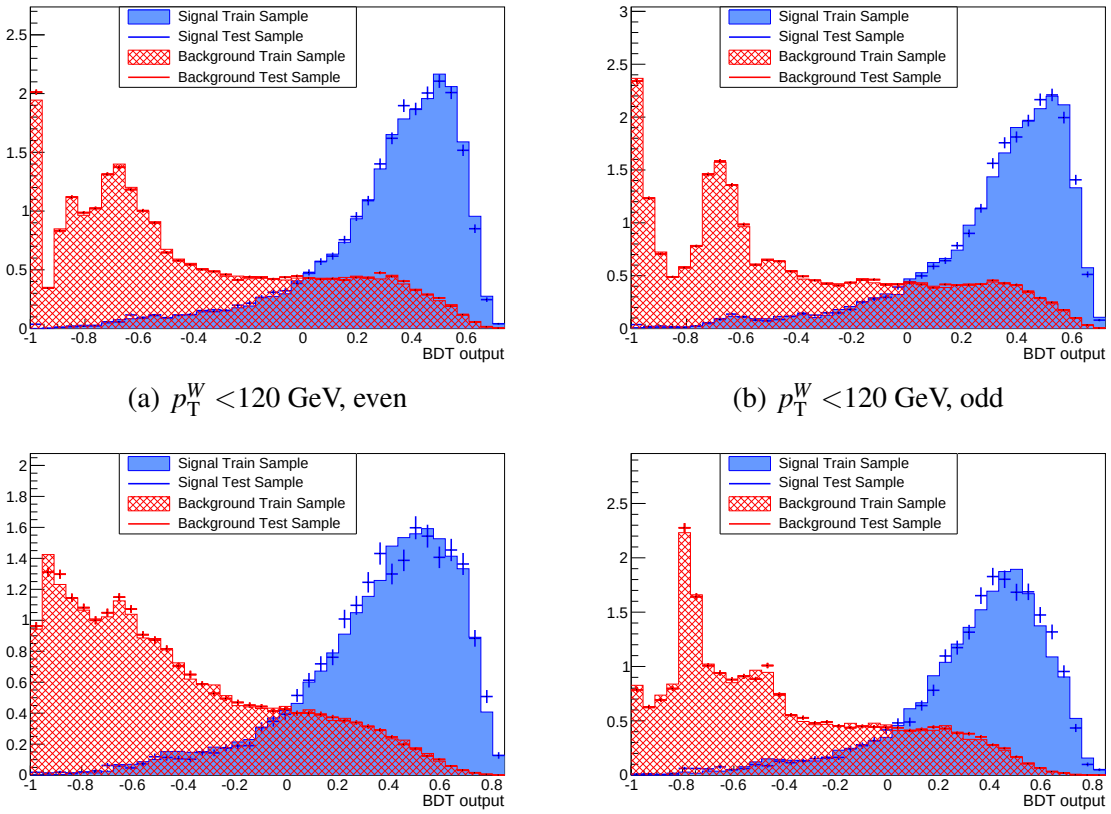
Even/Odd event number The training data is split into two samples based on the parity of the event number. Even events are used to train the BDT that will afterwards be applied to odd events and vice-versa. This is especially relevant to diminish any overtraining effect on MC.

Test sample The training data is further split in half at the beginning of the training phase. Half the sample is effectively used for the tree growth and then the resulting trained classifier is applied to the other data half, that serves as an independent test sample to check for overtraining.

Although this method is used to limit the overtraining effects, the most effective manner to avoid them is to use samples with large statistics in the training. The selection described in Section 6.4 deals exactly with this issue. But if on one hand the selection should keep as many events as possible to help the BDT understanding, on the other hand it should refine the selected sample to help the BDT to concentrate on difficult events. The ideal is to compromise between the two approaches. So, whenever a simple observable cut can be applied it should be done at event selection stage in order not to waste the BDT resources with a significant part of the phase space that can be dealt with a single cut. And in order to assure a rich statistical representation of data, other methods are complementarily used, such are the cases of truth tagging described at Section 6.3.4 and continuous tagging. By jointly using three efficiency working points of the $b-$tagging algorithm and having the MV1c output as a BDT input variable, the BDT is free to better explore the algorithm efficiency without compromising statistics.

Table 6.21 contains the total number of signal and background events used in the training of each WH BDT. The minimum number of events used for training occurs for the signal samples of the 3 jets and $p_{\mathrm{T}}^{W} > 120$ GeV BDT category, but the associated statistical uncertainty is below 1.5%.

Figures 6.33 and 6.34 show the output of the WH BDTs for the training and test sample for



(a) p_{T}^{W} <120 GeV, even

(b) p_{T}^{W} <120 GeV, odd

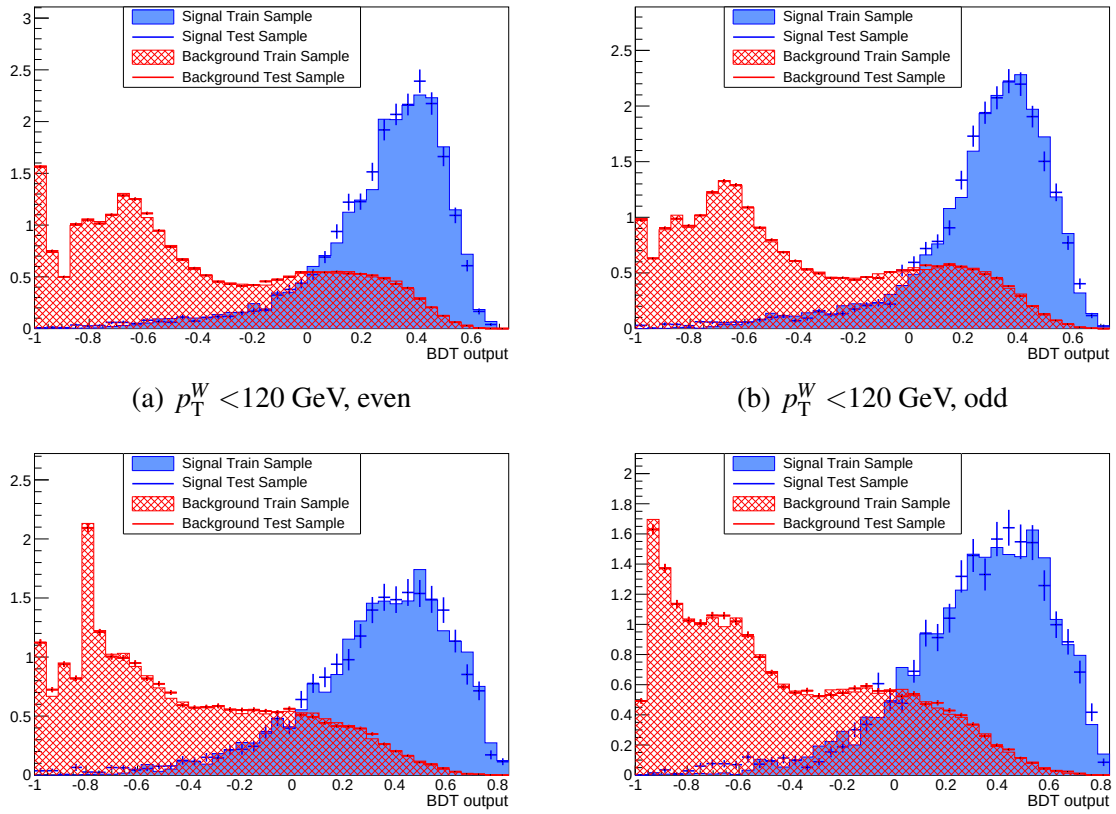
(c) p_{T}^{W} >120 GeV, even

(d) p_{T}^{W} >120 GeV, odd

Figure 6.33: Distribution of the BDT output trained from 2 jet events for the signal and background event types of the training and test simulated samples.The signal distribution is normalised to the background integral. The statistical uncertainty of the test sample distribution is displayed in the error bars.

2 jet and 3 jet events, respectively. They illustrate the ability of the method to form a powerful discriminant of signal and background events starting from the set of observables shown at Figure 6.30. The overtraining control is carried out by comparing the compatibility between the output weights for the test and training samples. The test and train BDT weight spectra are convergent within statistical uncertainties for both signal and background events and therefore overtraining issues are not observed.

The boost weight, defined as w in Eq. 6.5.1, resultant of the WH BDT training is shown in Figure 6.35 as a function of the decision tree number for events of the 2 jets and 2 $b-$tags category with $p_{\mathrm{T}}^{W} > 120$ GeV and even $k-$fold. The same distribution is shown for the error fraction ε. The boost weight is always greater than 1, meaning that more statistical relevance is put on mis-classified events. It starts at nearly 1.25 and decreases, asymptotically approaching 1. The error fraction increases with the decision tree number. This, however, should not be interpreted as a loss of performance of the new trees. The fact that mis-classified events, that are associated with the most indistinguishable ones, are systematically boosted increases the error rate despite the effective number of fails could be reduced.



(a) p_{T}^{W} <120 GeV, even

(b) p_{T}^{W} <120 GeV, odd

(c) p_{T}^{W} >120 GeV, even

(d) p_{T}^{W} >120 GeV, odd

Figure 6.34: Distribution of the BDT output trained from 3 jet events for the signal and background event types of the training and test simulated samples. The signal distribution is normalised to the background integral. The statistical uncertainty of the test sample distribution is displayed in the error bars.

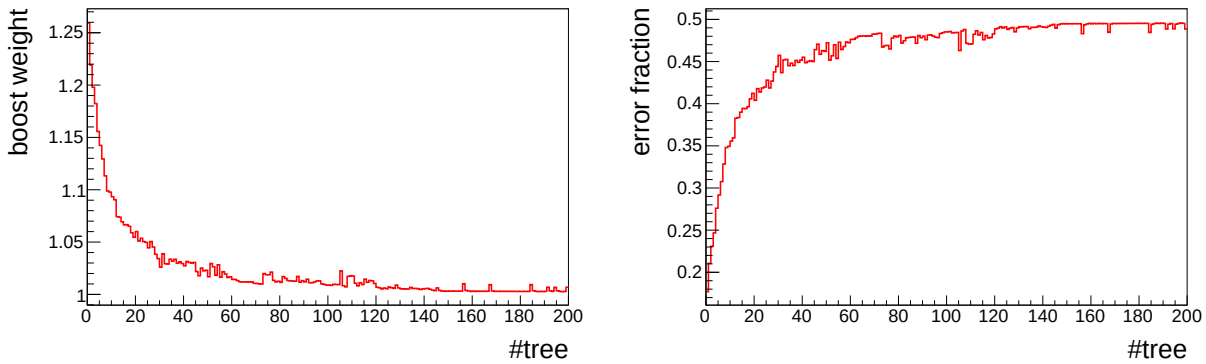


Figure 6.35: (Left) Adaptive boost weight α and (right) error fraction ε as a function of the decision tree number.

Example of a Decision Tree

Figure 6.36 shows an example of a single decision tree of the $WH \to \ell\nu b\bar{b}$ analysis. Although this monitoring procedure could not be done systematically, it is still enlightening to carry it out as an exercise.

The DT starts by selecting the Higgs mass peak by applying two subsequent cuts on $m_{b\bar{b}}$. With the first one, it is able to isolate a subset of events with $m_{b\bar{b}} > 147$ GeV with signal purity of 7.5%. This turns out to be considered a leaf, not because of any stopping criteria but because no other variable is able to separate more the two event classes. All events in this node are classified as background and attributed the encoded weight of -1. The same happens after the second $m_{b\bar{b}}$ cut for events with $m_{b\bar{b}} < 89.4$ GeV.

At the third layer, the MV1c of the p_{T}-leading jet is used to obtain a sample with signal purity of 0.8 from one of 0.72. The same holds for the subsequent cuts appearing at the DT depth 4 using $E_{\mathrm{T}}^{\mathrm{miss}}$ and the MV1c of the p_{T}-subleading jet, the first one being less relevant in terms of separation gain. Here, the tree growing stops for it has reached the maximum depth defined in the setup and the final leaves are classified.

Starting from a fully mixed sample, the decision tree algorithm is able to construct two regions of the event phase space where the signal purity is enriched to 81% and 69%. The DT classification can be interpreted through its output distribution showed also in Figure 6.36. For this single tree, the classifier weight admits only two values, ± 1. Only after the boost, the BDT output becomes a continuous value resultant from the weighted average of each tree classification. For this tree, the misclassification rate is nearly 20% as the plot shows. These misclassified events are the ones to boost on the next classifier growth.

Application

The customised BDTs resultant of the training with simulated signal and background samples are afterwards applied to MC and real data events. The $k-$fold method is adopted as stated before, where BDTs trained from even events are applied to odd events and vice-versa. As discussed, this mitigates residual overtraining effects on simulation. Figure 6.37 shows the BDT discriminant for data and MC for different analysis categories. Data and MC agree within the prediction uncertainty and the data statistical error. This results from the correct modelling by the MC of all variables used.

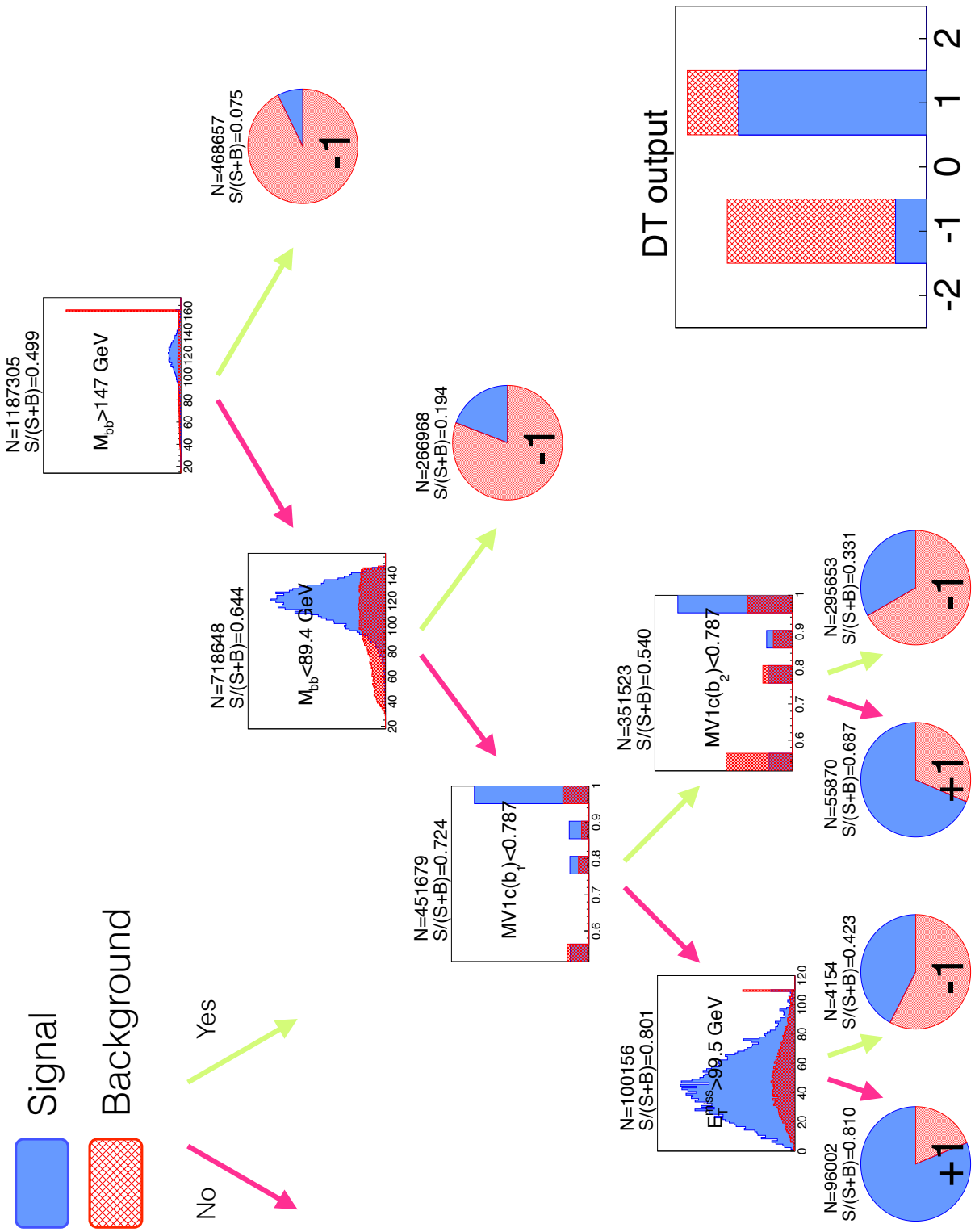


Figure 6.36: First decision tree resultant from the training with events of the 2 $b-$tagged jets, $p_{\mathrm{T}}^{W} < 120$ GeV and $k-$fold=0 category.

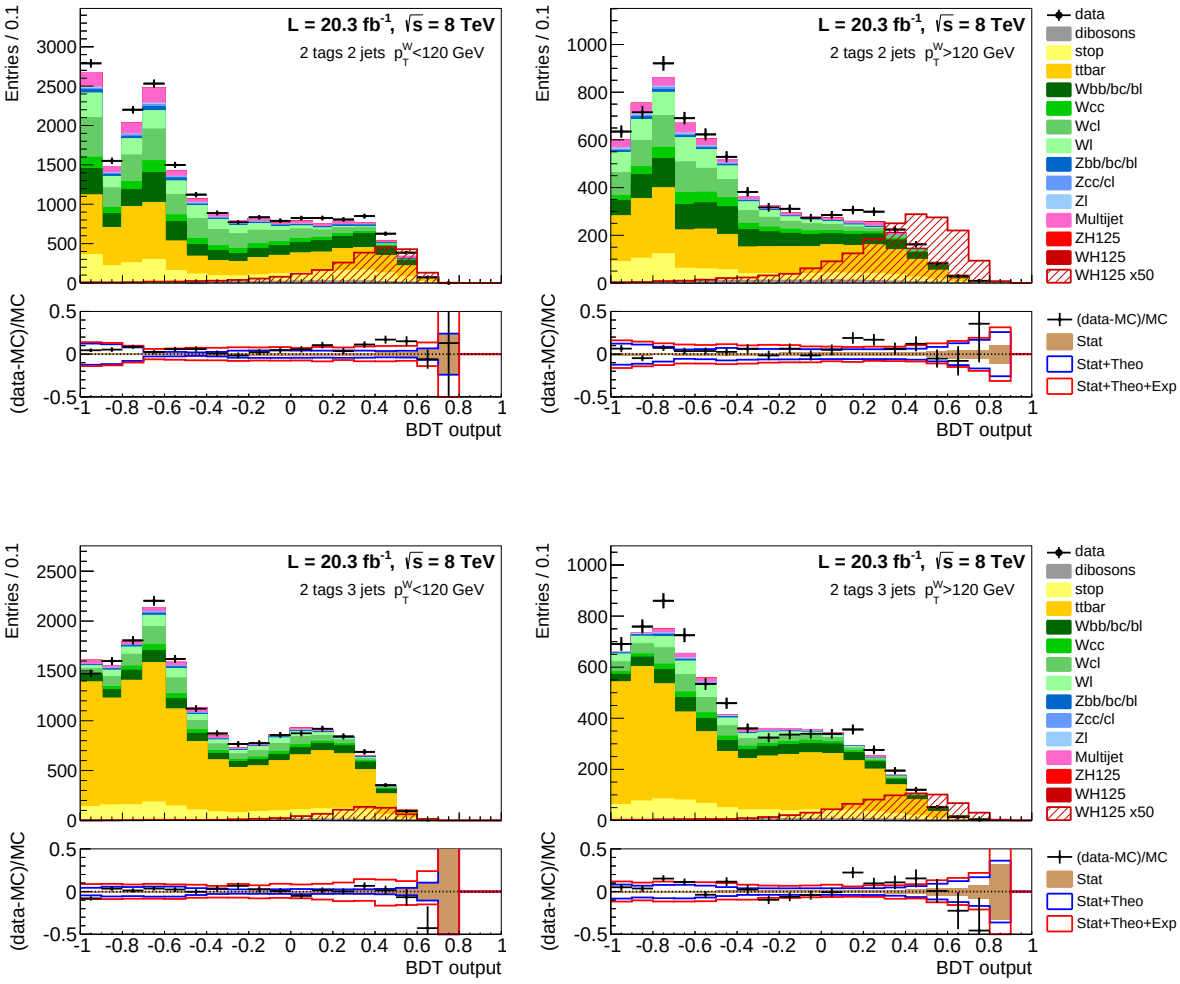


Figure 6.37: Distribution of the BDT output for data and simulation for the 2 and 3 jets and p_{T}^{W} below and above 120 GeV event categories. The uncertainty bands are exhibited in the bottom plot as detailed in Section 7.1.

Variable	Definition
Acoplanarity	$\lvert\pi - \lvert\Delta\phi(b_1,b_2)\rvert + \lvert\pi - \sum\theta_{b_i}\rvert$
Centrality	$\sum p_{\mathrm{T}}/E_{\mathrm{vis}}$
Aplanority/Aplanarity	determined from the Spherocity/Sphericity tensor
Spherocity/Sphericity	determined from the Spherocity/Sphericity tensor
Helicity($b_1(b_2)$)	defined in Figure 6.38
Azilicity($b_1(b_2)$)	defined in Figure 6.38
τ_{bb}	defined in Equation 6.11
$\tau_{\ell\nu}$	defined in Equation 6.11
$\Delta p_{\mathrm{T}}(W,\ell)$	W and ℓ p_{T}-difference: $p_{\mathrm{T}}^{W} - p_{\mathrm{T}}^{\ell}$
$p_{\mathrm{T}}^{b_1} + p_{\mathrm{T}}^{b_2}$	b_1 and b_2 p_{T}-scalar sum
$m_{Wb_{1(2)}}$	Invariant mass of the W + $b_{1(2)}$ system
$\Delta R(H,b_{1(2)})$	Separation between the Higgs candidate and $b_{1(2)}$
$\Delta Y(H,b_1)/\Delta Y(H,b_2)$	Rapidity difference between the H and $b_{1(2)}$
$\Delta Y(W,H)$	Rapidity difference between the Higgs and the W
θ Asymmetry	$(\theta_{b_1} - \theta_{b_2})/(\theta_{b_1} + \theta_{b_2})$

Table 6.22: Set of discriminant variables tested with the boosted decision tree method. b_1 and b_2 stand for the $p_{\mathrm{T}}-$leading and $p_{\mathrm{T}}-$subleading signal $b-$jets, respectively.

6.5.3 Optimisation of the $WH \to \ell\nu b\bar{b}$ BDT

The usage of the BDT technique increased the WH search sensitivity by 30%. This impact is a result of the method optimisation and customisation to this particular channel search. The input variable set is, out of many BDT parameters, the one that most defines its performance concerning signal and background separation. Therefore, the choice of the used observables appears as a natural subject for studies targeting the improvement of the BDT. Under this consideration, a study evaluating the impact of new observables in the BDT discriminant power was conducted. Each new variable was added at a time to the nominal set of input variables and the impact quantified in terms of additional background rejection.

Input variables

This optimisation follows closely a previous study intended to disclose the optimal variable set to use in MVA techniques in the WH and ZH search at LHC [84], based on signal and background processes simulated by MC at generator level. Therefore, the detector effects were not reflected in the results. The nominal variable set used in the WH BDT did not include any of the results of this study and so, these provided a guideline for the BDT optimisation work here described. The definitions of the new variables tested are presented in Table 6.22 and in the rest of this Section.

The set of input variables tested cover a wide range of different observables and enclose information about the event kinematics, shape and angular distributions. Some of the variables depend on the neutrino longitudinal momentum that can not be unambiguously determined from detector observables alone. For these cases, solutions described in Section 6.3.8 were used and

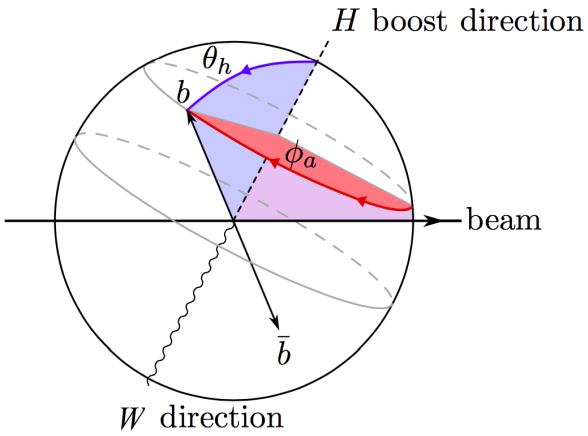


Figure 6.38: Helicity θ_h and Azilicity ϕ_a angle of the daughter $b-$jet defined for the $H \to b\bar{b}$ decay system. Adapted from [84].

tested.

Helicity and Azilicity Figure 6.38 exhibits the parametrisation of the Higgs decay products with two angles in the rest frame of the Higgs. The angles are measured with respect to the W and beam line directions as observed in this frame. The helicity angle θ_h measures the longitude, with the W direction defining the axis of the spherical coordinate system. The azilicity ϕ_a is a measurement of the azimuthal angle with origin agreed to be pointing to the beam line direction. Since the Higgs is a scalar, the distributions of these two angles are uniform in the case of signal events. The same is not expected for background events and therefore these angles can provide auxiliary discriminant information to the BDT.

Twist angle The twist angle is defined by the following:

$$\tau = \arctan\frac{\Delta\phi}{\Delta\eta} \tag{6.11}$$

where $\Delta\phi$ and $\Delta\eta$ represent the azimuthal distance and pseudorapidity difference between two final state objects. This variable explores the correlation between $\Delta\phi$ and $\Delta\eta$ for the bb system that is different for signal and background. For the signal process, jets are emitted mainly in the central region of the detector with low pseudorapidity, causing $\Delta\eta(b_1, b_2)$ to be very close to 0 and the τ_{bb} distribution to peak at $\pi/2$ as seen from Figure 6.40. Figure 6.39 shows a schematic view of the twist angle with a signal event with $\tau_{bb} = \pi/2$ and a background event where different configurations happen with the same probability.

Sphericity and Spherocity Tensors The sphericity and spherocity, aplanarity and aplanority are event shape variables build from the sphericity and spherocity tensors, respectively. The

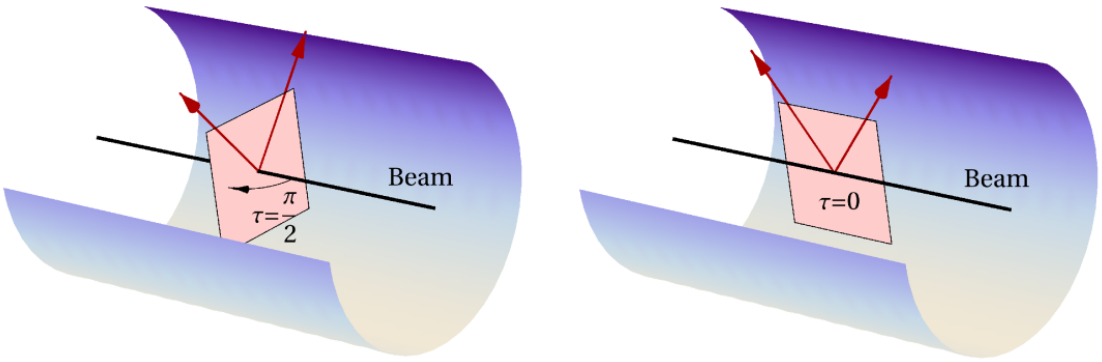


Figure 6.39: Twist angle of the two $b-$jets system for a signal and a background event. Taken from [84].

latter are defined by the momentum p of the four final state objects as follows:

$$\text{Sphericity Tensor} = \frac{1}{\sum_i |\overrightarrow{p}_i|^2} \sum_i \begin{pmatrix} p_x p_x & p_x p_y & p_x p_z \\ p_y p_x & p_y p_y & p_y p_z \\ p_z p_x & p_z p_y & p_z p_z \end{pmatrix} \tag{6.12}$$

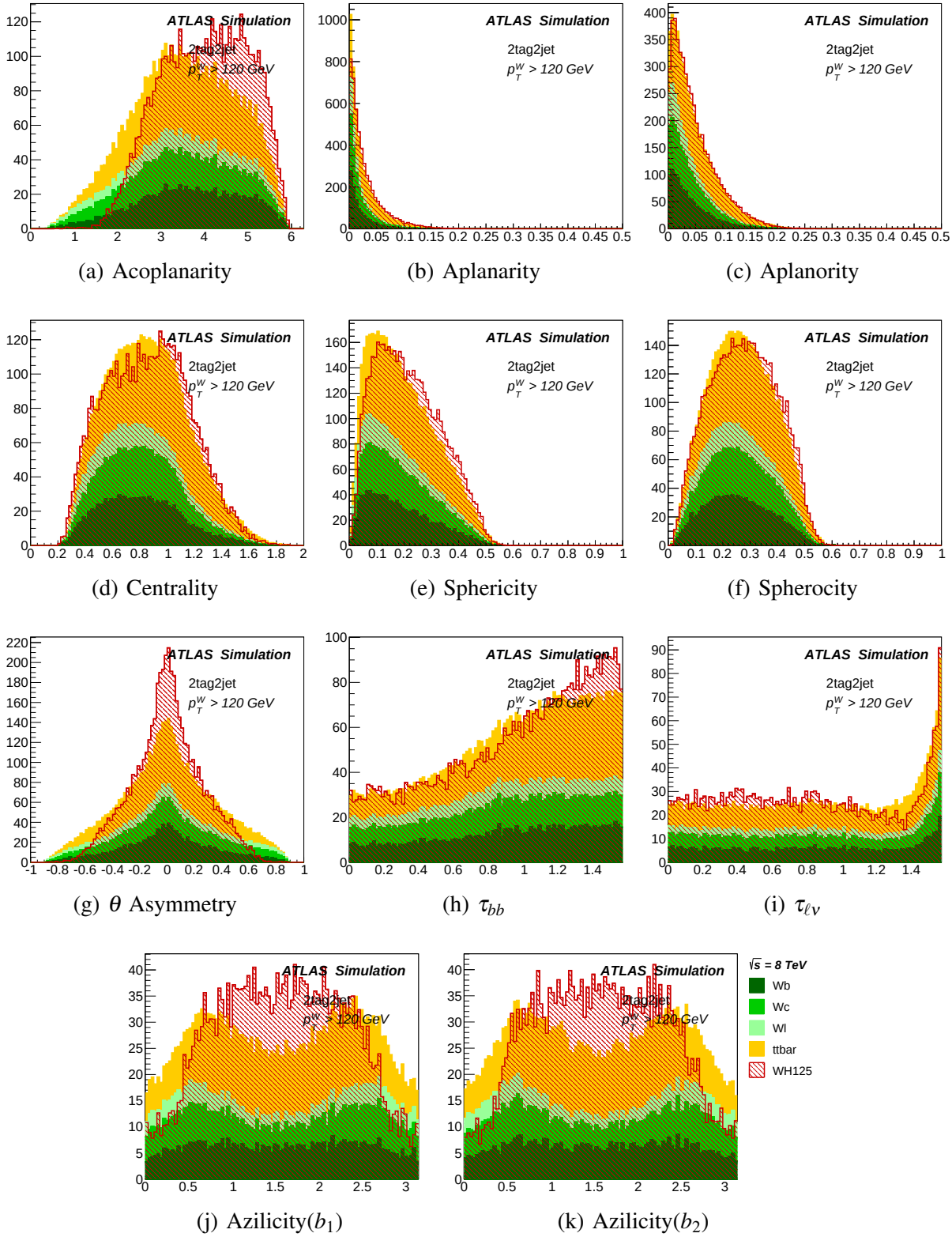
$$\text{Spherocity Tensor} = \frac{1}{\sum_i |\overrightarrow{p}_i|} \sum_i \frac{1}{|\overrightarrow{p}_i|} \begin{pmatrix} p_x p_x & p_x p_y & p_x p_z \\ p_y p_x & p_y p_y & p_y p_z \\ p_z p_x & p_z p_y & p_z p_z \end{pmatrix} \tag{6.13}$$

where the particle index i is omitted from the matrix elements for simplicity. The eigenvectors of these tensors, $\{\lambda_1, \lambda_2, \lambda_3\}$, after ordering ($\lambda_1 \geq \lambda_2 \geq \lambda_3$) and normalising ($\lambda_1 + \lambda_2 + \lambda_3 = 1$), are then used to calculate:

Sphericity and spherocity: $S = \frac{3}{2}(\lambda_2 + \lambda_3)$, where S can range from 0 to 1, from a non-spherical to a perfectly isotropic event, respectively.

Aplanarity and aplanority: $S = \frac{3}{2}(\lambda_3)$, with A ranging from 0 to 1/2, from a planar to a perfectly isotropic event, respectively.

Distributions Figures 6.40 and 6.41 show the distributions of the new variables for the signal and main backgrounds simulation, for events with 2 jets, both $b-$tagged, and $p_{\mathrm{T}}^{W} > 120$ GeV. Most of them do not seem to add much discriminating power as the signal and background shapes are always superimposed. However, any particular correlation with other BDT input variable can always benefit the analysis, and therefore none of the variables is discarded at this stage. Some of the most promising variables reported at [84] lose their power when the detector effects are considered here. Also, the cut-based pre-selection leads to background samples containing events that tightly match the signal topology. Other variables, however, such as acoplanarity, θ asymmetry, azilicity angles, the rapidity differences and radial differences still show some separation potential and are expected to impact the BDT discriminant positively.



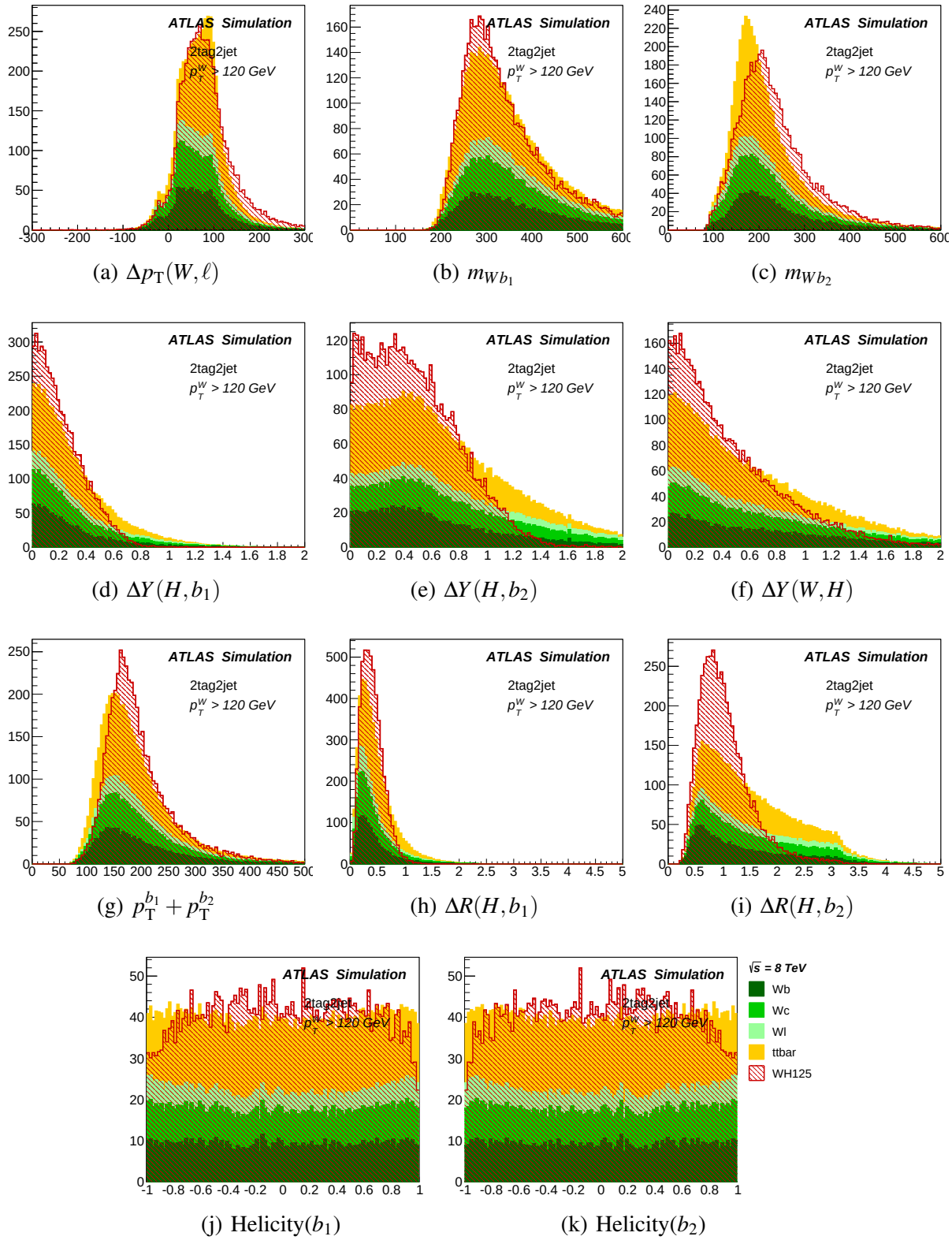
(a) Acoplanarity (b) Aplanarity (c) Aplanority

(d) Centrality (e) Sphericity (f) Spherocity

(g) θ Asymmetry (h) τ_{bb} (i) $\tau_{\ell\nu}$

(j) Azilicity(b_1) (k) Azilicity(b_2)

Figure 6.40: Distribution of the potential discriminant variables for BDT training for simulated events of signal and principal backgrounds. The samples shown correspond to the analysis category of two signal jets, both $b-$tagged, and $p_{\mathrm{T}}^{W} > 120$ GeV. The signal distribution is normalised to the total background integral.



(a) $\Delta p_{\mathrm{T}}(W,\ell)$ (b) m_{Wb_1} (c) m_{Wb_2}

(d) $\Delta Y(H,b_1)$ (e) $\Delta Y(H,b_2)$ (f) $\Delta Y(W,H)$

(g) $p_{\mathrm{T}}^{b_1}+p_{\mathrm{T}}^{b_2}$ (h) $\Delta R(H,b_1)$ (i) $\Delta R(H,b_2)$

(j) Helicity(b_1) (k) Helicity(b_2)

Figure 6.41: Distribution of the potential discriminant variables for BDT training for simulated events of signal and principal backgrounds. The samples shown correspond to the analysis category of two signal jets, both $b-$tagged, and $p_{\mathrm{T}}^{W} > 120$ GeV. The signal distribution is normalised to the total background integral.

The invariant mass of the W and leading or sub-leading jet can be useful to distinguish between signal and top background events since it can sometimes correspond to the top mass value in the case of a top quark event. The m_{Wb_2} distribution shows exactly this, with the top distribution peaking at a value similar to the m_{top}. The signal distribution of m_{Wb_2} is broader and peaks at a higher value.

Training

The BDT training scheme detailed in Section 6.5.2 is repeated for the following input variables set:

- baseline variables summarised in Table 6.20 + 1 new variable listed in Table 6.22 at a time

All the remaining BDT parameters are kept unchanged as the goal is to evaluate the impact of the new variable added on the final discriminant.

The impact of the variable in the structure of the trained BDT can be assessed via the variable importance index defined in what follows.

Variable importance: quantity proportional to the number of times a variable is used at a BDT node weighted by the signal and background separation gain resulting from that usage and by the number of events in that node. The larger this quantity, the more useful is the input discriminant to the BDT method and vice-versa.

Figure 6.42 shows the variable importance ranking for the baseline BDT training in one of the event categories. $m_{b\bar{b}}$, $p_{\mathrm{T}}^{b_1}$ and m_{T}^{W} are the variables that most contribute to the BDT structure. Figure 6.36 exhibiting a WH decision tree example already showed the efficient use that the BDT makes of $m_{b\bar{b}}$ to select the signal mass peak. The ranking varies between the odd and even event number $k-$fold used during training. This is a result of statistical fluctuations in the training samples, even more potentiated by the boost technique where the structure of each new decision tree depends on the previous one. Therefore, the variable importance index should be regarded as a qualitative hint of the variable performance, for it clearly distinguishes useful discriminants, but is unable to categorically quantify its impact.

Figure 6.43 shows the same ranking for two of the variables tested. These are two extreme examples of the results obtained. On one hand $\Delta Y(W,H)$ reaches the second place while τ_{bb} only the tenth out of 12 possibilities. The results shown so far have focused on the 2 $b-$jets and p_{T}^{W} >120 GeV analysis region since this has the largest signal significance, but Tables 6.23 and 6.24 summarise the results for all signal regions. These show that concerning the variable ranking $\Delta Y(W,H)$ and m_{Wb_1} stand out as potentially performant.

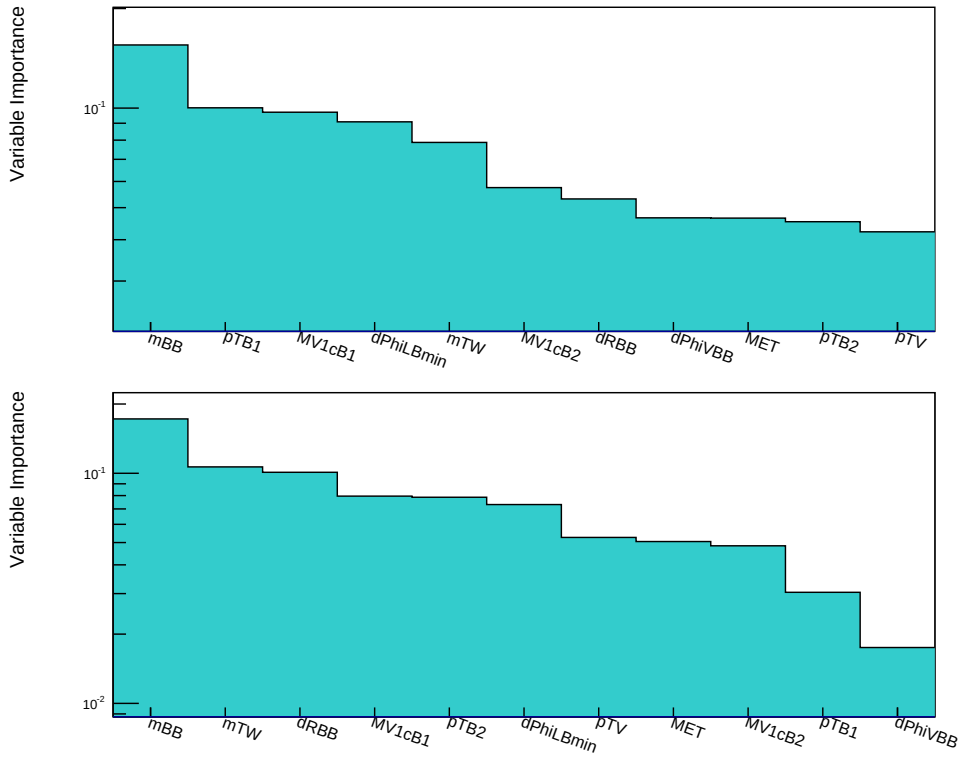


Figure 6.42: Variable importance ranking for the baseline BDT input variables summarised in Table 6.20 resulting from the training of event samples with 2 $b-$jets and $p_{\mathrm{T}}^{W} > 120$ GeV corresponding to the (top) even and (bottom) odd event number $k-$fold categories.

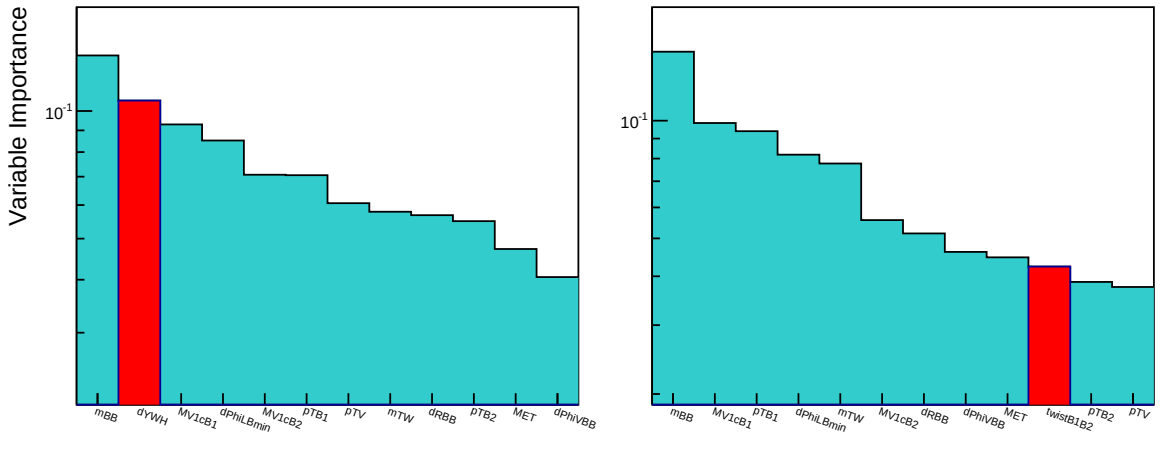


Figure 6.43: Variable importance ranking for the baseline BDT input variables summarised in Table 6.20 (left) $+\Delta Y(W,H)$ or (right) $+\tau_{bb}$ resulting from the training of event samples with 2 $b-$jets and $p_{\mathrm{T}}^{W} > 120$ GeV corresponding to the odd event number $k-$fold category.

	Rank position				Background rejection			
	$p_{\mathrm{T}}^{W} < 120$GeV		$p_{\mathrm{T}}^{W} > 120$GeV		$p_{\mathrm{T}}^{W} < 120$GeV		$p_{\mathrm{T}}^{W} > 120$GeV	
Variable	even	odd	even	odd	even	odd	even	odd
$\Delta Y(W,H)$	4	4	**3**	**2**	$\sim$ **+6%**	$\sim$ **+9%**	$\sim$ **+10%**	$\sim$ **+9%**
$\Delta Y(H,b_1)$	4	4	5	12	–	–	–	–
$\Delta Y(H,b_2)$	12	7	12	10	–	–	–	–
m_{Wb_1}	4	10	12	11	$\sim$ **+5%**	$\sim$ **+4%**	$\sim$ **+3%**	$\sim$ **+7%**
m_{Wb_2}	11	10	5	12	–	–	–	–
$p_{\mathrm{T}}^{b_1} + p_{\mathrm{T}}^{b_2}$	9	11	12	12	–	–	–	–
$\Delta p_{\mathrm{T}}(W,\ell)$	12	12	14	12	–	–	–	–
$\Delta R(H,b_1)$	11	12	3	1	–	–	–	–
$\Delta R(H,b_2)$	12	12	7	7	–	–	–	–
τ_{bb}	5	9	13	11	–	–	–	–
$\tau_{\ell\nu}$	6	12	5	12	–	–	–	–
Helicity(b_1)	11	8	8	10	–	–	–	–
Helicity(b_2)	11	8	8	10	–	–	–	–
Azilicity(b_1)	11	12	7	9	–	–	–	–
Azilizity(b_2)	8	10	11	12	–	–	–	–
θ Asymmetry	4	10	4	5	–	–	$\sim$ +2%	$\sim$ +2%
Acoplanarity	11	7	12	11	$\sim$ +2%	$\sim$ +2%	–	–
Centrality	11	10	12	12	–	$\sim$ +2%	–	–
Sphericity	12	10	8	12	–	–	–	–
Spherocity	12	10	2	10	–	$\sim$ +2%	–	–
Aplanarity	12	11	6	12	–	–	–	–
Aplanority	12	10	7	12	–	–	–	–
$\Delta Y(W,H) + m_{Wb_1}$	5/11	4/11	3/12	3/12	$\sim$ **+6%**	$\sim$ **+10%**	$\sim$ **+10%**	$\sim$ **+10%**

Table 6.23: Rank position resulting from adding each new variable to the BDT training and corresponding impact on the background rejection with respect to the nominal BDT for events with 2 jets, 2 b-tags.

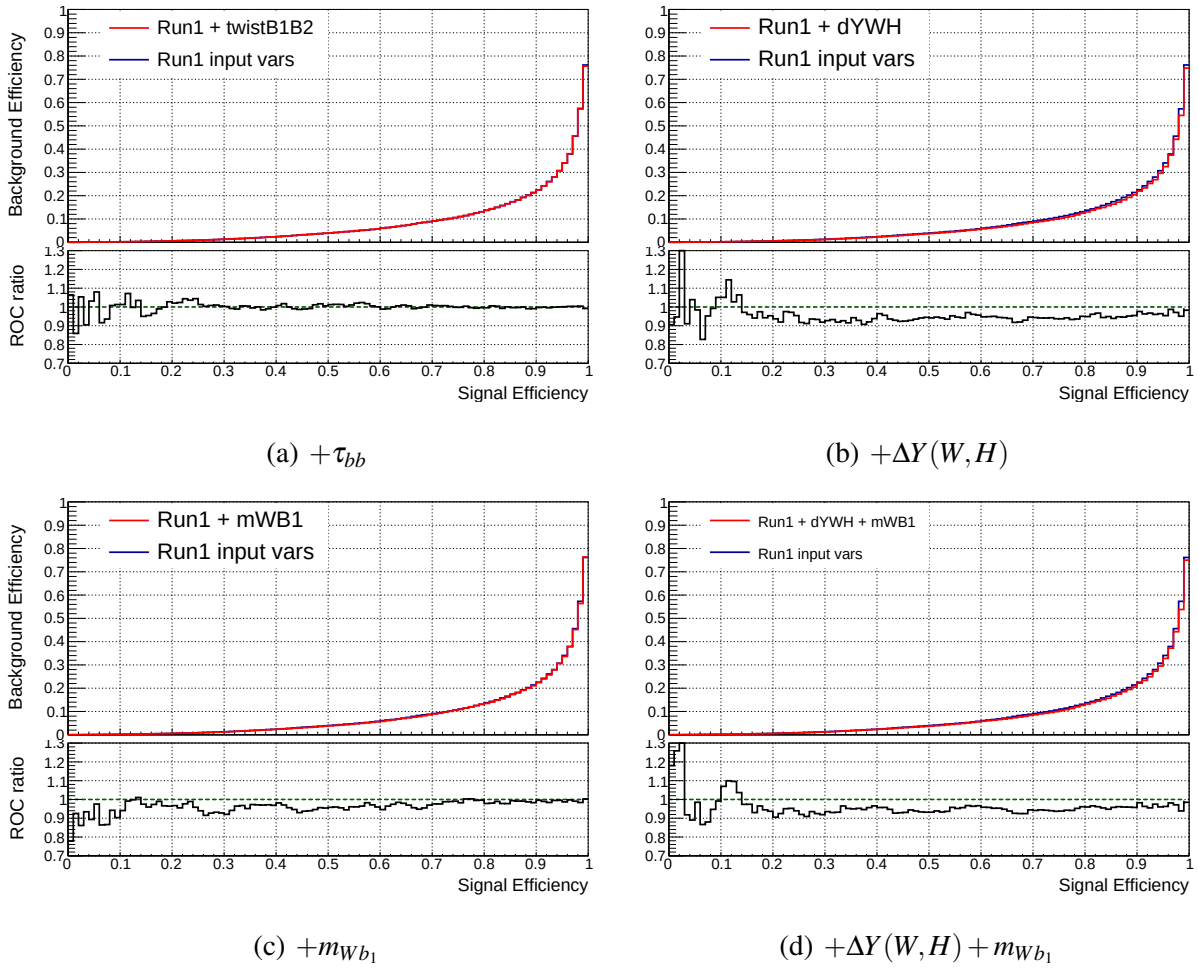
Impact on the WH BDT

As discussed, the impact of the variables under study on the final BDT discriminant can hardly be quantified by the variable importance during training. A more conclusive method is to compare the signal and background efficiencies for a given cut on the BDT output distribution for the baseline and each new BDT. For this purpose, the so-called Receiver Operating Characteristic (ROC) is used. The curve is constructed by plotting the efficiency for background as a function of the efficiency for the signal for different BDT output values. It displays the performance of the binary classifier as its discrimination threshold is varied.

The lower the ROC curve the better, for it means more background rejection capability for the same signal efficiency. These curves are shown for some training settings in Figure 6.44. It can be seen from the ratio between the new and nominal ROC curves that adding τ_{bb} to the set of BDT inputs does not impact the discriminating power of the BDT, while adding $\Delta Y(W,H)$ and m_{Wb_1} lead up to 10 and 7% higher background rejection, respectively, in the most sensitive part of the BDT output spectrum where the signal efficiency is above 40%. Given the performance of these two variables, already seen before on the variable importance ranking, both were added

	Rank position				ROC curve impact			
	$p_{\mathrm{T}}^{W} < 120$GeV		$p_{\mathrm{T}}^{W} > 120$GeV		$p_{\mathrm{T}}^{W} < 120$GeV		$p_{\mathrm{T}}^{W} > 120$GeV	
Variable	even	odd	even	odd	even	odd	even	odd
$\Delta Y(W,H)$	7	2	7	7	$\sim$ **+6%**	$\sim$ **+9%**	$\sim$ **+10%**	$\sim$ **+10%**
$\Delta Y(H,b_1)$	10	2	14	5	–	–	–	–
$\Delta Y(H,b_2)$	13	13	14	11	–	–	–	–
m_{Wb_1}	13	6	14	5	$\sim$ **+5%**	$\sim$ **+4%**	$\sim$ **+3%**	$\sim$ **+7%**
m_{Wb_2}	14	14	13	10	–	–	–	–
$p_{\mathrm{T}}^{b_1} + p_{\mathrm{T}}^{b_2}$	11	14	13	14	–	–	–	–
$\Delta p_{\mathrm{T}}(W,\ell)$	12	14	14	14	–	–	–	–
$\Delta R(H,b_1)$	14	8	14	1	–	–	–	–
$\Delta R(H,b_2)$	13	7	14	4	–	–	–	–
τ_{bb}	11	13	14	14	–	–	–	–
$\tau_{\ell\nu}$	14	4	14	4	–	–	–	–
Helicity(b_1)	4	13	7	14	–	–	–	–
Helicity(b_2)	4	13	7	14	–	–	–	–
Azilicity(b_1)	10	5	14	14	–	–	–	–
Azilizity(b_2)	9	7	14	5	–	–	–	–
θ Asymmetry	11	7	5	14	–	–	$\sim$ +2%	$\sim$ +2%
Acoplanarity	13	2	14	3	$\sim$ +2%	$\sim$ +2%	–	–
Centrality	14	3	14	14	–	$\sim$ +2%	–	–
Sphericity	14	14	14	5	–	–	–	–
Spherocity	14	13	14	8	–	$\sim$ +2%	–	–
Aplanarity	13	13	14	13	–	–	–	–
Aplanority	14	14	14	13	–	–	–	–
$\Delta Y(W,H) + m_{Wb_1}$	5/15	5/14	4/15	6/15	$\sim$ **+6%**	$\sim$ **+5%**	$\sim$ **+8%**	$\sim$ **+5%**

Table 6.24: Rank position resulting from adding each new variable to the BDT training and corresponding impact on the background rejection with respect to the nominal BDT for events with 3 jets, 2 b-tags.



(a) $+\tau_{bb}$ (b) $+\Delta Y(W,H)$

(c) $+m_{Wb_1}$ (d) $+\Delta Y(W,H)+m_{Wb_1}$

Figure 6.44: ROC curves (top panels) corresponding to the training of event samples with 2 jets and $p_{\mathrm{T}}^{W} > 120$ GeV with even event number kfold category for the baseline BDT (blue) and baseline BDT + a new variable (red). The bottom panel shows the ratio between the new and nominal ROC curves.

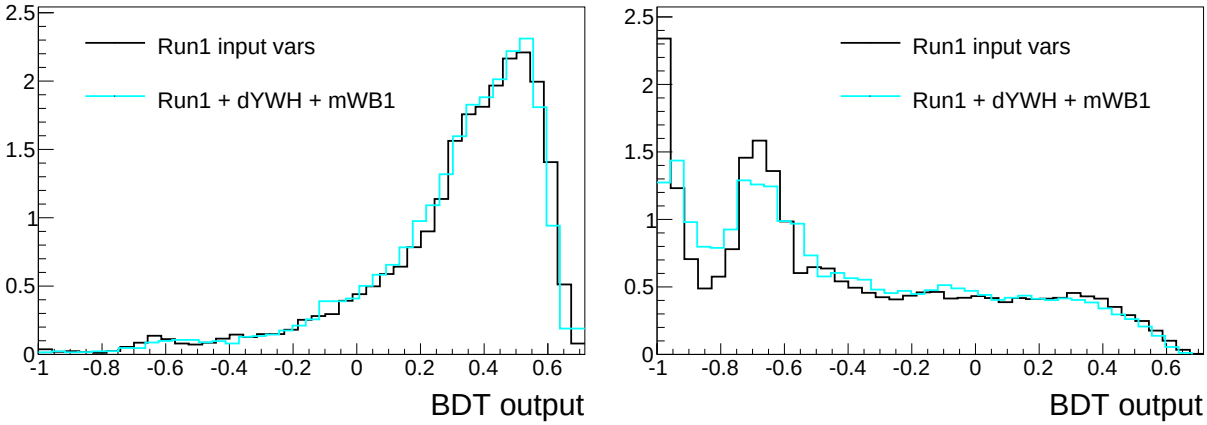


Figure 6.45: Distribution of the outputs of the baseline BDT and BDT+$\Delta Y(W,H)+m_{Wb_1}$ for (left) signal and (right) background events.

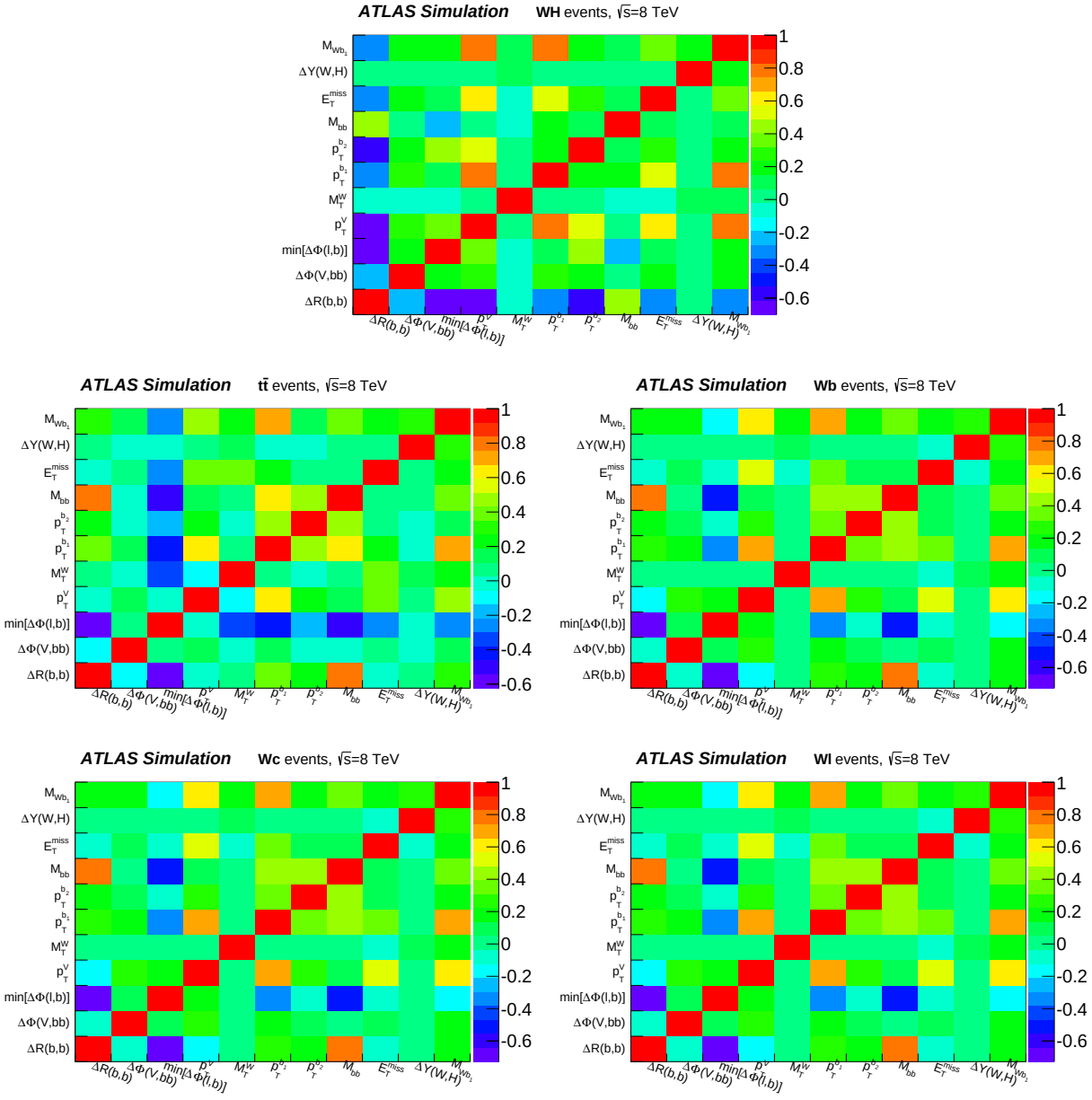


Figure 6.46: Correlation matrices of the BDT input variables for simulated WH, $t\bar{t}$ and $W+b-, c-$ and light-jets events with 2 $b-$tagged jets and $p_{\mathrm{T}}^{W} > 120$ GeV.

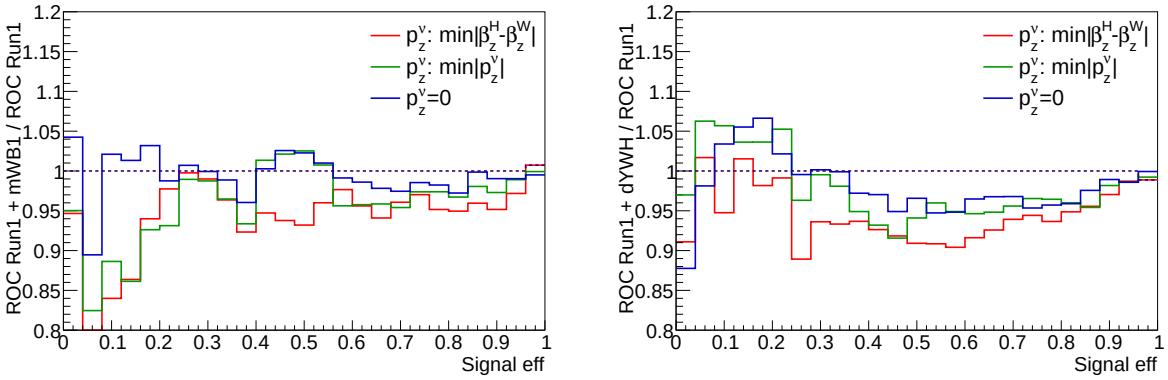


Figure 6.47: Ratio between the ROC curves of the (left) baseline BDT $+m_{Wb_1}$ (right) baseline BDT $+\Delta Y(W,H)$ and the baseline BDT for the different p_z^ν solutions, for the 2 jets and $p_T^W >$ 120 GeV category.

at the same time to the BDT training. Figure 6.44(d) shows that this slightly improves the BDT performance of the baseline $+\Delta Y(W,H)$ BDT. The results obtained for all the training conditions are summarised in Tables 6.23 and 6.24. The same conclusions hold for all the analysis categories.

Figure 6.45 shows the distributions of the BDT output weight for signal and background events with two signal $b-$jets and $p_T^W < 120$ GeV, obtained from training with the baseline set of variables and when adding both $\Delta Y(W,H)$ and m_{Wb_1} to the training. In the high signal efficiency region corresponding to the BDT weight in the interval between 0.3 and 0.6, the signal is more peaked when the new variables are added. The background is shifted towards smaller BDT output values for the new BDT. Thus, the added variables contribute both to improve the signal efficiency and reduce the amount of background in the region more sensitive to signal.

The observed impact of $\Delta Y(W,H)$ and m_{Wb_1} can be better understood from the variables correlation matrices shown in Figure 6.46. $\Delta Y(W,H)$ is not correlated with any other input variable meaning that it adds information about the event. On its turn, m_{Wb_1} is differently correlated for signal and background events and this is exactly the type of feature that the BDT explores. In general, the remaining variables were either strongly correlated with the baseline set of inputs, and for that reason their information was redundant, or did not add much separation power, leading to no impact on the BDT performance.

p_z^ν solutions The following p_z^ν solutions, referred in Section 6.3.8, were incorporated in the BDT improvement study by comparing the BDT performance when adding a new variable for each considered solution:

- $p_{\nu,z}$: min $|\beta_z^H - \beta_z^W|$
- $p_{\nu,z}$: min $|p_z^\nu|$

Additionally, setting $p_{\nu,z} = 0$ was tested for simplicity reasons. Figure 6.47 shows the ratio between the ROC curves of the BDT when a new variable was added, and the baseline BDT for

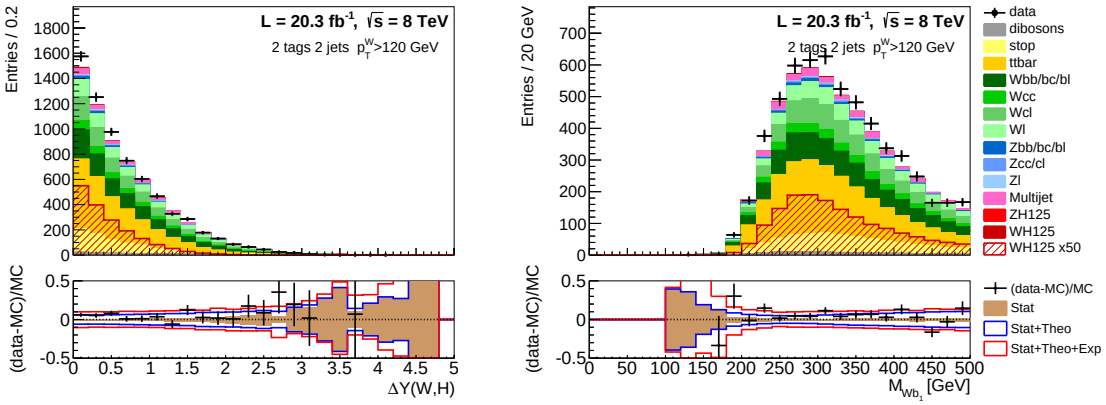
these three solutions. The two most performant variables are shown as example although all the cases were submitted to the same procedure. When this study was carried out, only the direct $b-$tagging was being used and the multijet background was not included at the BDT training. So, these plots do not match completely the ones shown in Figure 6.44, specially in what comes to the statistical effects, although the main conclusions are compatible. Nevertheless, they are useful to evaluate the BDT performance as a function of the p_z^ν solutions in a relative manner. They show that the solution that minimised the longitudinal boost difference between the reconstructed Higgs and the W bosons not only provided a more faithful description of the neutrino candidate, and therefore of the W candidate, as seen in Section 6.3.8, but also performed better on the BDT outputs. This was the solution used in all the results shown previously and will be used in what follows.

Modelling of the new input variables

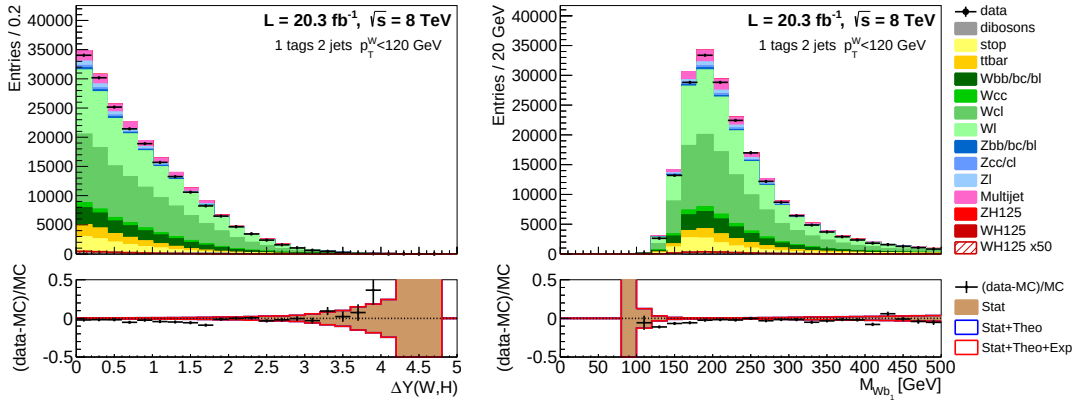
The modelling of $\Delta Y(W,H)$ and m_{Wb_1} was checked by comparing data and MC prediction for the different analysis regions. Appendix D contains the distributions of these variables for all the analysis categories. Examples are given in Figure 6.48, showing that data and prediction agree within the uncertainty band. On the ratio plot on the bottom panels no trend is observed for the present residual disagreement. The latter can still vanish during the statistical analysis of data, where the normalisation of the main backgrounds is constrained with real data. The regions enriched in the $t\bar{t}$ and W+jets backgrounds show the same level of agreement, revealing that $\Delta Y(W,H)$ and m_{Wb_1} is correctly modelled by MC for the dominant backgrounds.

Similarly, the impact of these variables on the BDT output distribution modelling is shown in Figure 6.49. Again, data and prediction converge within the uncertainties of the distributions for both the most sensitive signal region and a $t\bar{t}$ -enriched region, an expected feature given the discussion above.

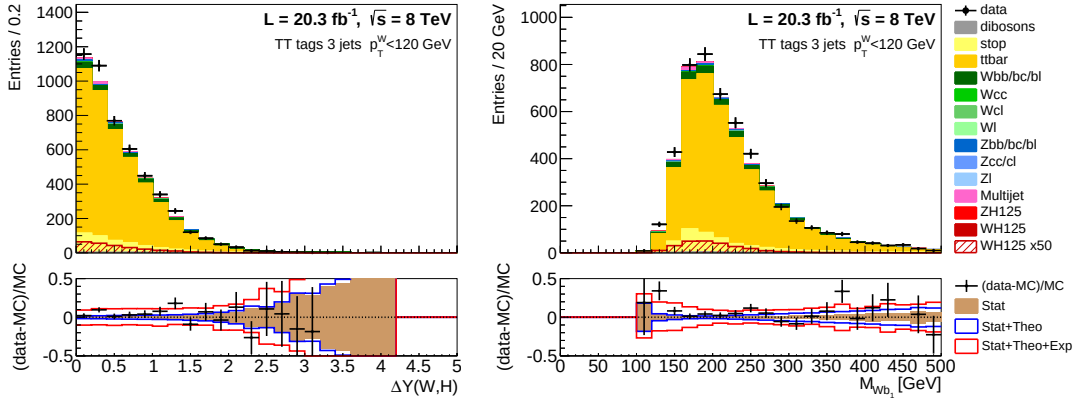
While for the largest backgrounds is possible to obtain enriched samples where simulation can be directly compared to data, the modelling checks for the signal process consist of comparing the predictions of different simulation models. Figures 6.50 show the $\Delta Y(W,H)$ and m_{Wb_1} distributions for the nominal LO generator PYTHIA8, and the alternative model POWHEG+PYTHIA8. The latter generates the hard-scatter at NLO in perturbative QCD with POWHEG but is interfaced with the PYTHIA8 hadronisation and showering models. Information about the alternative samples can be found in Appendix B, Table B.4. For the nominal sample, the signal modelling uncertainties, detailed in Section 7.1.3, are applied, as these are taken into account in the signal prediction used by this analysis. Theoretical uncertainties related for example with the $H \to b\bar{b}$ branching ratio or the WH production cross-section are excluded from this comparison since they affect both predictions equally. The discrepancies between the two models concerning the $\Delta Y(W,H)$ and m_{Wb_1} final spectra are covered by the uncertainty on the signal prediction. If this was not the case, a modelling systematic could be necessary to take into account the uncertainty on the variables shape. Moreover, their impact on the final



(a) Signal region

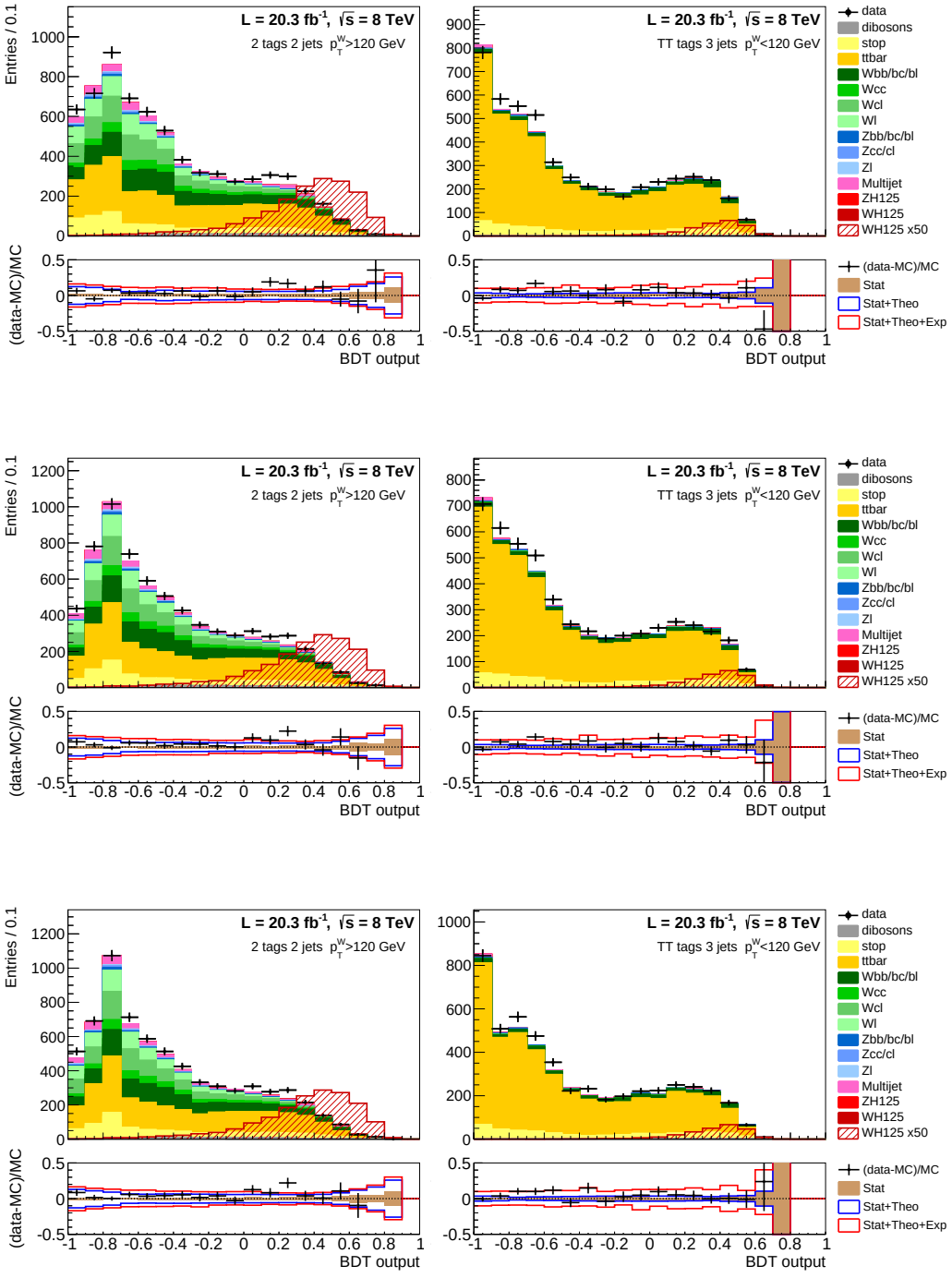


(b) W+jets-enriched



(c) $t\bar{t}$ -enriched

Figure 6.48: Distribution of the BDT training variables for data and prediction for events with two signal jets, both $b-$tagged, and $p_{\mathrm{T}}^{W} < 120$ GeV. The uncertainty bands in the bottom panel include the total statistical and systematic uncertainty as described in Chapter 7 Section 7.1.



(a) 2 jets, 2 $b-$tags, $p_{\mathrm{T}}^{W} > 120$ GeV (b) 3 jets, 2 TT $b-$tags, $p_{\mathrm{T}}^{W} < 120$ GeV

Figure 6.49: Distribution of the BDT discriminant for data and prediction in different analysis categories for (top) baseline BDT (middle) baseline $+\Delta Y(W,H)$ BDT and (bottom) baseline $+\Delta Y(W,H) + m_{Wb_1}$ BDT. The uncertainty bands in the bottom panel include the total statistical and systematic uncertainty as described in Chapter 7 Section 7.1.

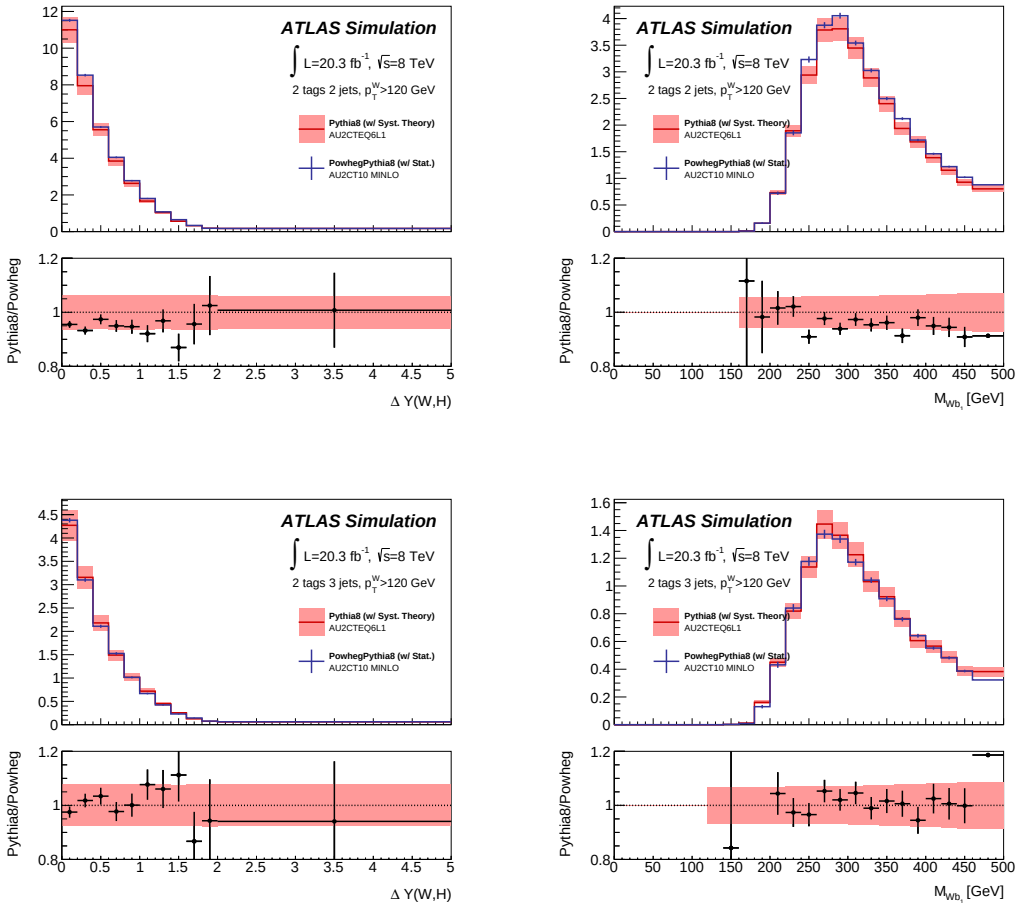
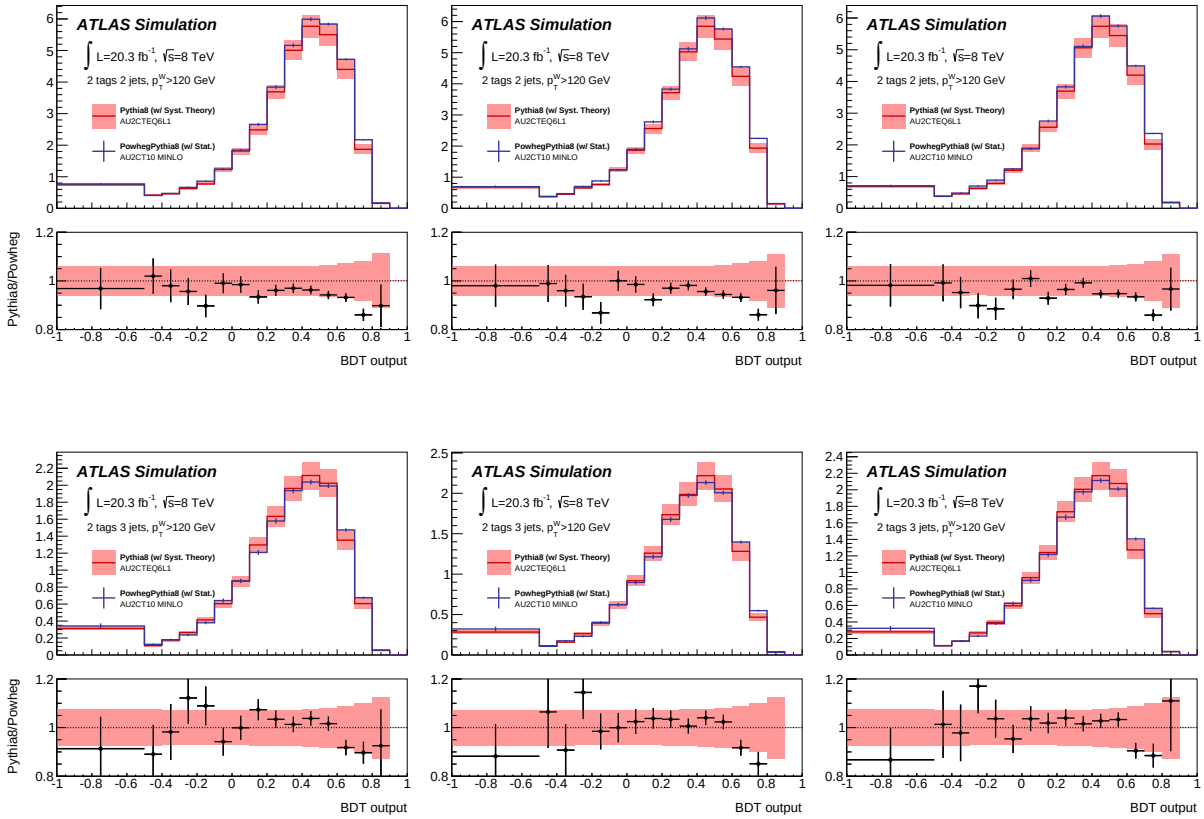


Figure 6.50: (left) $\Delta Y(W,H)$ and (right) m_{Wb_1} distributions for signal events simulated with the nominal PYTHIA8 and POWHEG+PYTHIA8 models in different analysis categories. The nominal distribution is exhibited with the signal modelling uncertainty band as detailed in Chapter 7 Section 7.1.3.

BDT discriminant does not change much between the two simulations, as can be observed from Figure 6.51, with the difference between the two predictions falling within the signal prediction uncertainty.

As the MC simulation correctly models the distribution of these two variables, they can be safely used as input discriminants to the BDT, with the benefit of increasing up to 10% the background rejection ability of the WH MVA analysis. Although the result seems promising, a final conclusion can only be reached when all the analysis uncertainties are taken into account and the statistical analysis of the data is performed, since a 10% improvement can be dissolved when all these aspects are considered. Chapter 7 will precisely give an answer to this question.



(a) baseline BDT (b) $+\Delta Y(W,H)$ (c) $+\Delta Y(W,H)+m_{Wb_1}$

Figure 6.51: Distribution of the BDT discriminant for signal events simulated with the nominal PYTHIA8 and POWHEG+PYTHIA8 models in different analysis categories for (left) baseline BDT (middle) baseline $+\Delta Y(W,H)$ BDT and (right) baseline $+\Delta Y(W,H)+m_{Wb_1}$ BDT. The nominal distribution is exhibited with the signal modelling uncertainty band as detailed in Chapter 7 Section 7.1.3.

WH BDT for the LHC pp collisions at 13 TeV

The BDT optimisation study described before was repeated using signal and background events resulting from simulated pp collisions at $\sqrt{s}$ =13 TeV.

The dataset used for this study corresponds to simulated samples of the $WH \to \ell\nu b\bar{b}$ signal and background processes corresponding to the LHC Run II conditions, i.e. pp collisions at $\sqrt{s} = 13$ TeV and integrated luminosity of 3.2 fb^{-1}. The details about the samples and MC generators and the cross-sections of each process at $\sqrt{s} = 13$ TeV pp collisions are listed in Tables 6.25 and 6.26 [85]. The multijet background is not taken into account in this study but, as it will be discussed ahead, this process does not have a significant impact on the analysis. By increasing the centre-of-mass energy of the collisions, the signal process cross-section is expected to increase by a factor of approximately 2, while $t\bar{t}$ increases by 3.3 and W+jets by 1.7.

Signal Process	Generator	σ [pb]	BR	N_{events}
$q'\bar{q} \to WH \to \ell\bar{\nu}_\ell b\bar{b}$	PYTHIA8	1.38	0.1886	500000

Table 6.25: Generators used in the $WH \to \ell\nu b\bar{b}$ signal event simulation at $\sqrt{s} = 13$ TeV pp collisions.

Background Process	Generator	$\sigma \times BR$ [pb]	N_{events}
Top quark			
$t\bar{t}$	POWHEG+PYTHIA6	831.76	19M
$t-$channel	POWHEG+PYTHIA6	69.51	1M
$s-$channel	POWHEG+PYTHIA6	3.31	1M
$Wt-$channel	POWHEG+PYTHIA6	68.00	1M
Vector Boson + jets			
$W \to \ell\bar{\nu}$	SHERPA 2.2	20080	68M
$Z/\gamma* \to \ell\bar{\ell}$	SHERPA 2.2	2107	11M
$Z/ \to \nu\bar{\nu}$	SHERPA 2.2	1914	42M
Diboson			
WW	SHERPA 2.1	49.74	4M
WZ	SHERPA 2.1	21.69	3.5M
ZZ	SHERPA 2.1	6.99	1M

Table 6.26: Generators used to simulate the background processes at $\sqrt{s} = 13$ TeV pp collisions.

Event Selection The signal and background samples used to train the BDT were obtained through an event selection procedure very similar to the one described in Sections 6.3 and 6.4, and provided by the VH analysis group. The event selection was adjusted to guarantee the best signal-to-background ratio at the new 13 TeV regime. These are listed in Table 6.27. One of the main differences with respect to the 8 TeV selection is the p_{T}^{W} region definition, with the threshold moving from 120 to 150 GeV. By increasing the pp collisions energy, the signal W and Higgs bosons are produced with larger momenta than at 8 TeV collisions and therefore the p_{T}^{W} limit is shifted upwards. In addition, only the $p_{\mathrm{T}}^{W} > 150$ GeV interval is now considered and for this reason, the multijet background contribution in the analysis loses relevance.

Another important distinction is the usage of a $E_{\mathrm{T}}^{\mathrm{miss}}$ trigger in the muon sub-channel, with a 70 GeV threshold, that the WH analysis described in this Chapter did not employ. Muons do not enter the $E_{\mathrm{T}}^{\mathrm{miss}}$ calculation at trigger level, so in this case $E_{\mathrm{T}}^{\mathrm{miss}}$ is equivalent to p_{T}^{W} for signal events. The advantage over the single muon trigger is the higher efficiency since the muon trigger chambers have limited coverage in certain η regions. The electron sub-channel uses unprescaled single electron triggers, with E_{T} thresholds of 24, 60 and 120 GeV. The 24 GeV threshold trigger includes isolation conditions.

The $E_{\mathrm{T}}^{\mathrm{miss}} > 30$ GeV cut is mostly intended to reject the multijet background that has fake $E_{\mathrm{T}}^{\mathrm{miss}}$, as Z+jets events are almost totally removed with the loose lepton veto, as seen in Section 6.4 for the Run I analysis. Therefore, this cut is only applied to the electron channel where the multijet background has larger contribution.

$WH \to \ell\nu b\bar{b}$ MVA Selection
Lowest unprescaled single electron trigger (e sub-channel)
$E_{\mathrm{T}}^{\mathrm{miss}}$ trigger (μ sub-channel)
1 Signal lepton
Trigger matching
Loose lepton veto
2 or 3 Signal jets
Forward jet veto
$E_{\mathrm{T}}^{\mathrm{miss}}$ >30 GeV (e sub-channel)
Exactly 2 b-tagged jets
leading jet p_{T} >45 GeV
p_{T}^{W} >150 GeV

Table 6.27: Event selection criteria used in the $WH \to \ell\nu b\bar{b}$ analysis at Run II [85].

Optimisation of the $WH \to \ell\nu b\bar{b}$ BDT The WH BDT was trained for the different analysis categories:

- jet multiplicity: 2 or 3 jets;
- one p_{T}^{W} intervals: above 150 GeV;
- two $k-$folds: samples with even or odd event number.

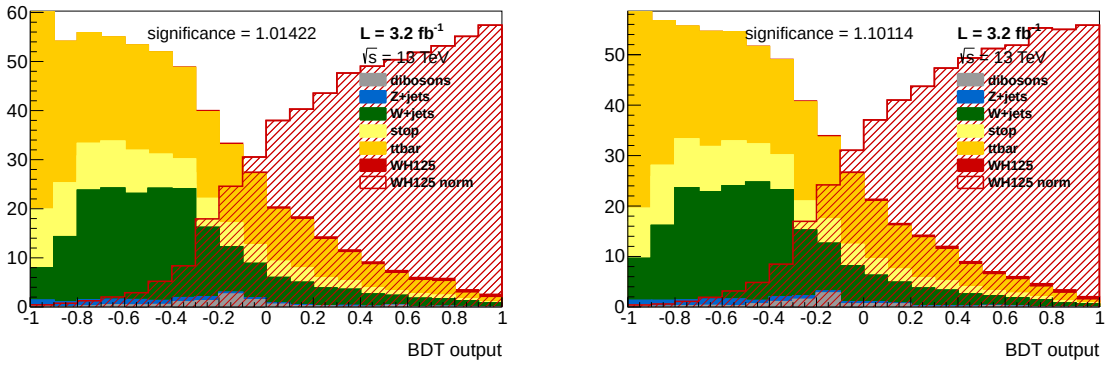
The BDT parametrisation was kept unchanged with respect to the Run I BDT configuration detailed in Table 6.19. The same applies to the set of input variables, listed in Table 6.20. The performance of the BDT was tested by comparing the performance of the BDT discriminant obtained from training the method with the nominal set of variables and with the addition of a new variable as before.

The results are shown in Figures 6.52 and 6.53 respectively for 2 and 3 jets events and for the different training cases. The BDT output distributions are shown for signal and background simulation, after the transformation described in Section 7.2.1. Table 6.28 shows the improvement in the associated cumulative significance, defined as:

$$\text{significance} = \sqrt{\sum_i \frac{s_i^2}{b_i}} \tag{6.14}$$

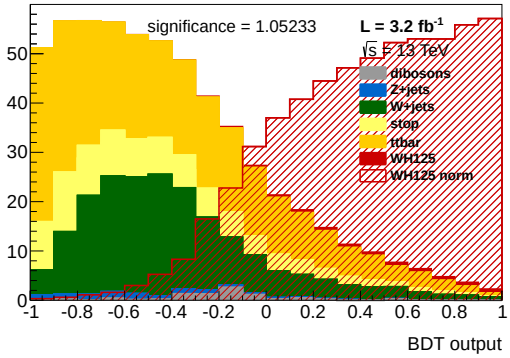
where i runs over the BDT histogram bins and s_i and b_i are the expected yield in the ith bin of the signal and background samples. Since other MVA improvement studies were ongoing on the VH analysis analysis group, the cumulative significance was adopted as metric, instead of the one presented in Section 6.5.3, to allow a more straightforward comparison.

The BDT output separation power benefits from the addition of $\Delta Y(W,H)$ and m_{Wb_1} to the input variable set. For events with two jets, an 8.6% gain is expected from the inclusion of $\Delta Y(W,H)$ alone, while for three jets the gain is 3.4%. m_{Wb_1} has smaller impact in the BDT and its addition on top of $\Delta Y(W,H)$ does not indicate any further improvement. For two jets events, the significance even drops when the second variable is considered. This should be

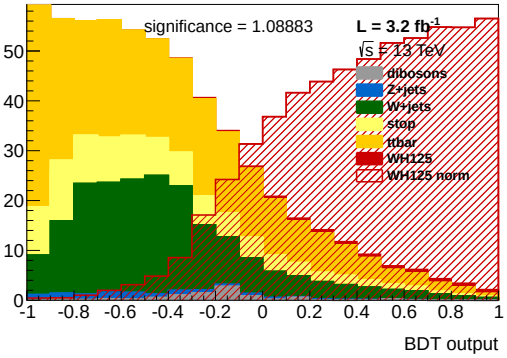


(a) baseline BDT

(b) $+\Delta Y(W,H)$



(c) $+m_{Wb_1}$

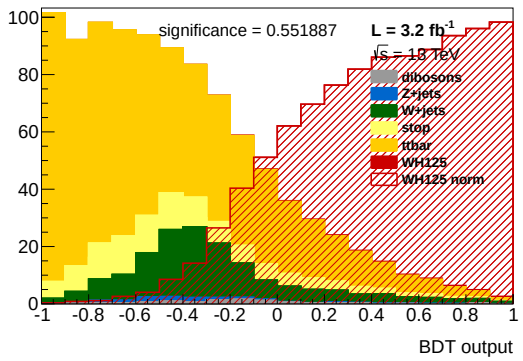


(d) $+\Delta Y(W,H)+m_{Wb_1}$

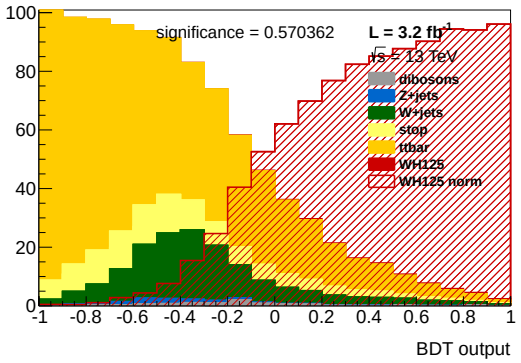
Figure 6.52: Distribution of the transformed BDT discriminant for simulated signal and background events with 2 jets, both $b-$tagged, and $p_{\mathrm{T}}^{W} > 150$ GeV for (a) baseline BDT (b) $+\Delta Y(W,H)$, (c) $+m_{Wb_1}$ and (d) $+\Delta Y(W,H)+m_{Wb_1}$. The cumulative significance is displayed in each plot.

due to statistical fluctuations in the training samples since the BDT method is in principle very robust with respect to the addition of less discriminant variables by simply not making use of them.

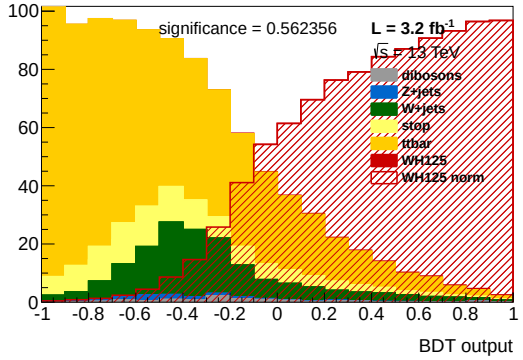
In summary, as for the Run I WH BDT, both variables do change the BDT response, increasing the signal/background separation. $\Delta Y(W,H)$ was included in the WH MVA analysis of the Run II data published by ATLAS [85]. m_{top}, a variable similar to m_{Wb_1}, but designed to identify the $b-jet$ from the top decay pairing the reconstructed W, under the hypothesis that the event is $t\bar{t}$, was added as well. Together, these variables increased the expected significance of the WH search in 7 %.



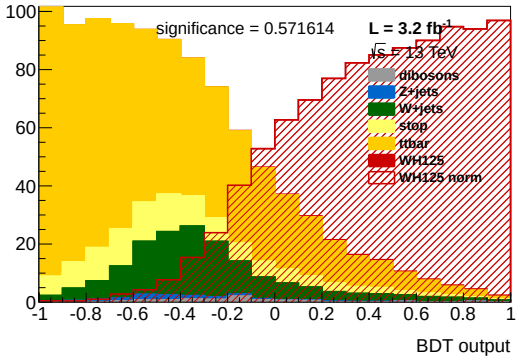
(a) baseline BDT



(b) $+\Delta Y(W,H)$



(c) $+m_{Wb_1}$



(d) $+\Delta Y(W,H)+m_{Wb_1}$

Figure 6.53: Distribution of the transformed BDT discriminant for simulated signal and background events with 3 jets, 2 $b-$tagged, and $p_{\mathrm{T}}^{W} > 150$ GeV for (a) baseline BDT (b) $+\Delta Y(W,H)$, (c) $+m_{Wb_1}$ and (d) $+\Delta Y(W,H)+m_{Wb_1}$. The cumulative significance is displayed in each plot.

Cumulative significance	2 jet	3 jet
baseline BDT	1.014	0.55
base BDT+$\Delta Y(W,H)$	+8.6%	+3.4%
base BDT+M_{Wb1}	+3.7%	+2.0%
base BDT+$\Delta Y(W,H)$+M_{Wb1}	+7.3%	+3.6%

Table 6.28: Cumulatve significance obtained with the baseline BDT output spectrum and relative difference obtained by adding the new variables to the baseline set.

Chapter 7

Uncertainties, Statistical Analysis and Results of the $WH \to \ell\nu b\bar{b}$ Search

This Chapter describes the uncertainties affecting the $WH \to \ell\nu b\bar{b}$ search, the statistical analysis of data and the results obtained. Besides the statistical error associated with the number of events contained in the data and simulated samples used, the uncertainties of the analysis can have experimental or theoretical nature. These are described in Section 7.1. The statistical analysis is detailed in Section 7.2. A maximum likelihood binned fit is used to measure the WH signal and simultaneously constrain the normalisation of the main backgrounds with observed data. Finally, the results are presented in Section 7.3.

7.1 Systematic Uncertainties

In the case of the WH analysis, the systematic uncertainties can have experimental or theoretical nature. The former are intrinsically related to the detector measurements and experimental set up, and the latter are associated with the prediction of the physical processes. Strictly speaking, even theoretical uncertainties may have an experimental basis as the physics models involved in simulation are fed by experimental results.

The impact of the systematic uncertainties in the analysis and on its results needs then to be evaluated.

7.1.1 Experimental Uncertainties

The experimental uncertainties that affect the $WH \to \ell\nu b\bar{b}$ search are essentially related to the energy scale and resolution of the final state objects used, their identification and reconstruction efficiencies and the integrated luminosity measurement. These uncertainties were generically quantified by the ATLAS physics performance groups but their impact on each specific analysis must be determined. In order to do so, the WH object and event selection chain is repeated varying each of the considered parameters from the nominal value to the

Name	Variation	Description
		Electrons
SysElecE	Up/Do	Energy scale
SysElecEResol	Up/Do	Energy resolution
SysElecEffic	Up/Do	Recontruction, ID and trigger efficiency
		Muons
SysMuonEResolID	Up/Do	ID momentum resolution
SysMuonEResolMS	Up/Do	MS momentum resolution
SysMuonEffic	Up/Do	Recontruction, ID and trigger efficiency
		$E_{\mathrm{T}}^{\mathrm{miss}}$
SysMETScaleSoftTerms	Up/Do	Soft terms scale
SysMETResoSoftTerms	Up/Do	Soft terms resolution
		Jets
SysJetNP1-6	Up/Do	Energy scale NP 1 to 6
SysJetNonClos	Up/Do	Energy scale non closure with respect to full simulation scheme
SysJetEtaModel	Up/Do	Energy scale for η intercalibration from η model
SysJetEtaStat	Up/Do	Energy scale for η intercalibration statistical uncertainty
SysJetNPV	Up/Do	Energy scale from pile-up, NPV-dependence
SysJetMu	Up/Do	Energy scale from pile-up, $< \mu >$ −dependence
SysJetPilePt	Up/Do	Energy scale from pile-up, p_{T}−dependence
SysJetPileRho	Up/Do	Energy scale from pile-up, energy density ρ−dependence
SysJetFlavComp	Up/Do	Energy scale from jet gluon and light-quark composition
SysJetFlavResp	Up/Do	Energy scale from different response to gluons and light-quarks
SysJetFlavB	Up/Do	Energy scale associated with b−jets
SysJetBE	Up/Do	b−jets energy scale from b−hadron decay
SysJetEResol	symmetric	Energy resolution
SysBJetReso	symmetric	b−jet energy resolution
SysJVF	Up/Do	JVF cut efficiency
		b−Tagging
SysBTagL0-9Effic	Up/Do	Light jets efficiency NP 0 to 9
SysBTagC0-14Effic	Up/Do	c−jets efficiency NP 0 to 14
SysBTagB0-9Effic	Up/Do	b−jets efficiency NP 0 to 9
SysTruthTagDR	Up/Do	Truth tagging $\Delta R(b_1, b_2)$ correction
		Pile-Up
SysMuScale	Up/Do	Pile-up scale

Table 7.1: List and brief description of the experimental systematic uncertainties affecting the $WH \to \ell\nu b\bar{b}$ analysis.

upper/lower limit defined by the uncertainty. This is done for each parameter at a time, for the MC prediction.

The experimental uncertainties that have an impact on this analysis are listed in Table 7.1, separated by the final state objects they are associated with, and will be discussed in detail in what follows.

Electrons

The systematic uncertainties associated with electrons correspond to the scale and resolution of the energy measurement and to the reconstruction, identification and trigger efficiency. The latter components of the efficiency are grouped in a single efficiency uncertainty parameter `SysElecEffic` that does not exceed 2% for signal electrons [34]. The energy scale and resolution uncertainties are evaluated separately and correspond to the uncertainty bands

of the scale and resolution correction factors presented at Chapter 4 in Figures 4.6 and 4.8, respectively, and do not exceed 0.4% and 5% of the electron energy [36].

Muons

For muons, momentum resolution uncertainties are considered. These are split into the MS and ID track measurement components and are of the order of 0.6% and 0.1%, respectively [38]. Muon momentum scale uncertainties are very small, $\mathcal{O}(0.01\%)$, and therefore neglected for not having a relevant impact on the analysis. As for the electrons, uncertainties associated with reconstruction and trigger efficiency are combined into a single uncertainty parameter of the order of 1%. Each of the parameters corresponds to an upwards or downwards shift of their nominal value.

Jets

As calorimeter jets are more complex objects than electrons or muons and the jet calibration chain, described in Section 4.4, involves many steps, a long list of uncertainties is associated with the jet energy scale. Just the EM to the EM+JES calibration step has 47 uncertainty components [41], differently correlated with the jet p_{T}. The uncertainties associated with the JES considered in this analysis, are related with the different steps of the jet calibration chain:

Pile-up The pile-up correction uncertainty is due to the correction dependence on the number of primary vertices (NPV), on the average number of interactions per bunch crossing $< \mu >$, on the jets p_{T} and on the calorimeter energy density. These are represented as `SysJetNPV` `SysJetMu`, `SysJetPt` and `SysJetPileRho`, respectively.

JES calibration The 47 JES uncertainty components are reduced to a set of 6 nuisance parameters (NP) that conserve the total JES uncertainty while keeping its stronger correlations with the jet p_{T}. This was reached by first diagonalising the total correlation matrix of the JES correction factors including the total uncertainty. The smaller set of NPs representative of the JES uncertainty components are the eigenvalues of the diagonalised matrix. Then, the set of NPs is further reduced by keeping separated the 5 dominant NPs, `SysJetNP1` to `SysJetNP5`, and grouping together the smaller and remaining ones representing the residual uncertainty: `SysJetNP6_rest` [41].

In-situ correction The in-situ corrections to the jet response ($p_{\mathrm{T}}^{\mathrm{reco}}/p_{\mathrm{T}}^{\mathrm{truth}}$) applied to real data are function of the jets η. This gives rise to uncertainties associated with the jets η modelling, `SysJetEtaModel`, and with the statistics used to determine this calibration `SysJetEtaStat`.

Additionally, the following uncertainties are considered [41]:

SysJetNonClos The JES was derived using the ATLAS full simulation scheme. A systematic uncertainty associated with the energy scale of jets is necessary to account for the calibration non-closure when fast simulation is used instead.

Jet flavour The jet response is different for quark or gluon-initiated jets and for different quark flavours. This is taken into account as a set of systematic uncertainties associated with the jet flavour:
SysJetFlavComp Energy scale from jet gluon and light-quark composition
SysJetFlavResp Energy scale from different response to gluons and light-quarks
SysJetFlavB Energy scale associated with $b-$jets
The usage of these systematics depend on the jet truth flavour, defined by the flavour of the closest hadron to the jet axis, with light jets being assigned the SysJetFlavComp and SysJetFlavResp uncertainties and truth $b-$jets the SysJetFlavB.

Missing Transverse Energy

The $E_{\mathrm{T}}^{\mathrm{miss}}$ is determined from the $p_{\mathrm{T}}-$sum of the other objects on an event basis. Therefore the $E_{\mathrm{T}}^{\mathrm{miss}}$ uncertainty comes from the electrons, muons and jets energy measurement uncertainties. The effects of these uncertainties are propagated into the $E_{\mathrm{T}}^{\mathrm{miss}}$ determination at the time they are individually considered. However, the impact of the uncertainty of the soft terms measurement on the $E_{\mathrm{T}}^{\mathrm{miss}}$ estimate must be considered separately, resulting in two systematic uncertainties, SysMETScaleSoftTerms and SysMETResoSoftTerms. These account for the soft term scale and resolution uncertainty, respectively. For each of the systematics, an up/down limit shift of the nominal value is considered.

$b-$Tagging

The efficiency of the $b-$tagging algorithm is different for data and MC and consequently $b-$tagging scaling factors are attributed to MC events on an event-by-event basis to restore data and MC agreement. This was detailed in Section 4.5. The calibration uncertainties have therefore impact on the signal and background MC prediction. Since the jet flavour tagging algorithms are complex, the calibration factors are affected by many uncertainties of different sources. For that reason, they are grouped in a set of NPs able to describe well the total uncertainty and the correlations between the different uncertainty sources, in a similar way that is used to parametrise the JES uncertainties described above. The set of NPs are further separated to decorrelate the uncertainty on the $b-$tagging efficiency (10 NPs SysBTagB0-9Effic), $c-$jet rejection (15 NPs SysBTagC0-14Effic), and light jet rejection (10 NPs SysBTagL0-9Effic). The usage of each group depends on the jet truth flavour, defined by the flavour of the closest hadron to the jet axis. The correspondent up/down variation from the nominal value is taken into account for each uncertainty component.

Pile-up and Luminosity

The pile-up reweighting procedure was used to fit the profile of the average number of interactions per bunch crossing $< \mu >$ of MC to real data, as described in Section 6.2.2. The uncertainty of this reweighting, `SysMuScale`, also influences the WH analysis and is therefore treated accordingly: the $< \mu >$ scale is shifted up/down (± 4 %) of its nominal value yielding to respective shifts of the nominal MC prediction.

An uncertainty of ± 2.8 % is applied to the signal and background estimated yields coming from the integrated luminosity measurement.

7.1.2 Validation of the Analysis Tools

The implementation of the systematic uncertainties in the analysis code was validated in a similar way as the object and event selection. The number of events selected by the analysis tools of the different groups participating in the WH analysis was compared for each systematic variation of the uncertainty parameters. A pre-defined sample containing 300 000 signal events was used.

The impact of each systematic uncertainty in the analysis has two distinguished possibilities: it either changes the set of events selected or just the statistical weight of the nominal sample. Systematic uncertainties affecting the objects definition, as energy scale or resolution uncertainties, belong to the first type, as the increase/decrease of the jet p_{T}, for instance, can make the event to be rejected or selected. Contrarily, uncertainties on the statistical weight of an event, as the uncertainties related to the $b-$tagging or electron or muon efficiency, belong to the second type.

Table E.1 in Appendix E shows the outcome of the validation of the analysis code with respect to the evaluation of systematic uncertainties. For the first type of uncertainties, the systematic variation outcome must be compared with the nominal number of selected events, no weights considered, in order to roughly evaluate the impact on the analysis. For the second type, the weighted number of events must be compared: the number of events with the weight shift and with the corresponding nominal weight alone. As shown in the Table E.1, the numbers obtained using the LIP analysis code differ in average 0.03% from the results of further groups. The most important difference, of the order of 0.5%, is associated with the electron efficiency systematics, for which a discrepancy is observed in the nominal weighted events of the same size.

7.1.3 Theoretical and Modelling Uncertainties

The prediction of signal and background involves many uncertainties. On one hand, the theoretical calculation of inclusive cross-sections and branching ratios of the different processes depends on the precise knowledge of the theory parameters, and on the order of the perturbation theory used to extract them. This reflects straightforward on the predicted yields through a

Name	Value	Description
TheoryBRbb	3.3%	Branching ratio
TheoryQCDScale	1%	QCD Scale uncertainty
TheoryAcc_J2	3.0%	Inclusive acceptance from QCD scale
TheoryAcc_J3	(-4.2) 4.2%	QCD scale acceptance relative to 2(3) jets
TheoryVPtQCD	p_{T}^{V}−dependent	Shape from QCD scale
TheoryPDF	2.4%	PDF uncertainty
TheoryAccPDF	3.5 (2.8)%	Acceptance from PDF uncertainty for 2(3) jets
TheoryVHPt	p_{T}^{V}−dependent	Shape from EW NLO correction
TheoryAcc_PS	7 to 13%	Acceptance from PS, UE and hadronization models

Table 7.2: List and brief description of the systematic uncertainties affecting the modelling of the signal process by MC.

scale uncertainty. On another hand, the specific models used to generate the processes, their differential cross-sections and the underlying event, the PDF descriptions of the colliding protons, the hadronisation and parton showering effects, have each their own uncertainties and rely on assumptions that can be made differently in an equally valid way. This leads to uncertainties on the background and signal acceptance and on the shape of observables, and must be determined in the analysis-specific context, as these are dependent on the topology that is searched for.

$WH \to \ell\nu b\bar{b}$ Signal

Uncertainties on the $pp \to WH \to \ell\nu b\bar{b}$ prediction are summarised in Table 7.2 [50]. These are decomposed in their different origins:

$H \to bb$ Branching Fraction The uncertainty on the predicted BR of the SM Higgs decay to bottom quarks is 3.3% for $m_H = 125$ GeV [15]. To this corresponds an uncertainty on the signal normalisation of the same order, TheoryBRbb.

QCD scale The uncertainty on the renormalisation and factorisation scales of the QCD interaction, μ_R and μ_F, affect the WH inclusive cross-section in $\pm 1\%$ [15] and are considered in the analysis as a signal normalisation uncertainty, TheoryQCDScale, of the same magnitude. The impact of the μ_R and μ_F uncertainties on the signal acceptance and observables shape was evaluated by comparing at MC truth-level the QCD NLO POWHEG samples generated with the nominal scales ($\mu_R = \mu_F = 1$) with other samples generated by shifting the μ_R and μ_F scales. Changing the QCD scales affects both the signal acceptance and the shape of the truth p_{T}^{W}. To account for this, two systematic uncertainties are assigned to the signal prediction, respectively TheoryAcc_J2/3 and TheoryVPtQCD. The first one is composed of an inclusive 2+3 jets bins acceptance uncertainty, TheoryAcc_J2, and of TheoryAcc_J3, describing the 3 jets component uncertainty relative to the 2 jets sample. The latter is applied to 3 and 2 jets events with opposite sign to explicitly anti-correlate their fluctuations, preserving the inclusive one. TheoryVPtQCD

envelopes the difference in shape observed in the p_{T}^{W} spectrum due to the QCD scale variation, and ranges from 1 to 3%.

PDF uncertainty The impact of the PDF uncertainties on the $pp \to WH$ total cross-section is $\pm 2.4\%$ [15] resulting in the `TheoryPDF` normalisation uncertainty. As for the QCD scale, the impact of this uncertainty must also be evaluated in terms of the signal acceptance or observables shape change. In order to do so, samples generated with POWHEG with nominal PDF sets were compared to alternative PDFs after full event selection performed at MC truth-level. This did not result in a significant shape variation for any relevant observable, and therefore a shape uncertainty was not justified. The differences in the obtained signal yields were found to be 3.5 (2.8)% for the 2(3) jets analysis bins and are taken as an acceptance uncertainty, `TheoryAccPDF`, correlated accros the two bins of jet multiplicity.

Parton shower, hadronisation and underlying event The effect of the showering, hadronisation model and underlying event uncertainties were examined by considering the nominal models contained in the PYTHIA8 generator and the POWHEG+HERWIG ones. The signal event yields obtained with either model were compared. An acceptance uncertainty on the signal prediction, `TheoryAcc_PS`, was necessary to take into account the limitations on the knowledge of these processes. It varies with the bins of p_{T}^{W} and jet multiplicity and ranges from 7 to 13%.

NLO EW correction uncertainty: The total cross-section of the WH production was corrected at NLO in the electroweak theory but the event simulation used an EW LO generator. Since the NLO differential cross-section has a strong dependence on p_{T}^{V}, a $p_{\mathrm{T}}^{V}-$dependent correction is applied to the simulated events, as addressed in Section 6.2.2. The uncertainty of this NLO dependence reflects on a systematic uncertainty on the shape of the p_{T}^{W} spectrum for the signal. This is represented in Table 7.2 as `TheoryVHPt` and ranges from 2% to 2.5%.

Vector Boson+jets

Table 7.3 lists all the systematics uncertainties associated with the prediction of the W or Z+jets backgrounds [50], described bellow.

The overall normalisation of the W/Z+jets processes is a dominant source of uncertainty in the analysis. But as these backgrounds are easy to control with pure samples, data is used to constrain their normalisation during the final data fit: the floating normalisation of the $W/Z+cl$ jets and $W/Z+bb$ jets samples are respectively described as the `norm_W/Zcl` and `norm_W/Zbb` NPs in Table 7.3.

Jet multiplicity The jet multiplicity and flavour composition of the W/Z+jets backgrounds are not well modelled by the SHERPA generator. This results in scale uncertainties prior to the fit coming from the different acceptance to 2 or 3 jets events. Using enriched samples of

Name	Value	Description
W/ZlNorm	10/5%	Scale uncertainty
W/ZlNorm_J3	10/5%	Acceptance from jet multiplicity, 3 to 2 jets ratio
norm_W/Zcl	Float	Scale to determine from profilled fit
W/ZclNorm_J3	10/26%	Acceptance from jet multiplicity, 3 to 2 jets ratio
norm_W/Zbb	Float	Scale to determine from profilled fit
W/ZbbNorm_J3	10/20%	Acceptance from jet multiplicity, 3 to 2 jets ratio
W/ZblW/ZbbRatio	35/12%	Acceptance from flavour composition
W/ZbcW/ZbbRatio	12/12%	Acceptance from flavour composition
W/ZccW/ZbbRatio	12/12%	Acceptance from flavour composition
SysW/ZDPhi	$\Delta\phi(j_1,j_2)-$dependent	$W/Z\ \Delta\phi(j_1,j_2)$ correction
SysW/ZPtW/Z	$p_{\mathrm{T}}^{W/Z}-$dependent	p_{T}^{W} modelling / p_{T}^{Z} correction
SysW/ZMbb	$m_{b\bar{b}}-$dependent	$W/Z\ m_{b\bar{b}}$ modelling

Table 7.3: List and brief description of the systematic uncertainties affecting the modelling of the W/Z+jets background processes by MC. The scale and acceptance uncertainties values are shown separatelly for the W/Z+jets samples.

W/Z+light jets, obtained with the 0 $b-$tagged jets requirement, the data and SHERPA yields are compared to determine the W/ZlNorm uncertainty on the overall normalisation of the W/Z+light component. W/ZlNorm_J3 is used to parametrise the uncertainty on the 2-to-3 jets yields ratio, and is applied only to the 3 jets prediction. The normalisation of the $W/Z+cl$ and $W/Z+bb$ is handled by the norm_W/Zcl and norm_W/Zbb floating NPs, and therefore, totally derived from data with no prior assumptions. However, the uncertainties on the ratios between the 2-to-3 jets yields for these samples was determined by comparing the SHERPA with the ALPGEN models. They are applied to 3 jets events of the cl and bb component: W/ZclNorm_J3 and W/ZbbNorm_J3, respectively.

Jet flavour fraction As said before, the simulation of W/Z+jets using the SHERPA generator does not model well the flavour composition of these backgrounds. Therefore, the prediction is affected by systematic uncertainties on the relative jet flavour fractions. Their values were determined from comparisons between the SHERPA and ALPGEN jet flavour models, and are parametrised as W/ZblW/ZbbRatio, W/ZbcW/ZbbRatio and W/ZccW/ZbbRatio, to describe the effect on the acceptance of the uncertainty on the bl, bc and cc to bb ratios, respectively.

Modelling of p_{T}^{V} and $\Delta\phi(j_1,j_2)$ The MC models of the p_{T}^{V} and $\Delta\phi(j_1,j_2)$ spectra do not fully resemble the data observations and reweighting corrections are applied to simulation to account for these effects, as described in Section 6.2.2. Half the $\Delta\phi(j_1,j_2)$ reweighting is used as systematic uncertainty in the Z/W+light and $W+cl$ components. For the $Z+c$ and $Z+b$ events, half the correction is assigned as systematic. Since the $W+cc/b$ are not corrected in regard of $\Delta\phi(j_1,j_2)$, the full reweighting derived for the W+light and cl component is adopted as the systematic on these components. These are collectively referred to as SysW/ZDPhi, and are separated into the 2 and 3 jets categories, with the heavy flavour component, $b+c$, treated

Name	Value	Description
norm_ttbar	Float†	Scale to determine from profilled fit
ttbarHighPtV	7.5%	High to low p_{T}^{W} acceptance ratio
ttbarNorm_J3	20%	3 to 2 jets acceptance ratio
TopPt	Top $p_{\mathrm{T}}-$dependent	Top p_{T} correction
TtbarMetCont	$E_{\mathrm{T}}^{\mathrm{miss}}-$dependent	$E_{\mathrm{T}}^{\mathrm{miss}}$ shape
TtbarMBBCont	$m_{b\bar{b}}-$dependent	$m_{b\bar{b}}$ shape

Table 7.4: List and brief description of the systematic uncertainties affecting the modelling of the $t\bar{t}$ background process by MC. † The floating normalization scale is considered decorrelated in the 0, 1 and 2 leptons analysis channels.

as uncorrelated from the light flavour. The SysW/ZPtW/Z systematic corresponds to half the p_{T}^{Z} reweighting and is assigned to the $Z+b,c$ and light components. It also includes a shape uncertainty on the p_{T}^{W} spectrum, derived by comparing different MC models, ranging from +9% to -23%.

Modelling of $m_{b\bar{b}}$ For Z+jets, a shape uncertainty on the $m_{b\bar{b}}$ spectrum is used to cover the difference observed between the SHERPA model and real data in pure samples of Z+jets events. The derived $m_{b\bar{b}}-$dependent uncertainty, ranging from -3% to 5%, is sufficient to cover the difference between the nominal SHERPA model and the alternative ALPGEN sample. For W+jets, a $m_{b\bar{b}}$ shape uncertainty was derived by comparing the nominal MC sample with other generators. The size of this uncertainty ranges from -23% to +28%, depending on $m_{b\bar{b}}$. These are collectively designated as SysW/ZMbb.

Top quark $t\bar{t}$

The $t\bar{t}$ production is, along with W+jets, the dominant background of the WH search. Therefore, the systematic uncertainties on this background prediction have a great impact on the final uncertainty of the measurement. The sources of uncertainties assigned to $t\bar{t}$ come from the parton showering, hadronisation and PDFs models. These are described below and are summarised in Table 7.4 [50].

Overall normalisation The overall $t\bar{t}$ normalisation and uncertainty is extracted from data using the profiled likelihood fit, by letting the norm_ttbar scale factor adjust to the observation. This can be done since the WH analysis benefits from the 3 jets region that is very pure in top pair events. norm_ttbar is considered decorrelated in the 0, 1 and 2 leptons search channels in case of the VH combined fit, and therefore the normalisations of $t\bar{t}$ are determined independently for the different channels. This choice was based on the fact that the three channels probe different and orthogonal phase space regions.

Parton shower, hadronisation and PDF sets The different possibilities to model the parton shower and hadronisation mechanisms, and the choice of the PDF description influence the

Name	Value	Description
	$s-$channel	
stopsNorm	$\pm 4\%$	cross section
SChanAcerMC	from +13 to +30%	acceptance from higher-order effects
SChanAcerMCPS	from +4 to +8%	acceptance from parton shower
	$t-$channel	
stoptNorm	$\pm 4\%$	cross section
TChan	from -18 to +52%	accep. from higher-order, hadron. and PS
	$Wt-$channel	
stopWtNorm	$\pm 7\%$	cross section
WtChanAcerMC	from -15 to +4%, $m_{b\bar{b}}-/p_{\mathrm{T}}^{b_1}-$dep.	accep. and shape from higher-order effects
WtChanHerwig	from -3 to +5%, $p_{\mathrm{T}}^{b_1}-$dependent	accep. and shape from hadronization and PS

Table 7.5: List and brief description of the systematic uncertainties affecting the modelling of the single top background process by MC.

prediction of the physical processes, resulting in acceptance and shape uncertainties. These uncertainties were evaluated for the $t\bar{t}$ background by comparing the nominal NLO generator POWHEG+PYTHIA with generators containing alternative models for each of the simulation chain steps. The largest of the observed yield differences are taken as systematic uncertainties on the acceptance of $t\bar{t}$: ttbarHighPtV and ttbarNorm_J3. The former reflects the uncertainty on the high-to-low p_{T}^{V} acceptance ratio and is only assigned to the $p_{\mathrm{T}}^{V} > 120$ GeV (high) bin. The second is applied to 3 jets events and represents the uncertainty on the 2-to-3 jets regions acceptance. Shape uncertainties on the BDT input variables were investigated with a similar method. A reduced set of shape uncertainties was achieved by taking into account the correlation between the observables in the $t\bar{t}$ sample, i.e. if differences were observed in a given variable, the difference was only considered as a shape uncertainty if it was not covered already by a shape uncertainty on another variable. From this procedure two shape systematics were identified, associated with the $E_{\mathrm{T}}^{\mathrm{miss}}$ and $m_{b\bar{b}}$ spectra shapes: TtbarMetCont and TtbarMBBCont, respectively.

Modelling of $p_{\mathrm{T}}^{\mathrm{top}}$ The top p_{T} modelling by MC does not follow exactly the data spectra. So, a correction was applied to improve the top p_{T} modelling, as described in Section 6.2.2. Half the correction weight is assigned as a systematic uncertainty on the top background, TopPt.

Single Top

The determination of the systematic uncertainties associated with the single top prediction is one of the work subjects of this thesis. A more extensive description of this study are therefore presented in Section 7.1.4. Table 7.5 lists the uncertainties associated with the single top modelling. With the exception of the cross-section uncertainties, all the systematics associated with the single top prediction of the VH analysis [50] are the result of the dedicated study mentioned before.

Name	WW	WZ	ZZ	Description
VVJetScalePtST1	p_{T}^{V}−dependent			Scale and shape from higher-order effects, 3 jets
VVJetScalePtST2	p_{T}^{V}−dependent			Scale and shape from higher-order effects, 2 jets
VVJetPDFAlphaPt	3%	4%	3%	Acceptance from PDFs and α_s, 3 jets
VVJetPDFAlphaPt	2%	2%	3%	Acceptance from PDFs and α_s, 2 jets
VVMbb	$m_{b\bar{b}}$−dependent			$m_{b\bar{b}}$ shape from PS and hadronization

Table 7.6: List and brief description of the systematic uncertainties affecting the modelling of the dibosons background processes by MC.

Dibosons

WW, WZ and ZZ production, jointly quoted as the dibosons background, have a final state topology very similar to the signal. But as these processes cross-sections are not as large as $t\bar{t}$ and W/Z+jets it is difficult to obtain samples with a high level of purity in the context of the phase space regions defined in this analysis. So, data to MC comparison techniques can not be applied. Thus, a careful effort must be put on the knowledge of these processes relying solely on MC. The nominal generator used to predict the dibosons contribution is the NLO POWHEG generator with the hadronisation given from the PYTHIA8 models. The hadronisation and parton shower models, PDFs set, α_s and perturbative effects are all potential sources of systematic uncertainties of the dibosons prediction. The final list of systematics affecting the prediction of these backgrounds is summarised in Table 7.6 [50] and explained ahead.

Higher-order The higher-order perturbative corrections to the inclusive and differential cross-sections result in systematic uncertaintis on the normalisation and shape of observables, if those corrections are significant. In the case of the dibosons, the differential cross-section correction exhibits a non-negligible relation with p_{T}^{V}. In this way, the uncertainty from higher-order corrections is parametrised as a shape uncertainty that does not preserve the normalisation, acting as a scale uncertainty simultaneously. It is determined separately for the 3 and 2 jets bins: VVJetScalePtST1 and VVJetScalePtST2, respectively.

α_s and PDFs The impact of the PDF choice and α_s uncertainty is examined by comparing the nominal diboson samples with simulations having alternative PDF sets and with shifted α_s constants. From this procedure two acceptance uncertainties were identified for the different dibosons processes, that have distinct values for the 2 and 3 jets categories. These are listed in Table 7.6 as VVJetPDFAlphaPt.

Parton shower and hadronisation The uncertainties originating from the incomplete knowledge of the showering and hadronisation processes are evaluated by comparing the nominal POWHEG+PYTHIA8 models with HERWIG, featuring alternative representations of the PS and hadronisation mechanisms. A clear modification of the $m_{b\bar{b}}$ spectrum is observed and a shape uncertainty is assigned to cover this effect: VVMbb. No significant change was observed on the

obtained event yields and for this reason, no uncertainty is assigned to the dibosons acceptance due to the hadronisation and PS modelling.

Multijet

The multijet background samples were obtained from data by requiring non-isolated leptons of lower purity, as described in Section 6.4.3. Therefore, the multijet background definition depends on the selection rules applied to leptons and on the isolation requirements. The corresponding modelling systematics are obtained by varying the thresholds of the multijet lepton selection conditions and by evaluating their impact on the multijet sample [50]. In addition, uncertainties on the normalisation of the multijet background have origin on the statistical uncertainty of the multijet fit and must also account for the uncertainty on the non-multijet background subtracted to build the multijet template. These amount to 11%, 14% and 22% in the LL, MM and TT $b-$tagging categories, respectively for 2 jets events of the electron sub-channel. As in the muon sub-channel the multijet sample has a smaller size, the corresponding uncertainties are three times larger.

7.1.4 Determination of the Single Top Modelling Uncertainties

Three main mechanisms can give rise to a top quark in pp collisions, as already sketched in Figure 6.4. The $t-$channel has the largest cross-section of the three channels but the $s-$channel is the most similar to the signal and therefore has the highest acceptance. Together, these three mechanism contribute 10%, in average, to the total analysis background, as shown in Table 6.4.4, and can be as high as 127 times larger than the WH signal in some analysis bins. Therefore the significance of this background to this analysis is clear and the correct determination of the systematic uncertainties that affect its prediction have a real impact on the signal measurement.

However, since it is difficult to obtain a pure control region for single top events, the evaluation of the systematic uncertainties associated with the prediction of this background rely on MC comparisons.

Cross-section

Cross-section uncertainties of 4%, 4% and 7%, derived at NNLO [86, 87, 88, 89], affect the $s-$, $t-$ and $Wt-$ channels normalisation, `stopsNorm`, `stoptNorm` and `stopWtNorm`, respectively, as listed in Table 7.5. These parameters take into account the uncertainties on the QCD renormalisation and factorisation scales, on α_s and on the PDFs.

Process	MC generator	Description
$s-$ch.	POWHEGPYTHIA	baseline generator
	ACERMCPYTHIA	higher order effects
	MCATNLOJIMMY	hadronization, showering and underlying event
	ACERMCPYTHIA more vs less PS	parton shower effects
$t-$ch.	ACERMCPYTHIA	baseline generator
	AMCATNLOJIMMY	higher order, hadronization, showering and underlying event
	ACERMCPYTHIA more vs less PS	parton shower effects
$Wt-$ch.	POWHEGPYTHIA DR	baseline generator
	POWHEGHERWIG	hadronization and showering models
	ACERMCPYTHIA	higher order effects
	POWHEGPYTHIA DS	Diagram Removal (DR) vs Subtraction (DS)
	ACERMCPYTHIA more vs less PS	parton shower effects

Table 7.7: Monte Carlo samples used to determine the single top systematics and corresponding event generation effects investigated.

Acceptance and Shape

Besides this cross-section uncertainty, the PDF definition, order on the perturbative theory at which the hard scatter is generated, the hadronisation models and showering processes are all potential sources of systematic uncertainties. In order to evaluate their impact on the top prediction, the nominal predictions were compared to the predictions of other generators.

MC simulation Table 7.7 lists the alternative MC generators used, together with the effect that is tested by comparing them to the nominal simulation. Table 7.8 shows the details of each generator.

For the $t-$channel, the baseline generator ACERMCPYTHIA, LO in QCD, was compared with the NLO AMCATNLO interfaced with the underlying event model provided by JIMMY to explore the effect of the higher-order QCD corrections and uncertainties on the underlying event description.

The baseline $s-$channel POWHEGPYTHIA sample, generated at NLO in QCD, was compared to the LO ACERMCPYTHIA to probe the impact of higher-order QCD corrections. The MCATNLO NLO generator interfaced with the JIMMY underlying event model and HERWIG-based mechanisms of parton shower and hadronisation, allows the possibility of testing alternative descriptions of these mechanisms.

For the $Wt-$channel, the baseline POWHEGPYTHIA sample was compared with ACERMCPYTHIA and POWHEGHERWIG to investigate the $Wt-$channel prediction dependence on higher order-effects, and on the definition of the hadronisation and showering models.

For all the channels, the effect of increasing or decreasing the amount of parton shower was also studied by comparing the ACERMCPYTHIAMOREPS sample with the AcerMCPythiaLessPS sample.

Single top production via the $Wt-$channel at NLO has an interference with top pair production at LO as Figure 7.1 shows. Two methods for resolving this interference, and avoid double counting this process, are considered at event generation level: diagram subtraction

Process	Generator, PDF set and MC tune	N_{events}
$s-$channel	AcerMCPythia, P2011CCTEQ6L1	1199999
	McAtNloJimmy, AUET2CT10 ($W \to e\nu$)	999998
	McAtNloJimmy, AUET2CT10 ($W \to \mu\nu$)	998000
	McAtNloJimmy, AUET2CT10 ($W \to \tau\nu$)	999998
	AcerMCPythia, P2011CMorePSCTEQ6L1	1199999
	AcerMCPythia, P2011CLessPSCTEQ6L1	1198998
$t-$channel	aMcAtNloJimmy, AUET2CT10	996999
	AcerMCPythia, P2011CMorePSCTEQ6L1	2978000
	AcerMCPythia, P2011CLessPSCTEQ6L1	2997999
$Wt-$channel	AcerMCPythia, P2011CCTEQ6L1	998997
	PowhegHerwig, AUET2CT10	998896
	PowhegPythia, P2011C, DS	994894
	AcerMCPythia, P2011CMorePSCTEQ6L1	998999
	AcerMCPythia, P2011CLessPSCTEQ6L1	1000000

Table 7.8: Monte Carlo samples and statistics used to determine the single top systematics.

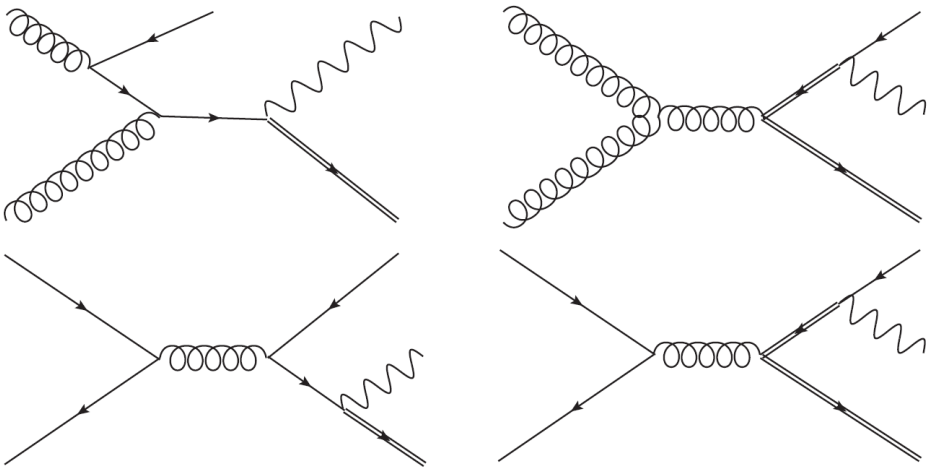


Figure 7.1: (Left) $Wt-$channel diagram at NLO. (Right) Top pair production at LO. Double fermionic lines represent the top quark. Taken from [90].

(DS) and diagram removal (DR). The DR method is the default technique to deal with the $t\bar{t}$ interference. Here, the $t\bar{t}$ contribution is not generated. On the DS method, the $t\bar{t}$ process is generated but afterwards subtracted at cross-section level [90]. So, for the $Wt-$channel, the interference treatment is a source of uncertainty and the two methods have to be compared to evaluate the size of the effect.

Acceptance In order to determine the systematic uncertainties affecting the single top prediction, the events are selected according to the 1 lepton selection described in Section 6.4. Acceptance uncertainties result from the difference between the yields of the alternative samples with respect to the baseline sample. The latter are presented in Table 7.9. Acceptance differences due to more or less parton shower are obtained through direct comparison between the AcerMCPythia samples with MorePS and LessPS. In general, the acceptance is larger

Sample	Acceptance Difference (%)			
	2 jets		3 jets	
	$p_{\mathrm{T}}^{W} < 120$ GeV	$p_{\mathrm{T}}^{W} > 120$ GeV	$p_{\mathrm{T}}^{W} < 120$ GeV	$p_{\mathrm{T}}^{W} > 120$ GeV
$s-$channel				
ACERMCPYTHIA	$+(13.0 \pm 0.2)$	$+(21.8 \pm 0.6)$	$+(17.6 \pm 0.4)$	$+(30.2 \pm 0.8)$
MCATNLOJIMMY	$+(1.8 \pm 0.3)$	$-(1.7 \pm 0.7)$	$+(6.4 \pm 0.4)$	$+(0.8 \pm 1.1)$
ACERMCPYTHIALess/MorePS	$+(11.8 \pm 0.5)$	$+(16 \pm 1)$	$+(6.7 \pm 0.8)$	$+(8 \pm 2)$
$t-$channel				
AMCATNLOJIMMY	$+(51.6 \pm 0.5)$	$+(25 \pm 1)$	$+(12.1 \pm 0.8)$	$-(18 \pm 3)$
ACERMCPYTHIALess/MorePS	$+(15 \pm 1)$	$+(16 \pm 3)$	$+(10 \pm 1)$	$+(19 \pm 3)$
$Wt-$channel				
ACERMCPYTHIA	$+(1.2 \pm 0.6)$	$-(2 \pm 1)$	$+(3.9 \pm 0.5)$	$-(15 \pm 1)$
POWHEGHERWIG	$+(4.9 \pm 0.6)$	$+(2.4 \pm 1)$	$+(4.7 \pm 0.5)$	$-(2.8 \pm 0.9)$
POWHEGPYTHIADS	$+(7.6 \pm 0.6)$	$+(0 \pm 1)$	$+(5.2 \pm 0.5)$	$-(16 \pm 1)$
ACERMCPYTHIALess/MorePS	$+(0 \pm 2)$	$+(4 \pm 4)$	$+(8 \pm 2)$	$+(4 \pm 4)$

Table 7.9: Acceptance difference for single top obtained by comparing different Monte Carlo samples to the baseline prediction. For the hadronization effects, the acceptance difference between the ACERMCPYTHIA samples with More and Less parton shower is shown instead. The statistical uncertainty related to the MC samples size is shown.

for events with less parton shower, no matter the top production channel, as these generate fewer additional jets and the analysis selection specifically requires only 2 or 3 signal jets.

For the $s-$channel, significant acceptance differences arise from higher-order effects. Events generated at NLO have up to 22% larger acceptance, shown in Table 7.9. This difference is taken as a systematic uncertainty on the normalisation of the $s-$channel prediction, `SChanAcerMC`, as was summarised in Table 7.5. Showering effects, evaluated by comparing more or less PS extreme conditions, result in a normalisation systematic, `SChanAcerMCPS`, corresponding to half the acceptance difference observed.

$t-$channel production yields after the 1-lepton selection differ up to 52% with respect to the baseline sample when using the AMCATNLOJIMMY generator. The difference is due to the fact that not only the generator order is different, the baseline being at LO and the alternative at NLO, but also the showering, the hadronisation and the underlying event models change from one sample to the other. AMCATNLOJIMMY has on average a lower jet multiplicity than ACERMCPYTHIA and therefore larger acceptance. The differences come also from the $b-$tagging selection and te jets p_{T}. The difference is taken as a systematic uncertainty on the normalisation of the $t-$channel, `TChan`. Differences coming from the showering alone are small compared to this systematic and considered to be contained already in this uncertainty, so no additional systematic from the more and less PS comparison is used.

For the $Wt-$ events, acceptance differences from the comparisons between the baseline generator and ACERMCPYTHIA and POWHEGHERWIG are used as the scale systematic uncertainties, `WtChanAcerMC` and `WtChanHerwig`, respectively. As for the $t-$channel case, the acceptance differences associated with the more/less PS tunes and to the DS method were not significant enough to justify the need of an additional systematic.

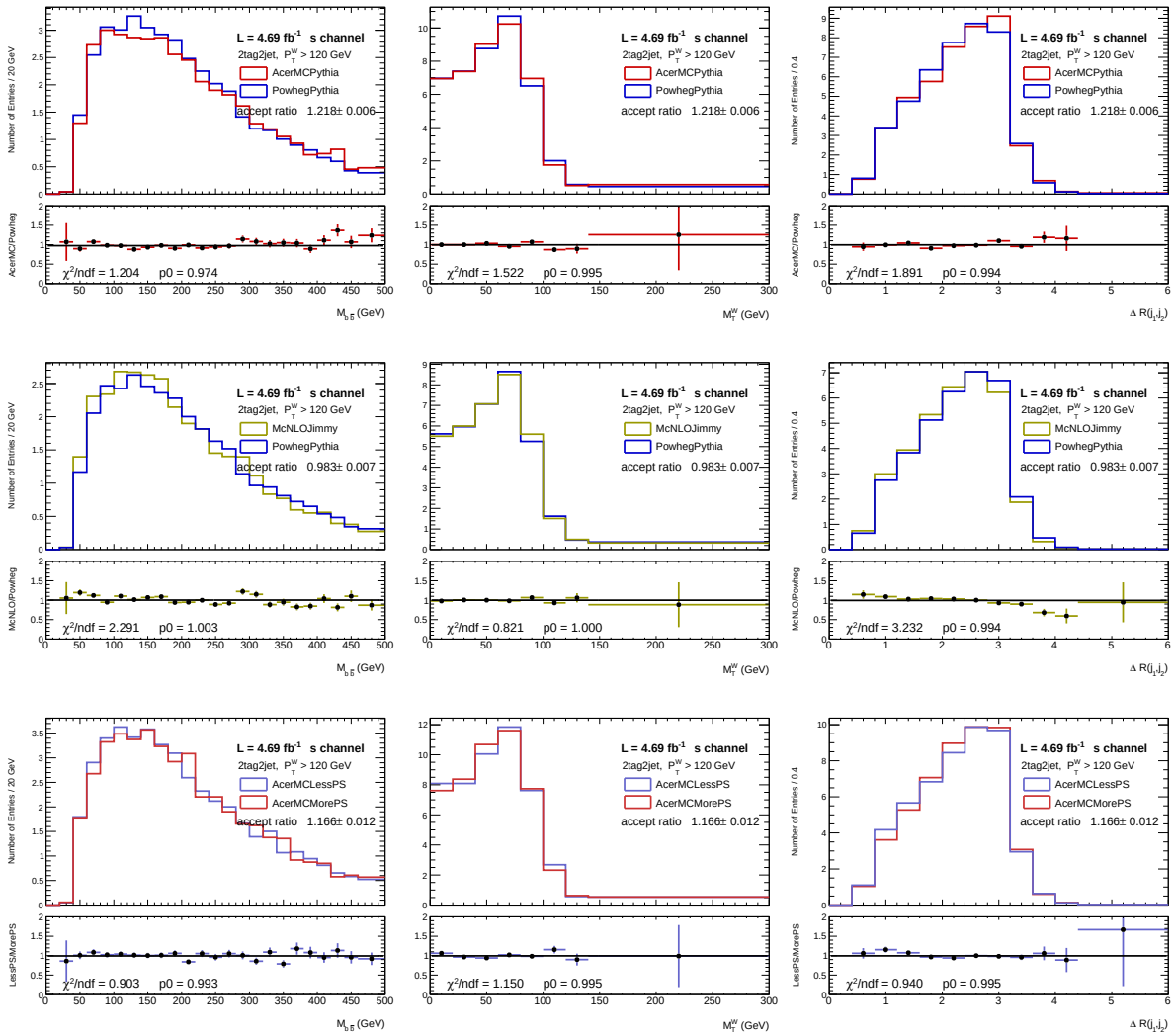


Figure 7.2: $m_{b\bar{b}}$, m_{T}^{W} and $\Delta R(b_1, b_2)$ distributions for the $s-$channel single top production, using different MC generators. The distributions contain 1-lepton events for $p_{\mathrm{T}}^{W} > 120$ GeV with 2 b-tagged jets, and are normalised to the same integral.

Sample	Category	Linear fit
	$Wt-$channel	
POWHEGHERWIG	2 jets, p_{T}^{W} <120 GeV	$\pm$(1.3 - 0.004$\times p_{\mathrm{T}}^{b_1}$)
ACERMCPYTHIA	2 jets, p_{T}^{W} >120 GeV	$\pm$(0.6 + 0.004$\times m_{b\bar{b}}$)
	3 jets, p_{T}^{W} <120 GeV	$\pm$(1.2 -0.003$\times p_{\mathrm{T}}^{b_1}$)
	3 jets, p_{T}^{W} >120 GeV	$\pm$(1.4 -0.003$\times m_{b\bar{b}}$)

Table 7.10: Single-top systematics obtained by comparing different Monte Carlo samples. $p_{\mathrm{T}}^{b_1}$ and $m_{b\bar{b}}$ are the p_{T} of the leading b-jet and the di-bjet invariant mass.

Observables shape The impact of the generators on the modelling of the set of variables used as input to the BDT method was investigated. If a statistically significant difference in shape is encountered, an uncertainty is assigned as a function of that same variable.

Figures 7.2, 7.3 and 7.4 show the distributions of $m_{b\bar{b}}$, m_{T}^{W} and $\Delta R(b_1, b_2)$ for the three top production channels for events with two $b-$tagged jets and p_{T}^{W} >120 GeV. These distributions are shown for different generator hypothesis and the baseline sample. The normalisations are fixed to agree, as their differences were already taken as the acceptance systematic. The bottom panel presents the ratio of these distributions and the χ^2 spread of the data relative to a linear fit hypothesis without slope. With the exception of $m_{b\bar{b}}$ in the comparison between ACERMCPYTHIA and the baseline sample in the $Wt-$channel, these variables present no systematic modelling difference and therefore no shape uncertainty must be considered. In the $m_{b\bar{b}}$ case, the shape difference is manifest for the ACERMCPYTHIA model with respect to the baseline model and a systematic uncertainty on the $m_{b\bar{b}}$ spectrum is derived.

Figure 7.5(a) shows the more prominent modelling differences encountered. These are the $p_{\mathrm{T}}^{b_1}$ for $p_{\mathrm{T}}^{W} < 120$ GeV events and $m_{b\bar{b}}$ for $p_{\mathrm{T}}^{W} > 120$ GeV, for 2 and 3 jets events. A linear function is fitted to the ratio plots to parametrise the differences as a function of each variable. The results of the fit are shown in Table 7.10. These are used as shape uncertainties for the $Wt-$channel and cover for less prominent modelling differences encountered on other variables.

The impact of the modelling difference on the final BDT output is presented in Figure 7.5(b), showing the BDT output distributions obtained with the baseline generator and the with the alternativee generators. The baseline sample is also exhibited after the shape systematic has been applied. The plots show that the BDT differences fall within the shape uncertainties derived, indicating that these are sufficient to parametrise the MC modelling differences for the various BDT inputs.

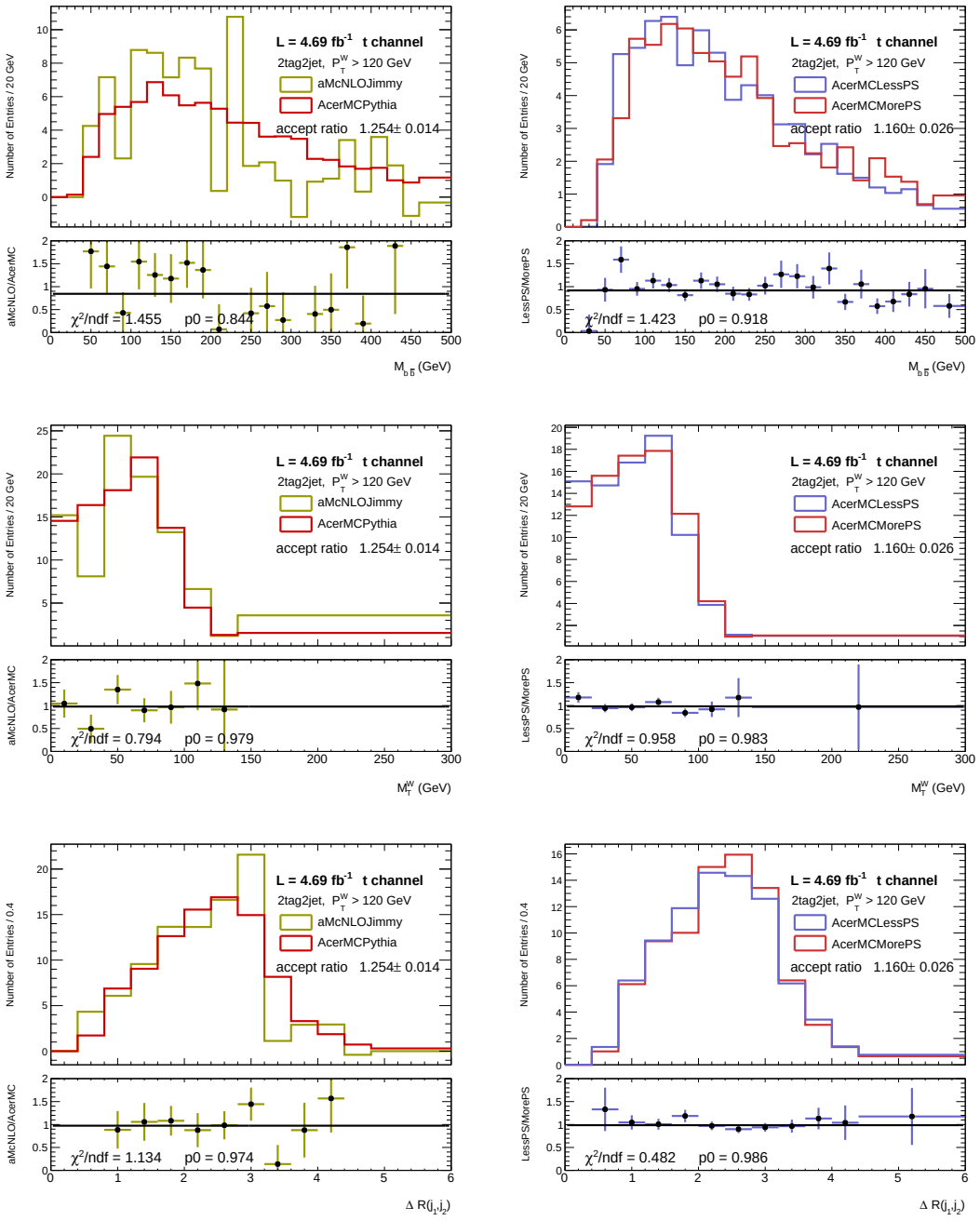


Figure 7.3: $m_{b\bar{b}}$, m_{T}^{W} and $\Delta R(b_1, b_2)$ distributions for the $t-$channel single top production, using different MC generators. The distributions contain 1-lepton events for $p_{\mathrm{T}}^{W} > 120$ GeV with 2 b-tagged jets, and are normalised to the same integral.

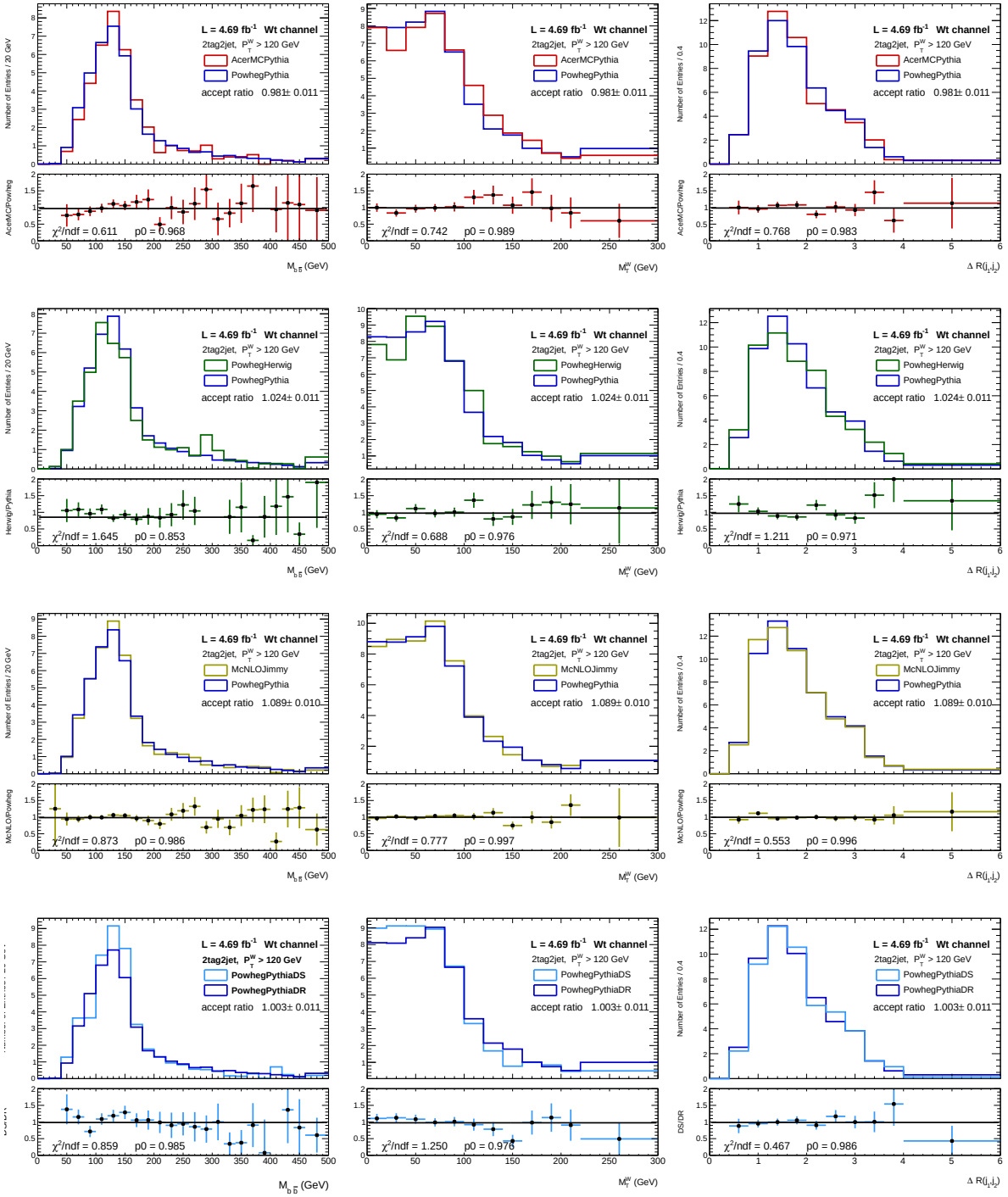
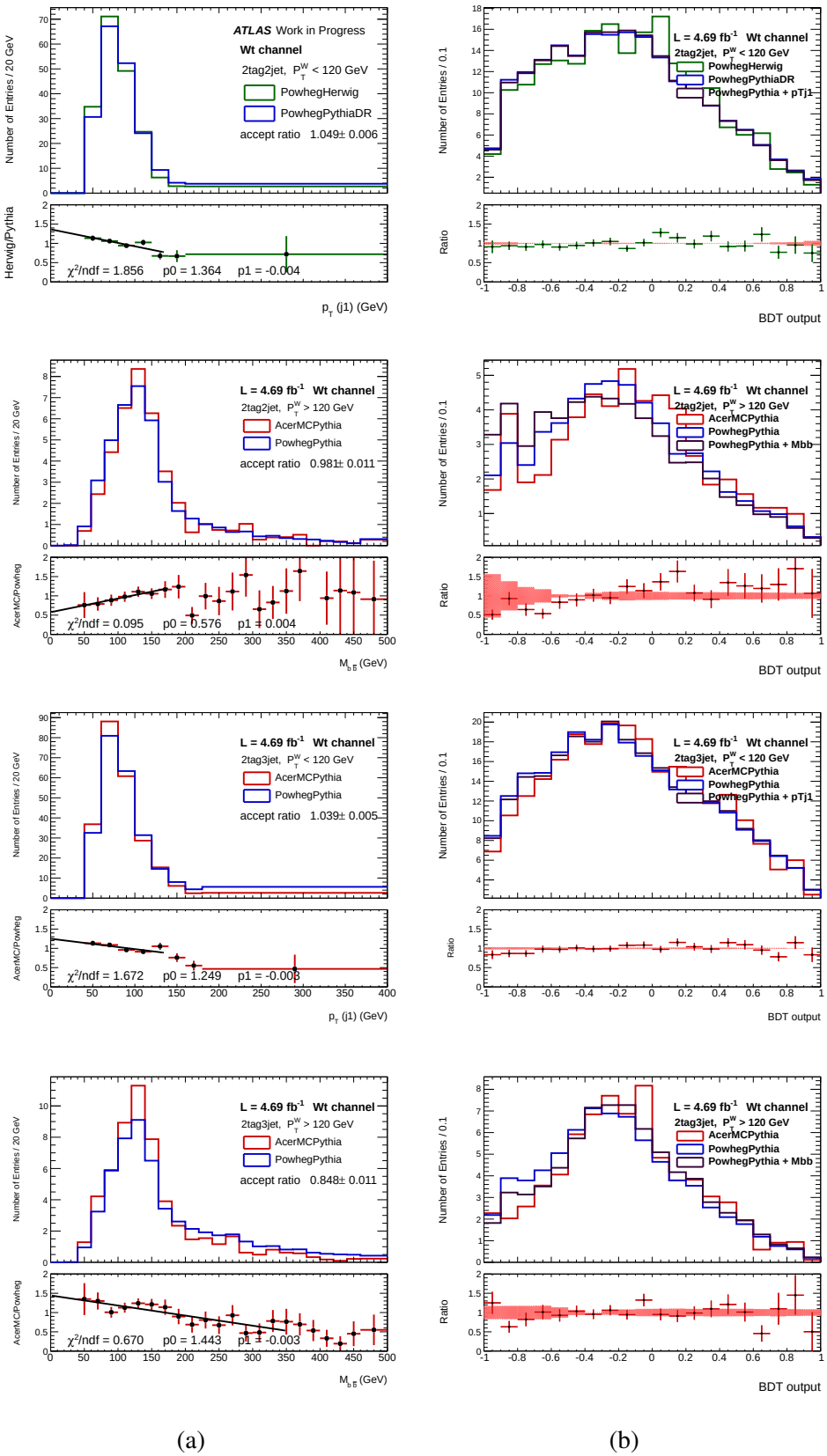


Figure 7.4: $m_{b\bar{b}}$, m_{T}^{W} and $\Delta R(b_1, b_2)$ distributions for the $Wt-$channel single top production, using different MC generators. The distributions contain 1-lepton events for $p_{\mathrm{T}}^{W} > 120$ GeV with 2 b-tagged jets, and are normalised to the same integral.



(a) (b)

Figure 7.5: *Wt*−channel shape systematics, obtained by comparing HERWIG with PYTHIA and ACERMC with POWHEG. (a) Linear fit to the ratio between the alternative and the baseline distributions. (b) BDT distribution for the nominal, nominal plus shape uncertainty and alternative samples. The bottom pad shows the ratio between the alternative and the baseline BDT shapes and the shape uncertainty.

7.2 Statistical Analysis of Data and Simulation

A likelihood fit is used to measure the probability of the data observation given the signal plus background hypothesis as predicted by MC. Systematic uncertainties are incorporated in the likelihood function as a set of nuisance parameters. The signal significance is determined as the probability of obtaining the data measurement given the background-only hypothesis.

7.2.1 Fit Regions

The analysis categories described in Section 6.4 are all given as input to the final data fit. These were defined for the VH 1 lepton analysis but are similarly defined for the 0 and 2 leptons cases. The distributions used are the BDT discriminant for the most sensitive regions of two $b-$tagged jets, and the output of the $b-$tagging algorithm for the leading jet, MV1c(b_1). In total 38 regions enter the fit procedure.

Inputs Transformation

The BDT distributions, originally with 1000 bins equally segmenting the $x-$axis ranging from -1 to +1, are re-binned in order to reduce the number of bins that the fit procedure must manage. The BDT bins are merged such that the backgrounds are smoothed and the signal significant parts of the BDT spectrum have fine granularity [50]. Starting from the upper limit of the BDT spectrum, the bins are consecutively merged if a condition based on the relative proportions of signal and background is satisfied. The statistical uncertainty associated with the background prediction must be smaller than 10% of the total merged bin yield. This condition is imposed in order to shield the output of this technique from statistical fluctuations.

Figure 7.6 shows the outcome of the binning transformation. The total background is smooth after the transformation and the statistical uncertainty is well distributed across the different bins.

Input Distributions

Figures 7.7 and 7.8 present all the distributions used as input to the fit of the 1 lepton channel. The BDT output distributions are transformed accordingly to what was discussed before.

7.2.2 Likelihood and Significance

The statistical analysis of data and MC provides the final results of the analysis. In the VH search, the most important parameters of the measurement are the signal strength μ, that reveals the compatibility of the measured signal to the model prediction, and the signal significance, the probability of obtaining the data measurement in the background-only hypothesis. These measurements are accomplished by a maximum binned likelihood fit.

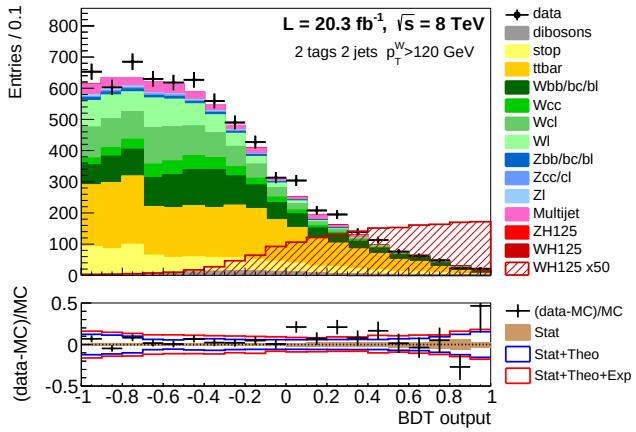


Figure 7.6: Distribution of the BDT output after the transformation for events with 2 jets and $p_{\mathrm{T}}^{W} > 120$ GeV.

Likelihood Function

Equation 7.1 represents the likelihood function used to determine the parameters of interest.

$$\mathscr{L} = \underbrace{\prod_{i,b}^{\text{Regions bins}} f\left(N_{ib}\middle|\mu\cdot S_{ib}\cdot \prod_{r}^{\text{Sys + Stat}} \nu_{br}(\theta_r) + \sum_{k}^{Bkg}\beta_k\cdot B_{kib}\cdot \prod_{s}^{\text{Sys + Stat}} \nu_{bs}(\theta_s)\right)}_{\text{Poisson with signal strength } \mu\text{; predictions S, B}} \cdot \underbrace{\prod_{t}^{\text{Sys + Stat}\ \in\{r,s\}} g\left(\vartheta_t\middle|\theta_t\right)}_{\text{Gaussian for Syst. and MC Stat.}} \cdot \underbrace{\prod_{n}^{\text{Norm Sys}\ \in\{r,s\}} \ell\left(\vartheta_n\middle|\theta_n\right)}_{\text{Log-normal for Norm. Syst.}} \tag{7.1}$$

- The first term represents the probability of having obtained N_{ib} data events in bin b of the ith analysis region distribution, and is parametrised by a Poisson function where the mean parameter is the signal plus background prediction for the given bin and analysis region. The prediction for the bin b and analysis region i is the sum of the signal yield S_{ib} and each of the backgrounds S_{kib}, for k running over the different backgrounds.

- The signal contribution is multiplied by the signal strength parameter μ, to be determined through the maximisation of the likelihood function.

- In a similar way, some of the backgrounds are associated with a floating scaling factor β_k, allowing their contribution to better fit the data. This is done for the major backgrounds, where a large purity is reached in certain analysis regions, and data can be used to constrain MC. For the remaining, β_k is simply set to 1.

- The systematic and MC statistical uncertainties are parsed as a set of nuisance parameters (NPs) here represented as θ, and the dependence of the expected event yields on these uncertainties is parametrised through the $\nu_b(\theta)$ functions. The set of θs is also left as almost free parameters of the fit to let the data dictate their best values.

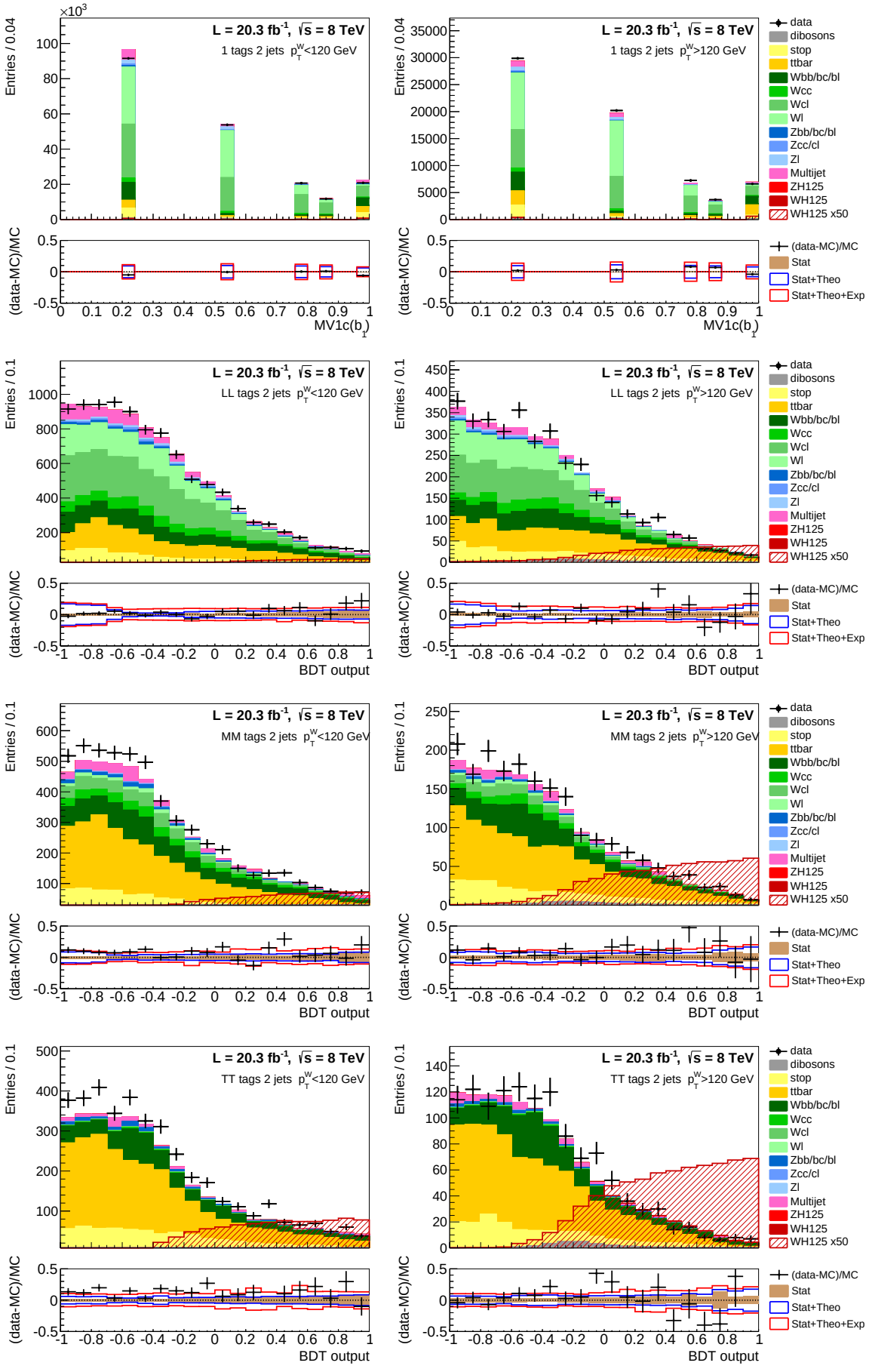


Figure 7.7: Distributions used as input to the VH statistical analysis for the 1 lepton channel, for the 2 jets category.

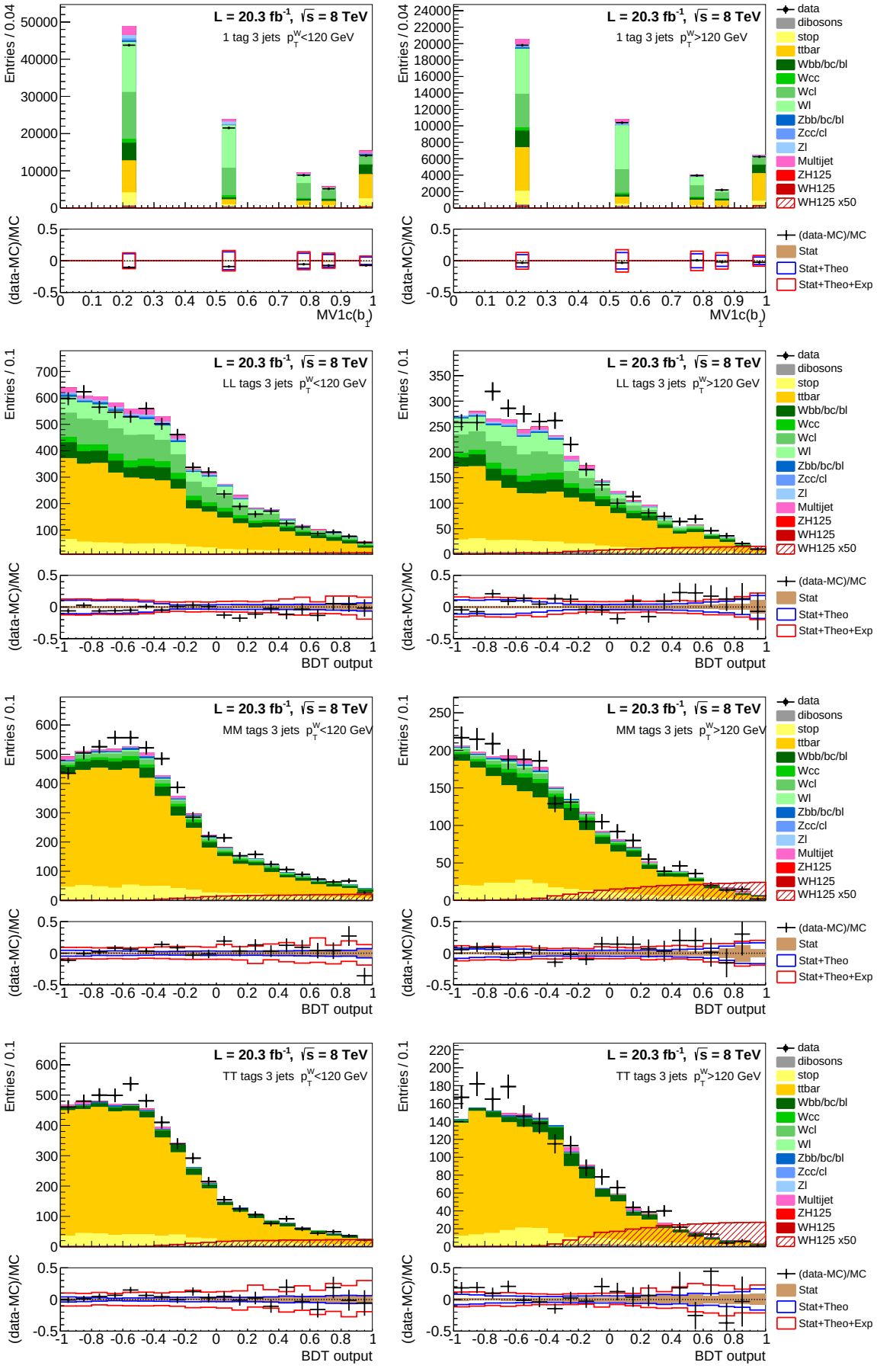


Figure 7.8: Distributions used as input to the VH statistical analysis for the 1 lepton channels, for the 3 jets category.

- The second term of the likelihood function employ a gaussian probability density function to incorporate the prior knowledge on the NPs θ_t central value ϑ_t, acting as a penalty term in the likelihood preventing θ_t to be very different from its prior value, where t runs over all the MC statistical uncertainties and systematics other than normalisation.

- The NPs associated with normalisation systematics θ_n are constrained with the third term of the likelihood function, where a log-normal function is used to prevent normalisation factors to assume negative values. ϑ_n represent the prior knowledge on the NP θ_n, where n runs over all the normalisation systematics.

Parameters of Interest

The measurements of the analysis are achieved by maximising the likelihood function with respect to all its parameters, the value of μ that maximises the likelihood being the observed signal strength and equally to the remaining factors.

The β_k scaling factors were already presented in Section 7.1.3 and will act on the $t\bar{t}$, $W/Z+bb$ and $W/Z+cl$ samples normalisations. The $t\bar{t}$ scaling factor is explicitly allowed to vary uncorrelated across the 0, 1 and 2 lepton channels of the VH analysis under the argument that each of the channels are probing very different regions of the phase space. Nevertheless, Figures 7.7 and 7.8 illustrate how data is used to put constraints on the normalisation of the backgrounds, by adjusting the β_k factors, in the 1 lepton channel. Some regions containing events with 3 jets are almost pure samples of $t\bar{t}$ and therefore its scale can be directly measured from data. Simultaneously, $W+cl$ can be constrained with the 1-tag data samples if $t\bar{t}$ is more strongly fixed elsewhere as it is the case. In the $W+bb$ jets case, none of the regions are as pure as in the $t\bar{t}$ case but still, given the number of regions and bins given to the fit, this background ends up to be controlled across the fit phase space. The Z+jets-related normalisation is evidently more important to the 0 and 2 leptons channels and constrained there with much more statistical power.

In the case of the VH analysis, since the search is done using three signal channels as described, the measurement of the signal can be done in the following alternative ways:

- Combining all information into a single signal strength parameter, μ_{VH}.
- Split by production mode into a WH and a ZH measurement, yielding μ_{WH} and μ_{ZH}, respectively.
- Split according to the three channels, from where results μ_{0lep}, μ_{1lep} and μ_{2lep} associated with the total signal observed in the 0, 1 and 2 leptons channels, respectively.

To be able to perform such analyses, the likelihood function must be defined accordingly. As the WH signal will dominantly fall within the 1 lepton channel, and not many of the ZH signal ends misidentified as a 1 lepton event, μ_{WH} and μ_{1lep} are practically measuring the same signal. The only reason why this is not exactly true is because of the residual ZH (WH) contamination on the 1 (0 and 2) lepton signal sample.

Statistical Test

The data is tested with respect to a given hypothesis using the test statistic q_μ, defined as a function of the signal strength μ parameter:

$$q_\mu = -2\ln\frac{\mathscr{L}(\mu,\theta)}{\mathscr{L}_{max}} \tag{7.2}$$

where $\mathscr{L}_{max}$ is the likelihood function value maximised unconditionally with respect to all parameters, and $\mathscr{L}(\mu,\theta)$ is the maximum likelihood value maximised with respect to all θ parameters but fixing μ to a chosen value.

The agreement between data and an hypothesis is quantified by the $p-$value:

$$p_\mu = \int_{q_\mu,\text{obs}}^{\infty} f(q_\mu|\mu)dq_\mu \tag{7.3}$$

$q_{\mu,\text{obs}}$ is the observed value of the test statistic and $f(q_\mu|\mu)$ its probability density function, approximated to be given by a χ^2 distribution. The probability of the background-only hypothesis, commonly designated p_0, is tested by setting $\mu = 0$. Conversely, the background plus signal hypothesis probability is evaluated with $\mu = 1$.

Significance

The signal significance S is expressed in terms of standard deviations of a gaussian distribution, and is evaluated by:

$$S = \sqrt{2}\phi^{-1}(1-2p_0) \tag{7.4}$$

where ϕ^{-1} is the inverse of the cumulative Gaussian distribution, also called error function. The significance can be defined as an expected or observed significance. In the case of the expected significance, no real data is used and an Asimov data set is constructed based on the MC prediction. This data set is generated from simulation setting all the analysis NPs to their best-fit values provided by a fit to the data, with the exception of the signal strength that is set to the nominal value of 1.

7.3 Results

The statistical treatment described above is used to extract the observed signal strength and significance. The results of the search for the VH 1 lepton channel, using the outcome of the LIP analysis code will be presented and discussed here. Furthermore, the MVA improvement study detailed in Section 6.5 will be probed here by testing the impact on the expected signal significance of adding new discriminant variables to the default BDT.

The maximum likelihood fit performed to data and prediction is implemented in a framework developed and maintained by the VH analysis group. This framework was shared

Scale Factor (β_k)	ATLAS Run 1	1 lepton LIP
$t\bar{t}$ 0ℓ	1.33 ± 0.14	1.16 ± 0.17
$t\bar{t}$ 1ℓ	1.13 ± 0.09	1.14 ± 0.12
$t\bar{t}$ 2ℓ	1.00 ± 0.04	1.00 ± 0.05
Wbb	0.79 ± 0.15	0.88 ± 0.16
Wcl	1.14 ± 0.10	1.09 ± 0.05
Zbb	1.09 ± 0.05	1.10 ± 0.05
Zcl	0.87 ± 0.12	0.82 ± 0.12

Table 7.11: Background-specific scale factors obtained from the VH combined fit to the ATLAS Run 1 distributions and replacing the 1 lepton distributions by the ones obtained with the LIP analysis code.

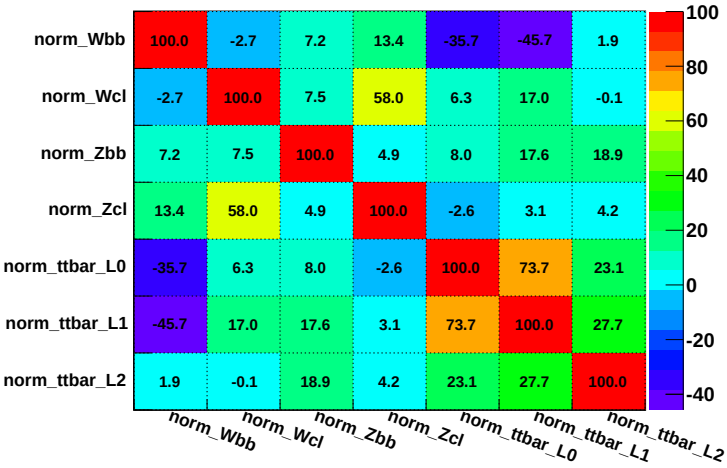


Figure 7.9: Correlation of the background scale nuisance parameters, in %.

between the group users, allowing an easy analysis of each group distributions and an easy interpretation and comparison of the results obtained.

The results of the analysis described in this thesis use the 1 lepton channel distributions produced through the LIP analysis chain, and inputs of the other channels prepared by other ATLAS groups. These are compared to the ATLAS Run 1 results obtained by a private fit of the 8 TeV data set. The number of parameters of interest is configurable in the fit, to measure either the total VH signal or the WH and ZH signals independently. The choice can also be to measure the signal present in the 0, 1 or 2 leptons channels distributions. In either case, the fit is performed in a combined fashion, meaning that the information of the three channels is exploited. Additionally, the fit framework was configured to measure the signal in the 1 lepton channel, using 1 lepton distributions alone.

Table 7.11 shows the background-specific scale factors defined by the β_k parameters of the likelihood function in Equation 7.1. According to the data measurement $t\bar{t}$ is underestimated in the 1 lepton channel, resulting in a scale factor with a central value greater than 1. The opposite is observed for $W + bb$. The scale factors obtained by the two fitted sets are compatible within the associated uncertainty. The most significant changes, of -12% and 11%, occur for $t\bar{t}$ in

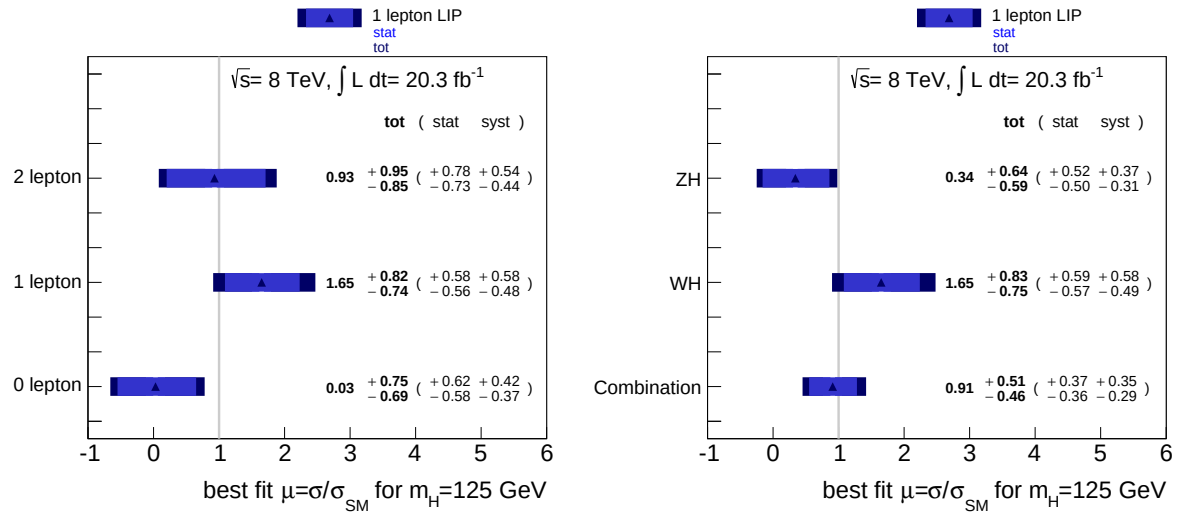


Figure 7.10: Strength of (left) the signal in the 0, 1 and 2 lepton channels and of (right) the WH, ZH and combined VH signal. The results were obtained from fitting the 1 lepton channel distributions from LIP and the 0 and 2 lepton inputs from the ATLAS 8 TeV publication.

Significance [σ]	Expected	Observed
VH	2.24	2.03
1ℓ 1 poi	1.25	1.61

Table 7.12: Expected and observed significance for the VH combined fit and for the 1 lepton fit with 1 parameter of interest (poi).

the 0 lepton channel and for $W+bb$ respectively. Since $W+bb$ is mostly normalised in the 1 lepton channel, replacing the 1 lepton inputs has a direct influence on the measurement of the scale of this process. Figure 7.9 shows that the $W+bb$ and the $t\bar{t}$ 0 lepton scale factors are anti-correlated, so the increase of $W+bb$ by 11% contributes to a decrease of the $t\bar{t}$ 0 lepton normalisation. Also lowering the $t\bar{t}$ 0 lepton scale is the $W+cl$ change by -4%, given the fact that these two nuisance parameters are correlated.

An observation can not yet be claimed, as both the expected and observed signal significance are well below 5σ: the observed (expected) significance is of the order of 2σ. These are shown in Table 7.12 for the VH combined fit and for the 1 lepton channel alone.

Figure 7.10 shows the signal strength results. For the 1 lepton channel, a signal strength of $\mu_{1lep} = 1.65^{+0.82}_{-0.74}$ was obtained. The WH search resulted in $\mu_{WH} = 1.65^{+0.83}_{-0.75}$, almost the same as μ_{1lep}, as expected from the very low contamination of ZH events in the 1 lepton signal sample.

All the signal strengths are compatible with the standard model prediction of $\mu = 1$, within uncertainties. But given the error size, the results are also compatible with the background-only hypothesis of $\mu = 0$. The VH uncertainties associated with the data set statistics are larger than the systematic uncertainties, and therefore the analysis can clearly benefit from a larger integrated luminosity. This comes mostly from the ZH side, since for the WH search, the

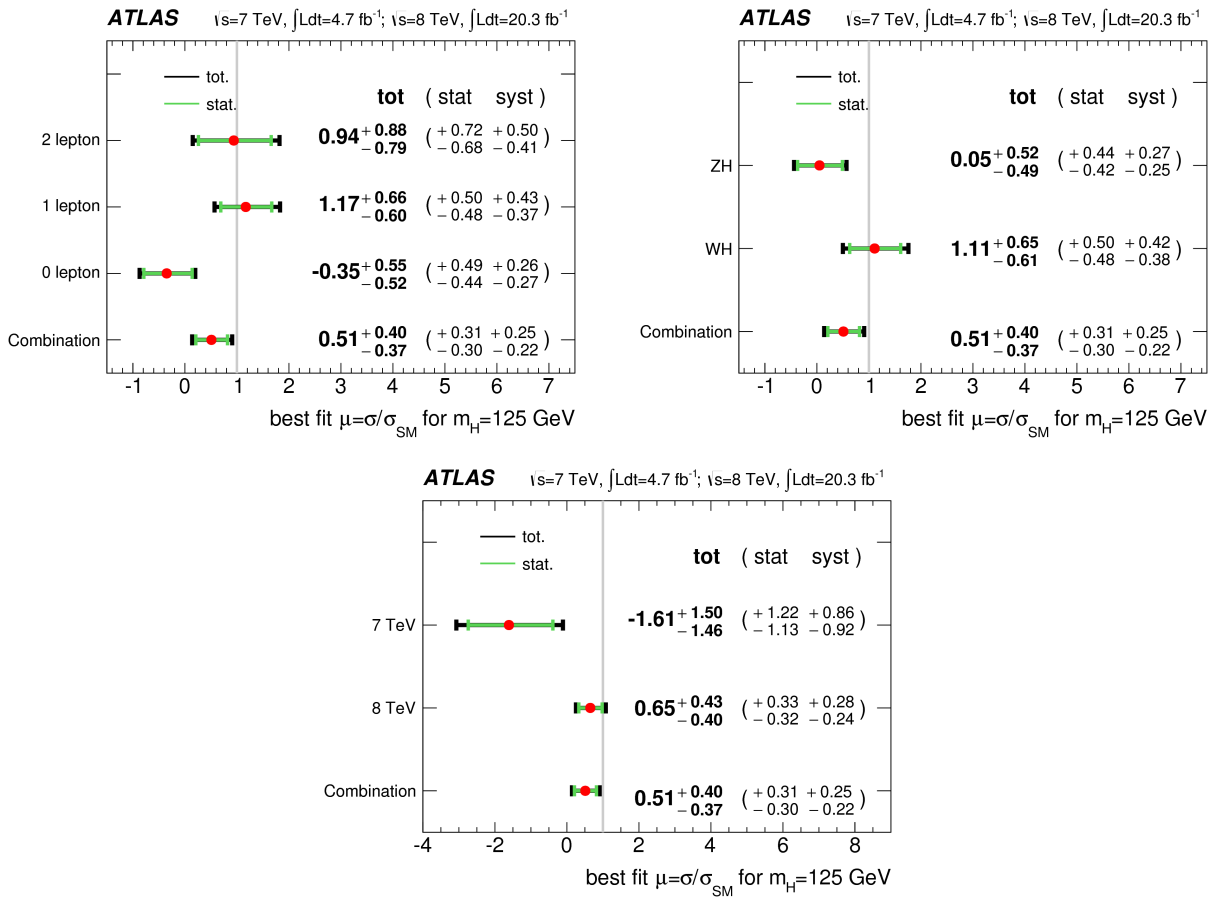


Figure 7.11: Strength of the signal in the 0, 1 and 2 lepton channels and of the WH and ZH for the 7 and 8 TeV data set. The signal strength of the combined VH signal is also shown separately per data set. Taken from [50].

systematic and statistical errors are similar.

Figure 7.11 shows the ATLAS results for the 7 and 8 TeV data sets [50]. The results obtained with the 1 lepton channel inputs from LIP do not differ much from the ATLAS publication. Although the latter do not use the 7 TeV data set, this is expected to have a lower impact on the final result than the 8 TeV data, given the much lower statistics: 4.7 fb^{-1}against the 20.3 fb^{-1}at 8 TeV.

The main conclusions are similar to those drawn already: the signal is compatible with the SM Higgs prediction within uncertainties, but also with the background-only hypothesis. The observed (expected) VH significance was 1.63σ (2.49σ). In the 7 TeV data set, the signal strength is negative, expressing that a downwards fluctuation of the background occurred in data.

Figure 7.12 shows the results obtained by the CMS experiment in a similar analysis [91]. The $W \to \tau\nu$ decay is additionally considered in the 1 lepton channel and a BDT analysis is also performed. The central values of the signal strength are very close for CMS and ATLAS in the 1 and 2 lepton channels, while for the 0 lepton they are compatible within uncertainties. The combined VH signal strength obtained by CMS was $\mu_{VH} = 1.0 \pm 0.5$ for an observed and

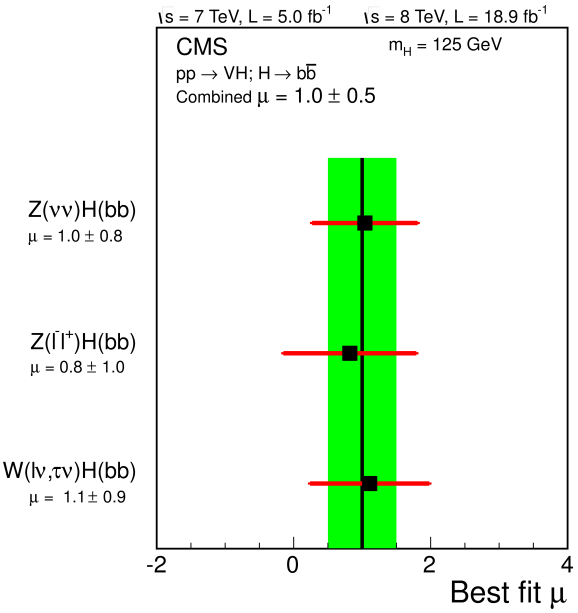


Figure 7.12: Strength of the signal in the 0, 1 and 2 lepton channels for the 7 and 8 TeV data set obtained by the CMS experiment. The signal strength of the combined VH signal is $\mu_{VH} = 1.0 \pm 0.5$. Taken from [91].

expected significance of 2.1σ, also in agreement with the ATLAS VH search.

The combined ATLAS and CMS analysis of the LHC pp collision data at 7 and 8 TeV [23] resulted in a $H \to b\bar{b}$ signal strength measurement of $\mu_{bb} = 0.70^{+0.29}_{-0.27}$, assuming that the Higgs production cross-section is the same as predicted by the SM. Conversely, assuming the $H \to b\bar{b}$ branching fraction to be the SM prediction, the $pp \to WH$ signal strength measured was $\mu_{WH} = 0.89^{+0.40}_{-0.38}$.

7.3.1 Impact of the Systematic Uncertainties

The impact of the data statistics and the uncertainties described in Section 7.1 on the final signal strength uncertainty is summarised in Table 7.13. The impact is determined for groups of uncertainties that are related with each other, for instance all systematics associated with the modelling and normalisation of the $t\bar{t}$ background are grouped into the set named $t\bar{t}$. The absolute impact of each set is evaluated by turning off all its elements during the fit, determining a new error on μ and quadratically subtracting it from the total uncertainty.

This evaluation was done for the combined VH analysis. As stated previously, the data statistics has a larger impact on the $\Delta\mu$ for the 0 and 2 lepton channels than all systematic uncertainties. The latter are dominated by the signal and background modelling. Amongst the modelling uncertainties, the W+jets background prediction is the most relevant to the total signal strength uncertainty, followed by the signal prediction itself, that is responsible for up to 9% of the signal measurement uncertainty.

The uncertainties associated with the multijet background modelling were estimated using the 1 lepton inputs from the ATLAS 8 TeV dataset and contribute only up to 2.4% to the total

Uncertainty	Absolute impact on $\Delta\mu$		Relative impact $\Delta\mu$ (%)	
Data Statistics	+0.37	-0.36	+53	-61
Full Systematics	+0.35	-0.29	+47	-40
Total	+0.51	-0.46	+100	-100
Modelling				
W+jets	+0.15	-0.16	+9	-12
Signal	+0.15	-0.06	+9	-1.7
$t\bar{t}$	+0.10	-0.08	+3.8	-3.0
Z+jets	+0.09	-0.09	+3.1	-3.8
Multijet	+0.06	-0.07	+1.6	-2.4
Single Top	+0.04	-0.03	+0.6	-0.4
Diboson	+0.02	-0.02	+0.2	-0.2
Total	+0.27	-0.23	+28	-25
Experimental				
Jets	+0.12	-0.10	+5.5	-4.7
$b-$Tagging	+0.11	-0.08	+4.6	-3.0
$E_{\mathrm{T}}^{\mathrm{miss}}$	+0.06	-0.05	+1.4	-1.1
Luminosity	+0.04	-0.02	+0.6	-0.2
Leptons	+0.02	-0.01	+0.2	-0.05
Total	+0.18	-0.14	+12	-9

Table 7.13: Absolute and relative (%) impact of each uncertainty component on the signal strength uncertainty $\Delta\mu$ for the combined VH statistical analysis. The relative impact is the square of fraction relative to the total uncertainty.

$\Delta\mu$ obtained from the combined VH fit.

The derived single top modelling uncertainties add only up to 0.6% to the signal strength uncertainty. Despite the importance of this background to the 1 lepton channel, the ZH search is less affected by it and therefore the total impact of its systematics to the combined analysis is small.

Between the experimental uncertainties, accounting to 12% of $\Delta\mu$ in total, $b-$tagging and jet energy scale and resolution are by far the most relevant sources. This is not only due to the complexity of the $b-$tagging algorithm and of the jet calibration chain, but also due to the importance that jets and $b-$tagging have in a $H \to b\bar{b}$ signature search.

7.3.2 Post-Fit distributions

Figures 7.13 and 7.14 show the post-fit distributions of the 1 lepton channel analysis for events with 2 and 3 jets, respectively. The main backgrounds are normalised including the scale factors determined during the fit, presented in Table 7.11. The uncertainty bands result from the fit constrain of each nuisance parameter. The W/Z+jets samples are redefined merging the $W/Z+bb/bc/bl/cc$ samples into a single W/Z+hf sample, with hf=bb, bc, bl, cc standing for

heavy flavour jets. From direct comparison with the pre-fit total background, also represented in the plots, it is visible how the agreement between the data and simulation prediction improves with the fit procedure.

7.3.3 Impact of adding the MVA input variables

In order to evaluate the real impact of the new input variables on the BDT performance, reported back in Section 6.5.3, including all the analysis uncertainties, an independent fit was performed for the following conditions:

- BDT + $\Delta Y(W,H)$
- BDT + m_{Wb_1}
- BDT + $\Delta Y(W,H) + m_{Wb_1}$ (both)

Since the aim is to isolate the effect of the new variables, the fit results are compared to the ones obtained with the LIP 1 lepton distributions with the default BDT. The most important parameter to measure the performance of the new variables is the expected significance. With it is possible to estimate which option has more sensitivity to do the signal measurement. All the other parameters, signal strength and observed significance, are the answer given by the data and must not dictate any decision on the analysis strategy.

The expected significance is shown, for the different fit cases, in Figure 7.15. The expected significance improves by up to 12% when both new variables are added to the BDT training, while performing the fit with one signal strength parameter using the 1 lepton channel alone. The impact on the expected significance of the combined analysis is non-negligible, up to 6%, even if the improvement contribution comes only from the 1 lepton channel.

Figures 7.16 and 7.17 show the signal strength obtained for the same cases. The signal strength increases when adding $\Delta Y(W,H)$ to the BDT training, specially for the WH and 1 lepton signals. This tendency affects the total VH signal, although more smoothly due to the contributions of the other channels. Since the modelling of the BDT output and input variables under test were carefully checked for the signal and main backgrounds, not revealing any MC mis-modelling of their spectra, this fluctuation is most likely due to the real data.

Nevertheless, all results are compatible with each other and with the standard model expectation of a $\mu = 1$ signal, if the large uncertainties are taken into account. It is interesting to note that, in average, the total uncertainty on the WH and 1 lepton signal strengths decrease when the new variables are added, and that it comes mostly from the systematic side. This can be due to the fact that the new BDTs are better to reduce the backgrounds. So, the uncertainties associated with the background prediction, a dominant source in this analysis, affect less the measurement of the signal, reducing the systematic uncertainty.

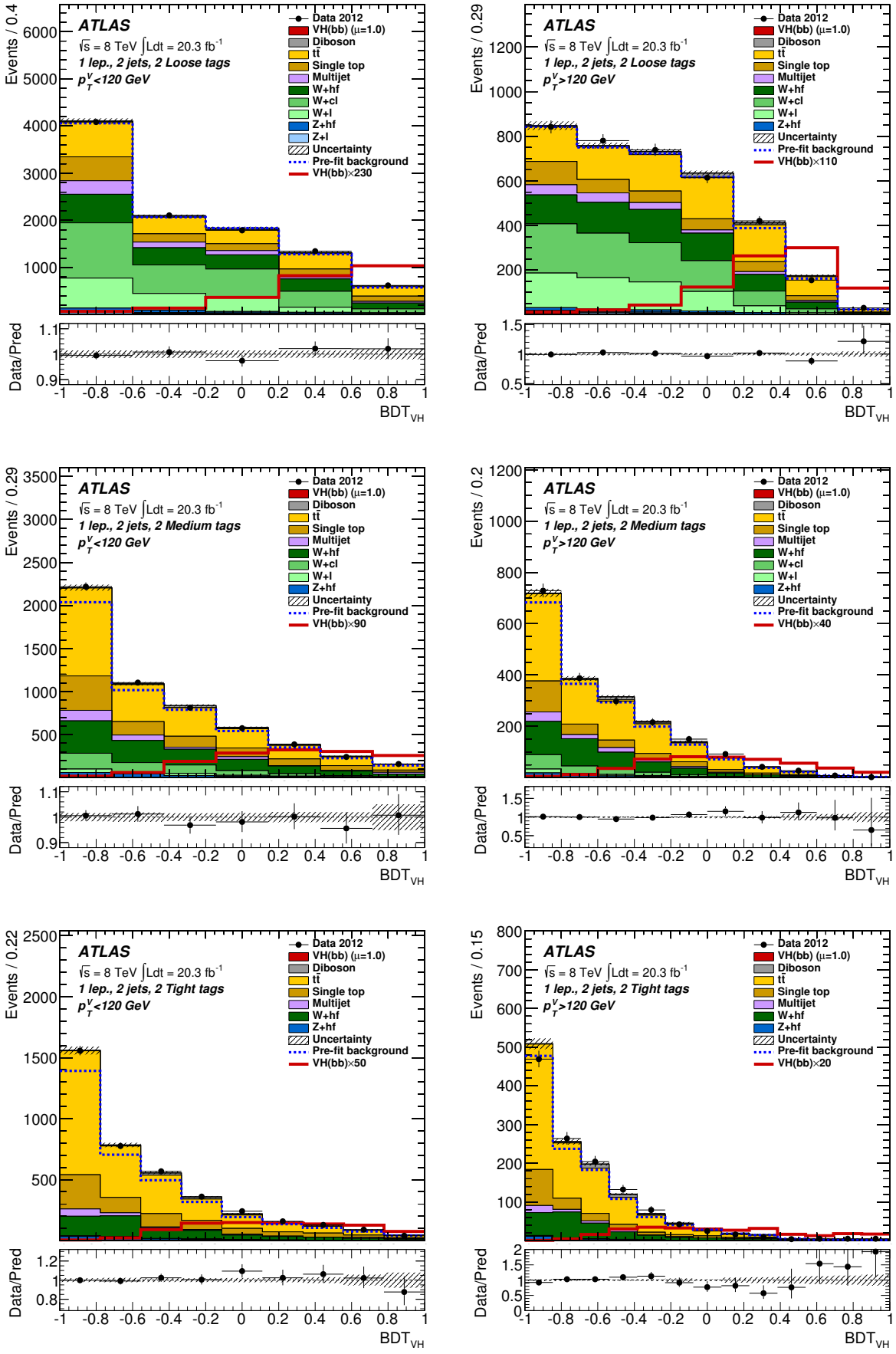


Figure 7.13: Post-fit distributions obtained from the VH statistical analysis for the 1 lepton channel and events with 2 jets.

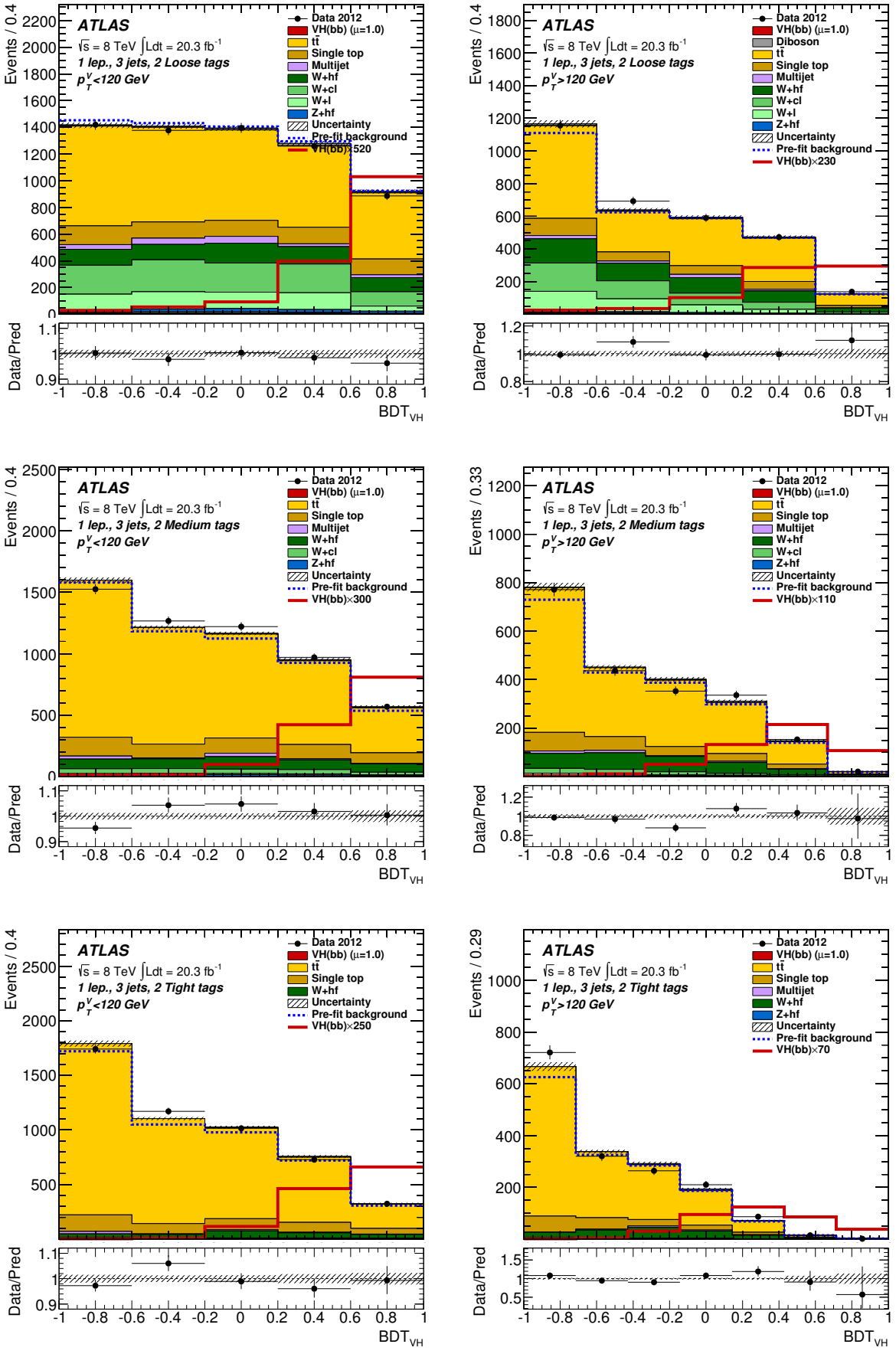


Figure 7.14: Post-fit distributions obtained from the VH statistical analysis for the 1 lepton channel and events with 3 jets.

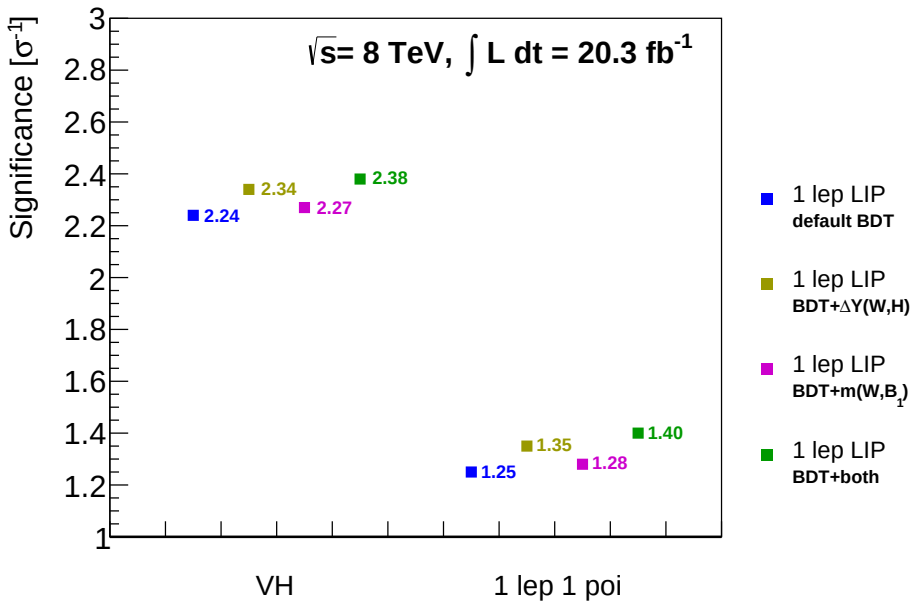


Figure 7.15: Expected significance for the VH combined fit and for the 1 lepton fit with 1 parameter of interest (poi). The results were obtained with the 1 lepton distributions from LIP for the default BDT and the BDT + $\Delta Y(W,H)$, + m_{Wb_1} and + $\Delta Y(W,H) + m_{Wb_1}$ (both).

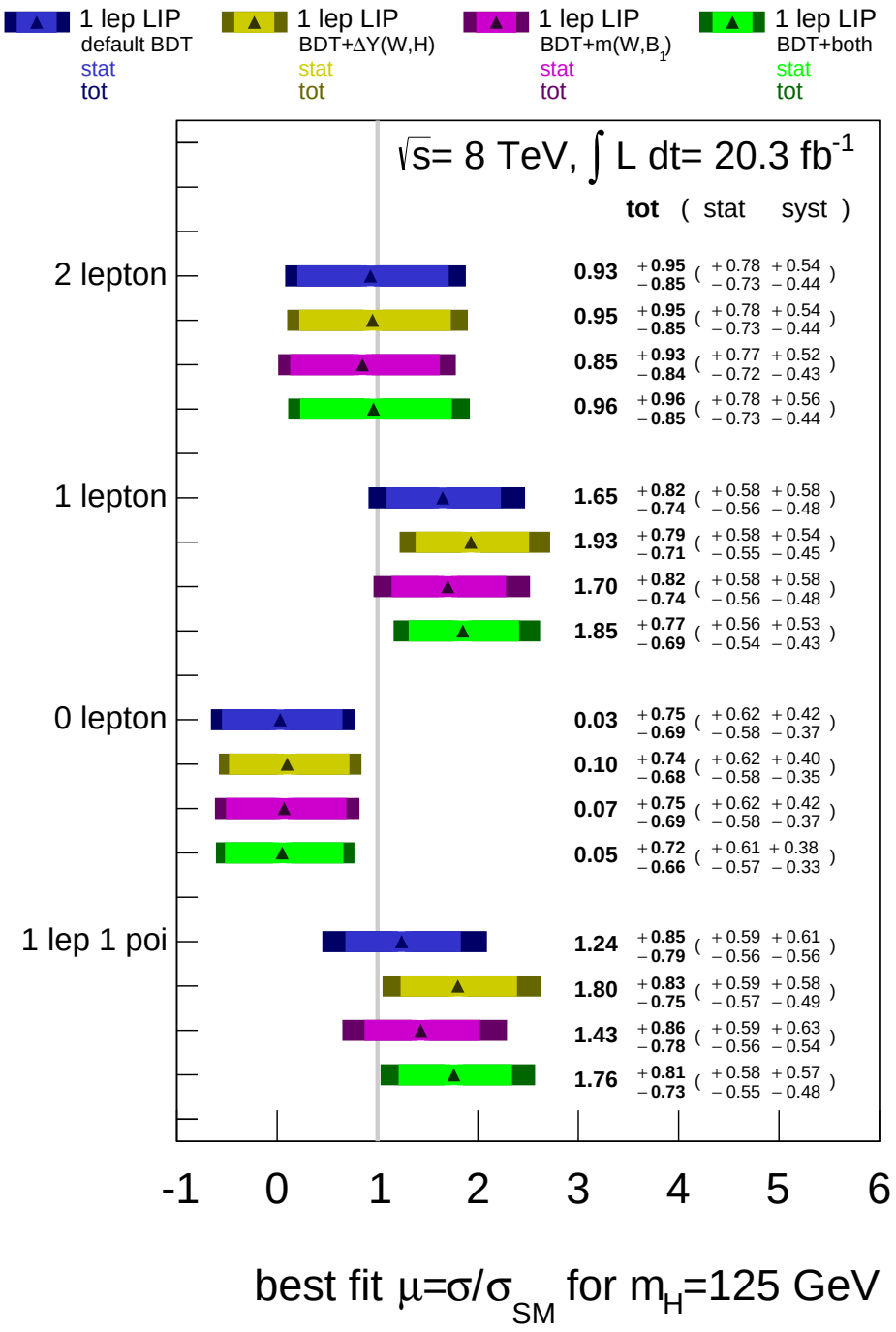


Figure 7.16: Signal strength for the signal in the 0, 1 and 2 lepton channels. The results obtained with the 1 lepton distributions from LIP are shown for the default BDT and the BDT + $\Delta Y(W,H)$, + m_{Wb_1} and + $\Delta Y(W,H) + m_{Wb_1}$ (both).

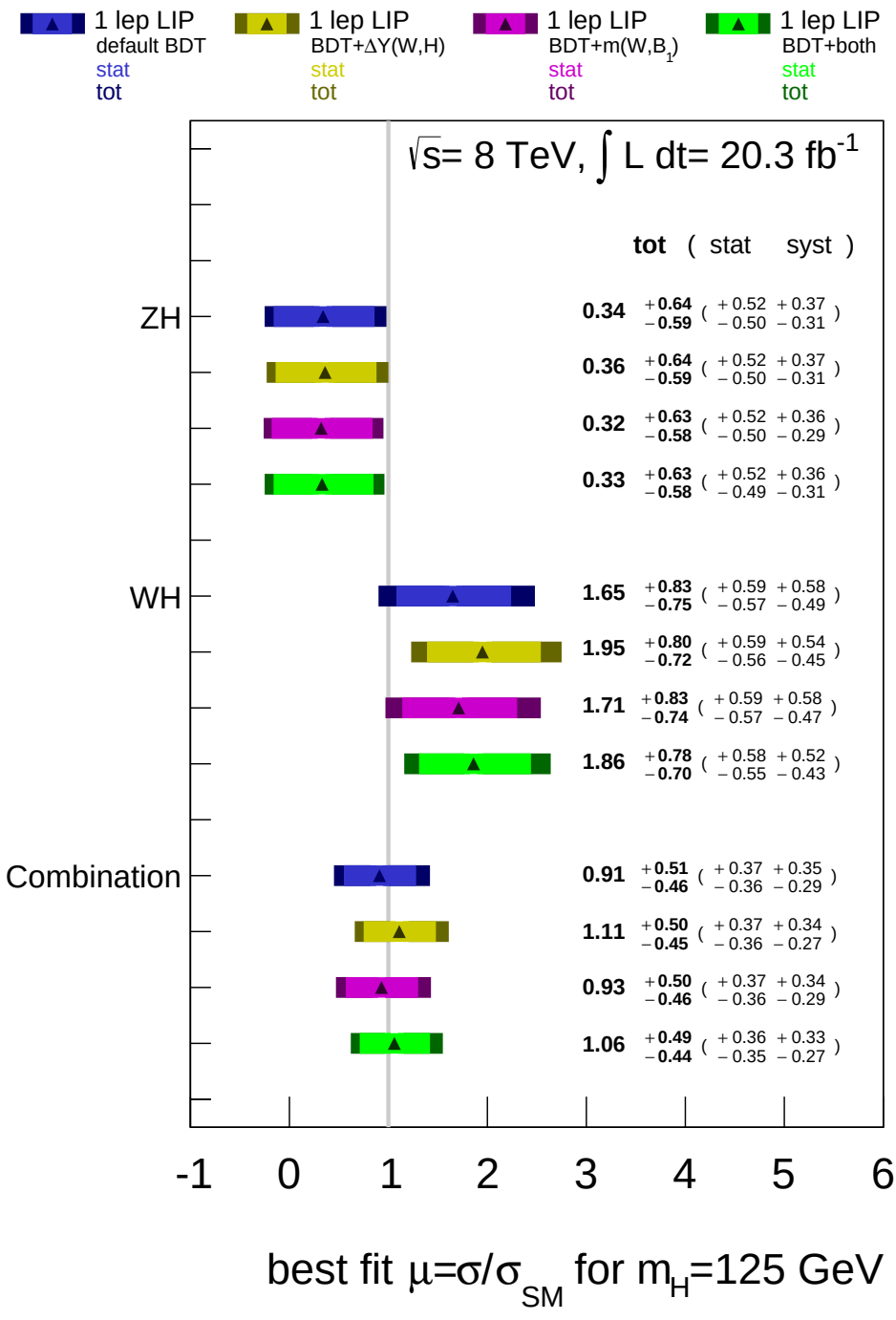


Figure 7.17: Signal strength for the WH, ZH and combined VH signals. The results obtained with the 1 lepton distributions from LIP are shown for the default BDT and the BDT + $\Delta Y(W,H)$, + m_{Wb_1} and + $\Delta Y(W,H) + m_{Wb_1}$ (both).

Chapter 8

Conclusions

This thesis described the search for the SM Higgs signal in the $WH \to \ell\nu b\bar{b}$ channel at the LHC/CERN with the detector ATLAS, using 20.3 fb^{-1}of pp collision data at a center-of-mass energy of $\sqrt{s} = 8$ TeV.

The analysis consisted of an event selection designed to efficiently select the final state topology of the signal process and reject as much as possible background-like events. It searched for events containing one isolated and high-energy electron or muon, missing transverse energy and two jets passing the $b-$tagging requirement, triggered by the single electron or muon trigger chain. $t\bar{t}$ and $W+$jets constitute the main background in a search where the signal significance $S/\sqrt{B}$ is as low as 0.3. The event selection was followed by a multivariate analysis technique that exploits the set of selected events and globally increases the sensitivity to the signal.

The signal and all the backgrounds to the analysis, with the exception of multiple jets processes, are simulated using the Monte-Carlo method. The simulation of the main backgrounds is validated with high-purity samples and normalised with data. This establishes a prediction of the background composition allowing to measure the signal parameters in data.

The full event selection software was independently coded and cross-checked against the outcome of the codes of other groups participating in the WH analysis. First, the physical objects in the event, such as jets, electrons and muons, are calibrated and categorised according to their reconstruction quality. This step was executed with a full agreement within the different groups, with small differences arising only for a maximum of 0.1% events from a 300 000 event reference sample containing the MC signal. Then, the event selection was validated in the same MC reference sample and in real data samples generated with the electron and muon triggers. A maximum deviation of 0.08% was found in the number of selected events.

The events passing the $WH \to \ell\nu b\bar{b}$ selection were analysed with a Boosted Decision Tree (BDT) multivariate technique. Previous studies have shown that the analysis benefits from a 30% improvement in sensitivity by doing so. This method employs a sequence of binary cuts on observables with the objective of classifying the events in two categories: signal or background. The BDTs are constructed from simulation and exploit the correlations between

the different variables of the input set to access regions of the phase space of larger signal purity. This method was chosen due to its performance, simplicity and straightforward interpretation.

A study to further enhance the BDT performance was carried out. It aimed to identify new BDT input observables to increase the signal sensitivity. About 20 new variables were tested by adding one at a time to the baseline input variable set and by evaluating their impact on the background versus signal efficiency curve of the output. Amongst the studied set of variables, two resulted in a significant improvement on the BDT capability to reject background: the rapidity difference between the Higgs and W boson candidates $\Delta Y(W,H)$ and the invariant mass of the W candidate and the $p_{\mathrm{T}}-$leading jet m_{Wb_1}. When added to the BDT, these two variables can increase the background rejection up to 10% for the same signal efficiency, when compared to the default set of variables. The correct modelling of these variables by MC is a necessary condition to the reliability of the BDT when applied to real data. This was carefully checked by comparing the spectra of the new variables for data and prediction. Special effort was put into making sure that the most important backgrounds of the analysis were well modelled by simulation. In that sense, data and MC were compared using samples enriched in the dominant backgrounds, namely $t\bar{t}$ and $W+$jets. The differences encountered were within the uncertainties of the analysis. For the signal process, an additional cross-check consisted of comparing different simulation models of $\Delta Y(W,H)$ and m_{Wb_1}. The predictions agreed within the systematic uncertainty associated with the signal modelling considered in the $WH \to \ell\nu b\bar{b}$ analysis.

Experimental uncertainties on the energy scale and resolution of jets, electrons and muons, and associated with the $b-$tagging procedure were taken into account. Furthermore, theory uncertainties related to the MC prediction of the background and signal processes were also considered. The impact of these systematic uncertainties was evaluated by re-running the full event selection chain, varying the parameter under evaluation to the up/down uncertainty limit. In total, a set of 65 experimental systematic uncertainties and 50 theoretical systematic uncertainties was considered. The implementation of the systematic variations on the selection code was verified, and the numbers obtained using the LIP analysis code differ in average 0.03% from the results of further groups.

The derivation of the uncertainties on the single top background simulation was part of the work of this thesis. The effect of higher-order corrections and of the uncertainties on the hadronisation, parton shower and underlying event models were investigated by checking the modelling of the most important observables to the WH search by different MC event generators. Single top modelling uncertainties resulted in a maximum uncertainty of 30%, 52% e 15% in the yields of the $s-$, $t-$ and $Wt-$channel production, respectively. Moreover, it was found that higher-order corrections and uncertainties on the hadronisation model impact the $m_{b\bar{b}}$ and $p_{\mathrm{T}}^{b_1}$ spectra shape systematically for the $Wt-$channel, and a systematic uncertainty was considered in the analysis to account for it.

The signal was measured through a maximum likelihood fit of prediction to data using the BDT output discriminant, where the normalisation of the main backgrounds are adjusted

by the fit to the data yield. Systematic and MC statistical uncertainties are incorporated into the likelihood function through nuisance parameters that model the impact of each source of uncertainty on the yields of the different physical processes, and are also constrained by data.

The signal strength obtained was $\mu_{WH} = 1.65^{+0.58}_{-0.56}(\text{stat})^{+0.58}_{-0.48}(\text{syst}) = 1.65^{+0.82}_{-0.74}$. The value is compatible with the SM prediction within the large uncertainties. From the experimental systematics side, the uncertainties related to the energy scale and resolution of jets and to $b-$tagging are the largest sources contributing to the uncertainty on the signal strength. Uncertainties on the simulation of the $W+$jets, of the signal and $t\bar{t}$ have more impact amongst the theoretical set.

The significance of the signal, representing the probability of the background-only hypothesis, corresponds to the gaussian probability of observing a value larger than 2.0σ, and is insufficient to claim the observation of the $WH \to \ell\nu b\bar{b}$ process. It is also insufficient to interpret the nominal $\mu > 1$ as an excess of signal relative to the SM prediction.

Additionally, the expected significance of the signal was measured to evaluate the effect of adding the new observables to the BDT on the signal sensitivity of the analysis. The same statistical procedure was performed using the different BDTs outputs. The usage of $\Delta Y(W,H)$ and m_{Wb_1} together in the BDT method improves the expected WH signal significance in 12%. For the combined VH search, including the ZH production channel, the improvement obtained was 6%, even if the new variables are used only in the WH BDT analysis.

The LHC is now colliding protons at a center-of-mass energy of $\sqrt{s} = 13$ TeV and ATLAS was able to record more than 80 fb^{-1}of data. The BDT performance was studied as a function of the same 20 new observables using 13 TeV MC samples and the same conclusions were reached. $\Delta Y(W,H)$ ended up by being included in the WH baseline MVA, as well as m_{top}, a variable similar to m_{Wb_1} proposed by a different study. Together, they increased the sensitivity of the search in 7%.

The search for the Higgs production associated with a W/Z boson in the $H \to b\bar{b}$ decay channel was already published for the first 36.1 fb^{-1}of integrated luminosity [92]. The results include the combination with the Run I data analysis and the measured signal strength was $\mu = 0.90 \pm 0.18(\text{stat})^{+0.21}_{-0.19}(\text{syst})$. The observed signal significance is 3.6σ to be compared with an expectation of 4.0σ, providing evidence for the Higgs decay to b-quarks and for its production in association with a vector boson. CMS reports a very similar analysis, yielding the same conclusions [93].

Appendices

Appendix A

Acronyms

2HDM Two Higgs Doublet Model

ATLAS A Toroidal LHC Apparatus

BDT Boosted Decision Tree

BR Branching Ratio

BSM Beyond the Standard Model

CERN European Organization for Nuclear Research

CIS Charge Injection System

CKM Cabibo-Kobayashi-Maskawa quark-mixture matrix

CMS Compact Muon Solenoid

CSC Cathode Strip Chamber

CTP Central Trigger Processor

DAQ Data Acquisition

DCS Detector Control System

DQ Data Quality

EF Event Filter

EM Electromagnetic

EMB Electromagnetic Calorimeter barrel

EMEC Electromagnetic Calorimeter End-Cap

FCal Forward Calorimeters

FSR Final State Radiation

GRL Good Runs List

GSC Global Sequential Calibration

HEC Hadronic End-Cap

HEP High Energy Physics

ID Inner Detector

IP Interaction Point

ISR Initial State Radiation

JES Jet Energy Scale

JVF Jet Vertex Fraction

LEP Large Electron-Positron collider

LH Likelihood

LHC Large Hadron collider

LLR Log-Likelihood Ratio

LO Leading Order

MC Monte-Carlo

MDT Monitored Drift Tube

MVA Multivariate Analysis

MS Muon Spectrometer

NLO Next-to-Leading Order

NN Neural Network

PCB Printable Circuit Board (PCB)

PDF Parton Density Function

PMT Photo-Multiplier Tube

PS Proton Synchrotron

PSB Proton Synchrotron Booster

PV Primary Vertex

QCD Quantum Chromodynamics

QED Quantum Electrodynamics

QFT Quantum Field Theory

RF Radio-Frequency

ROD Readout Driver

RoI Regions of Interest

ROS Readout System

RPC Resistive Plate Chamber

ROC Receiver Operating Characteristic

SM Standard Model

SUSY Super Symmetry

SV Secondary Vertex

SVM Support Vector Machines

TGC Thin Gap Chamber

TileCal Tile Calorimeter

TRT Transition Radiation Tracker

vev Vacuum Expectation Value

Appendix B

Monte-Carlo Samples

Tables B.1, B.2 and B.3 list the complete set of nominal Monte-Carlo samples used in the $WH \to \ell\nu b\bar{b}$ analysis. The statistics of the samples and their cross-sections are presented.

The top background events, except for the single top production in the $Wt-$channel, are generated with a charged lepton filter, imposing that at least one charged lepton exists in the event final state. The WZ and ZZ samples are split according to the decay channel of the electroweak gauge bosons. The fully hadronic final state sample is discarded from analysis, given the diminished selection efficiency and cross section of the WZ and ZZ process.

The $W/Z+$jets samples are separated according to the decay channel of the W/Z boson and to the flavour of the jets. Events with $b-$jets are obtained with a b filter, $c-$jets with c filter plus b veto and light jets have b and c veto. In order to increase statistics, the $W/Z+$jets are generated in the following slices of the $p_{\mathrm{T}}^{W/Z}$ spectrum: $[0,\infty[$, $[40,70[$, $[70,140[$, $[140,280[$, $[280,500[$, $[500,\infty[$ GeV. As the first sample is inclusive in the vector boson p_{T}, the overlap with the remaining samples is resolved at the event selection level, by removing events for which the truth $p_{\mathrm{T}}^{W/Z}$ is above 40 GeV.

Table B.4 contains the information about the POWHEG+PYTHIA8 samples used as an alternative model for the WH signal prediction, used for the signal modelling checks.

MC ID	Process	N_{events}	$\sigma\times$BR [pb]
	$p_T^W > 0$ GeV		
167740	$\mathrm{W}(\to e\nu)+b$	14997980	154.4
167741	$\mathrm{W}(\to e\nu)+c$	9998989	533.9
167742	$\mathrm{W}(\to e\nu)+l$	49885967	11363
167743	$\mathrm{W}(\to \mu\nu)+b$	14989485	154.4
167744	$\mathrm{W}(\to \mu\nu)+c$	9992484	533.9
167745	$\mathrm{W}(\to \mu\nu)+l$	49846965	11363
167746	$\mathrm{W}(\to \tau\nu)+b$	14925982	154.4
167747	$\mathrm{W}(\to \tau\nu)+c$	9993984	533.9
167748	$\mathrm{W}(\to \tau\nu)+l$	49920968	11363
	$40 \leq p_T^W < 70$ GeV		

180534	$W(\to e\nu)+b$	2999998	24.81
180535	$W(\to e\nu)+c$	4499994	121.3
180536	$W(\to e\nu)+l$	16997491	572.1
180537	$W(\to \mu\nu)+b$	2996996	24.81
180538	$W(\to \mu\nu)+c$	4498998	121.3
180539	$W(\to \mu\nu)+l$	16988984	572.1
180540	$W(\to \tau\nu)+b$	2998997	24.81
180541	$W(\to \tau\nu)+c$	4498999	121.3
180542	$W(\to \tau\nu)+l$	16996492	572.1
	$70 \le p_T^W < 140$ GeV		
167761	$W(\to e\nu)+b$	2000000	12.66
167762	$W(\to e\nu)+c$	2996497	54.67
167763	$W(\to e\nu)+l$	4998998	208.3
167764	$W(\to \mu\nu)+b$	1998999	12.66
167765	$W(\to \mu\nu)+c$	2995999	54.67
167766	$W(\to \mu\nu)+l$	4998992	208.3
167767	$W(\to \tau\nu)+b$ (FS)	1999893	12.66
167768	$W(\to \tau\nu)+c$ (FS)	2999890	54.67
167769	$W(\to \tau\nu)+l$ (FS)	4999786	208.3
	$140 \le p_T^W < 280$ GeV		
167770	$W(\to e\nu)+b$	4998995	2.165
167771	$W(\to e\nu)+c$	1999997	7.526
167772	$W(\to e\nu)+l$	2000000	24.57
167773	$W(\to \mu\nu)+b$	4983993	2.165
167774	$W(\to \mu\nu)+c$	1995998	7.526
167775	$W(\to \mu\nu)+l$	1993999	24.57
167776	$W(\to \tau\nu)+b$	3998996	2.165
167777	$W(\to \tau\nu)+c$ (FS)	1998688	7.526
167778	$W(\to \tau\nu)+l$ (FS)	1999994	24.57
	$280 \le p_T^W < 500$ GeV		
167779	$W(\to e\nu)+b$	899999	0.1680
167780	$W(\to e\nu)+c$ (FS)	199898	0.4690
167781	$W(\to e\nu)+l$ (FS)	499891	1.386
167782	$W(\to \mu\nu)+b$	898000	0.1680
167783	$W(\to \mu\nu)+c$ (FS)	199998	0.4690
167784	$W(\to \mu\nu)+l$ (FS)	499698	1.386
167785	$W(\to \tau\nu)+b$	898999	0.1680
167786	$W(\to \tau\nu)+c$ (FS)	199998	0.4690
167787	$W(\to \tau\nu)+l$ (FS)	499998	1.386

$p_T^W \geq 500$ GeV			
167788	$W(\rightarrow e\nu)+b$	100000	1.115×10^{-2}
167789	$W(\rightarrow e\nu)+c$	10000	2.698×10^{-2}
167790	$W(\rightarrow e\nu)+l$	10000	7.360×10^{-2}
167791	$W(\rightarrow \mu\nu)+b$	90000	1.115×10^{-2}
167792	$W(\rightarrow \mu\nu)+c$ (FS)	10000	2.698×10^{-2}
167793	$W(\rightarrow \mu\nu)+l$ (FS)	49700	7.360×10^{-2}
167794	$W(\rightarrow \tau\nu)+b$	90000	1.115×10^{-2}
167795	$W(\rightarrow \tau\nu)+c$ (FS)	10000	2.698×10^{-2}
167796	$W(\rightarrow \tau\nu)+l$ (FS)	49998	7.360×10^{-2}

Table B.2: Monte-Carlo samples, statistics and cross-section of the W+jets processes. 'FS' stands for Full simulation of the ATLAS detector. The default case corresponds to usage of the ATLAS Fast simulation scheme.

MC ID	Process	N_{events}	$\sigma\times$BR [pb]
$p_T^Z > 0$ GeV			
167749	$Z(\rightarrow ee)+b$	3999000	34.75
167750	$Z(\rightarrow ee)+c$	2999995	352.3
167751	$Z(\rightarrow ee)+l$	4978999	853.9
167752	$Z(\rightarrow \mu\mu)+b$	3997997	34.75
167753	$Z(\rightarrow \mu\mu)+c$	2937995	352.3
167754	$Z(\rightarrow \mu\mu)+l$	4993999	853.9
167755	$Z(\rightarrow \tau\tau)+b$	3997994	34.75
167756	$Z(\rightarrow \tau\tau)+c$	2998998	352.3
167757	$Z(\rightarrow \tau\tau)+l$	4989999	853.9
167758	$Z(\rightarrow \nu\nu)+b$	24932972	197.1
167759	$Z(\rightarrow \nu\nu)+c$	19997479	1879
167760	$Z(\rightarrow \nu\nu)+l$	24919979	4634
$40 \leq p_T^Z < 70$ GeV			
180543	$Z(\rightarrow ee)+b$	1199999	5.581
180544	$Z(\rightarrow ee)+c$	600000	26.98
180545	$Z(\rightarrow ee)+l$	1399998	46.40
180546	$Z(\rightarrow \mu\mu)+b$	1199000	5.581
180547	$Z(\rightarrow \mu\mu)+c$	599000	26.98
180548	$Z(\rightarrow \mu\mu)+l$	1398999	46.40
180549	$Z(\rightarrow \tau\tau)+b$	1198999	5.581
180550	$Z(\rightarrow \tau\tau)+c$	600000	26.98
180551	$Z(\rightarrow \tau\tau)+l$	1399996	46.40
$70 \leq p_T^Z < 140$ GeV			

167797	$Z(\to ee)+b$	1366999	2.727
167798	$Z(\to ee)+c$	999999	11.72
167799	$Z(\to ee)+l$	1999998	18.58
167800	$Z(\to \mu\mu)+b$	1394999	2.727
167801	$Z(\to \mu\mu)+c$	1000000	11.72
167802	$Z(\to \mu\mu)+l$	1996998	18.58
167803	$Z(\to \tau\tau)+b$ (FS)	1399396	2.727
167804	$Z(\to \tau\tau)+c$ (FS)	999998	11.72
167805	$Z(\to \tau\tau)+l$ (FS)	1969693	18.58
167806	$Z(\to \nu\nu)+b$ (FS)	5998993	15.69
167807	$Z(\to \nu\nu)+c$ (FS)	2998998	65.71
167808	$Z(\to \nu\nu)+l$ (FS)	4999996	105.2
	$140 \le p_T^Z < 280$ GeV		
167809	$Z(\to ee)+b$	999999	0.4262
167810	$Z(\to ee)+c$	399999	1.650
167811	$Z(\to ee)+l$	600000	2.384
167812	$Z(\to \mu\mu)+b$	987999	0.4262
167813	$Z(\to \mu\mu)+c$	399000	1.650
167814	$Z(\to \mu\mu)+l$	599500	2.384
167815	$Z(\to \tau\tau)+b$	798998	0.4262
167816	$Z(\to \tau\tau)+c$ (FS)	399999	1.650
167817	$Z(\to \tau\tau)+l$ (FS)	598897	2.384
167818	$Z(\to \nu\nu)+b$	4999995	2.442
167819	$Z(\to \nu\nu)+c$ (FS)	1999998	9.278
167820	$Z(\to \nu\nu)+l$ (FS)	2999999	13.48
	$280 \le p_T^Z < 500$ GeV		
167821	$Z(\to ee)+b$	180000	0.02903
167822	$Z(\to ee)+c$ (FS)	49899	0.1043
167823	$Z(\to ee)+l$ (FS)	49999	0.1371
167824	$Z(\to \mu\mu)+b$	175000	0.02903
167825	$Z(\to \mu\mu)+c$ (FS)	50000	0.1043
167826	$Z(\to \mu\mu)+l$ (FS)	50000	0.1371
167827	$Z(\to \tau\tau)+b$	180000	0.02903
167828	$Z(\to \tau\tau)+c$ (FS)	50000	0.1043
167829	$Z(\to \tau\tau)+l$ (FS)	49899	0.1371
167830	$Z(\to \nu\nu)+b$	1799997	0.1644
167831	$Z(\to \nu\nu)+c$ (FS)	249999	0.5830
167832	$Z(\to \nu\nu)+l$ (FS)	999892	0.7676
	$p_T^Z \ge 500$ GeV		

167833	$Z(\to ee)+b$	90000	1.721×10^{-3}
167834	$Z(\to ee)+c$ (FS)	10000	5.960×10^{-3}
167835	$Z(\to ee)+l$ (FS)	50000	7.228×10^{-3}
167836	$Z(\to \mu\mu)+b$	100000	1.721×10^{-3}
167837	$Z(\to \mu\mu)+c$	10000	5.960×10^{-3}
167838	$Z(\to \mu\mu)+l$	10000	7.228×10^{-3}
167839	$Z(\to \tau\tau)+b$	90000	1.721×10^{-3}
167840	$Z(\to \tau\tau)+c$ (FS)	10000	5.960×10^{-3}
167841	$Z(\to \tau\tau)+l$ (FS)	149900	7.228×10^{-3}
167842	$Z(\to \nu\nu)+b$	450000	9.600×10^{-3}
167843	$Z(\to \nu\nu)+c$ (FS)	50000	3.256×10^{-2}
167844	$Z(\to \nu\nu)+l$ (FS)	199699	3.975×10^{-2}

Table B.3: Monte-Carlo samples, statistics and cross-section of the Z+jets processes. 'FS' stands for Full simulation of the ATLAS detector. The default case corresponds to usage of the ATLAS Fast simulation scheme.

MC ID	Process	N_{events}	$\sigma\times$BR [pb]
	VH signal with Higgs mass of 125 GeV		
161805	qqWH($\to \ell\nu b\bar{b}$)	2988997	0.1317
161827	qqZH($\to \ell\ell b\bar{b}$)	2998998	0.02231
189340	ggZH($\to ee b\bar{b}$)	100000	6.300×10^{-4}
189341	ggZH($\to \mu\mu b\bar{b}$)	100000	6.304×10^{-4}
189342	ggZH($\to \tau\tau b\bar{b}$)	100000	6.312×10^{-4}
	Top backgrounds		
117050	$t\bar{t}$ (lepton filter)	99930891	137.3
110101	top $t-$channel (lepton filter)	8996990	28.43
110119	top $s-$channel (lepton filter)	5995993	1.818
110140	top $Wt-$channel	19937980	22.37
	Diboson backgrounds		
181966	ZZ($\to \ell\ell hh$)	3999995	1.207
181967	ZZ($\to \nu\nu hh$)	11518492	2.081
181968	WZ($\to hh\ell\ell$)	1500000	1.594
181969	WZ($\to hh\nu\nu$)	2999997	2.777
181970	WZ($\to \ell\nu hh$)	9999988	4.870
181971	WW	9999994	52.44

Table B.1: Monte-Carlo samples, statistics and cross-section of the signal, top background and dibosons processes. The $t\bar{t}$ and single top event generation have a charged lepton filter, except for the single top production in the $Wt-$channel.

MC ID	Process	N_{events}	$\sigma\times$BR [pb]
	WH signal with Higgs mass of 125 GeV, POWHEG+PYTHIA8		
189420	$qq \to W^+H \to e\nu b\bar{b}$	650000	0.028
189421	$qq \to W^+H \to \mu\nu b\bar{b}$	650000	0.028
189422	$qq \to W^+H \to \tau\nu b\bar{b}$	650000	0.028
189423	$qq \to W^-H \to e\nu b\bar{b}$	350000	0.016
189424	$qq \to W^-H \to \mu\nu b\bar{b}$	350000	0.016
189425	$qq \to W^-H \to \tau\nu b\bar{b}$	350000	0.016

Table B.4: Monte-Carlo ID, statistics and cross-section of the signal alternative samples generated with POWHEG+PYTHIA8 model.

Appendix C

MVA Input Variable Distributions

The distributions of the BDT input variables used in the 1 lepton MVA analysis are exhibited for data and prediction for all the analysis categories:

Figure C.1 shows the event category of 2 signal jets, 1 $b-$tagged and $p_{\mathrm{T}}^{W} < 120$ GeV.
Figure C.2 shows the event category of 2 signal jets, 1 $b-$tagged and $p_{\mathrm{T}}^{W} > 120$ GeV.
Figure C.3 shows the event category of 2 signal jets, both $b-$tagged and $p_{\mathrm{T}}^{W} < 120$ GeV.
Figure C.4 shows the event category of 2 signal jets, both $b-$tagged and $p_{\mathrm{T}}^{W} > 120$ GeV.
Figure C.5 shows the event category of 2 signal jets, LL tags and $p_{\mathrm{T}}^{W} < 120$ GeV.
Figure C.6 shows the event category of 2 signal jets, LL tags and $p_{\mathrm{T}}^{W} > 120$ GeV.
Figure C.7 shows the event category of 2 signal jets, MM tags and $p_{\mathrm{T}}^{W} < 120$ GeV.
Figure C.8 shows the event category of 2 signal jets, MM tags and $p_{\mathrm{T}}^{W} > 120$ GeV.
Figure C.9 shows the event category of 2 signal jets, TT tags and $p_{\mathrm{T}}^{W} < 120$ GeV.
Figure C.10 shows the event category of 2 signal jets, TT tags and $p_{\mathrm{T}}^{W} > 120$ GeV.
Figure C.11 shows the event category of 3 signal jets, 1 $b-$tagged and $p_{\mathrm{T}}^{W} < 120$ GeV.
Figure C.12 shows the event category of 3 signal jets, 1 $b-$tagged and $p_{\mathrm{T}}^{W} > 120$ GeV.
Figure C.13 shows the event category of 3 signal jets, 2 $b-$tags and $p_{\mathrm{T}}^{W} < 120$ GeV.
Figure C.14 shows the event category of 3 signal jets, 2 $b-$tags and $p_{\mathrm{T}}^{W} > 120$ GeV.
Figure C.15 shows the event category of 3 signal jets, LL tags and $p_{\mathrm{T}}^{W} < 120$ GeV.
Figure C.16 shows the event category of 3 signal jets, LL tags and $p_{\mathrm{T}}^{W} > 120$ GeV.
Figure C.17 shows the event category of 3 signal jets, MM tags and $p_{\mathrm{T}}^{W} < 120$ GeV.
Figure C.18 shows the event category of 3 signal jets, MM tags and $p_{\mathrm{T}}^{W} > 120$ GeV.
Figure C.19 shows the event category of 3 signal jets, TT tags and $p_{\mathrm{T}}^{W} < 120$ GeV.
Figure C.20 shows the event category of 3 signal jets, TT tags and $p_{\mathrm{T}}^{W} > 120$ GeV.

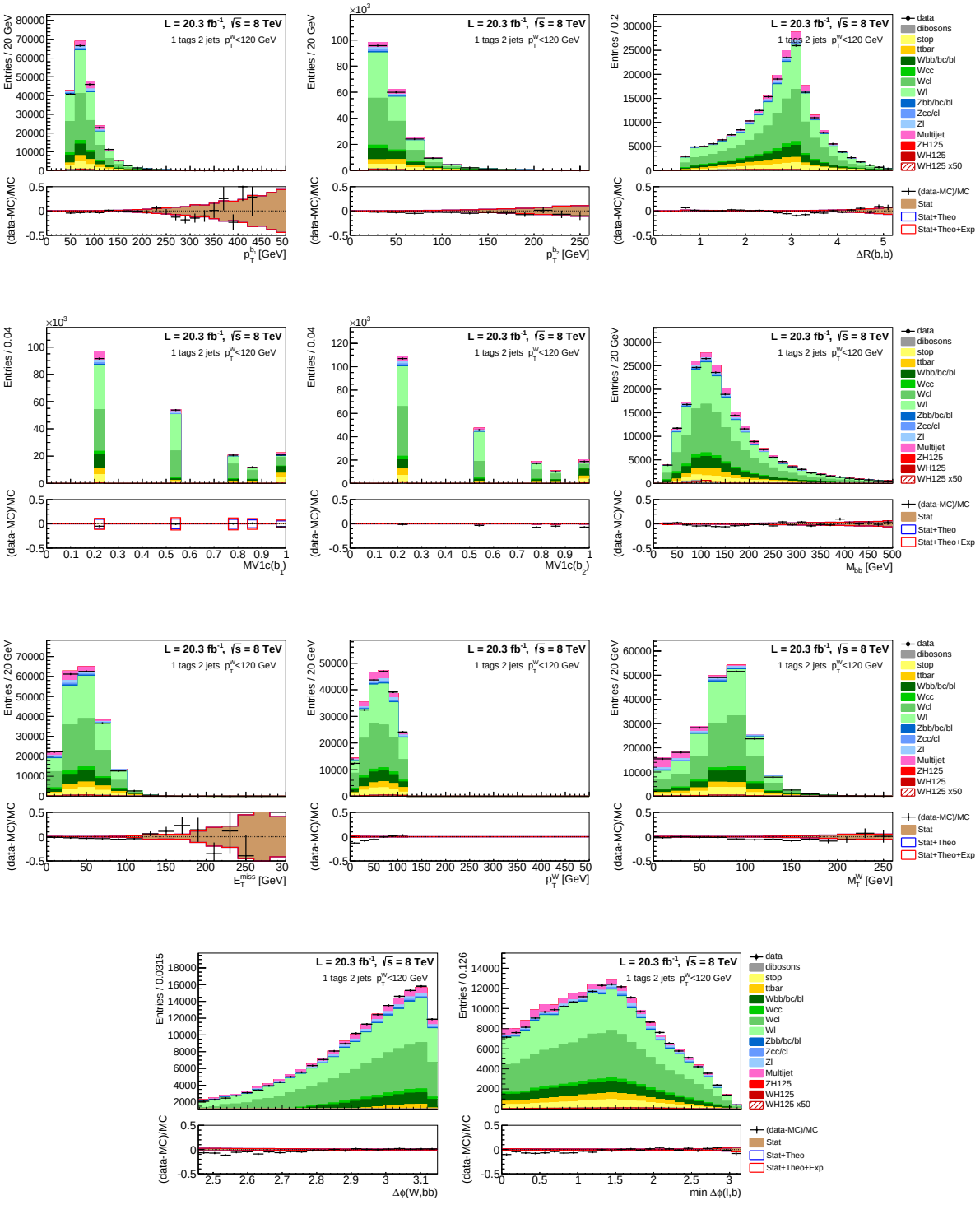


Figure C.1: Distribution of the BDT training variables for data and prediction for events with 2 signal jets, 1 $b-$tagged and $p_{\mathrm{T}}^{W} < 120$ GeV. The uncertainty bands are exhibited in the bottom plot as detailed in Section 7.1.

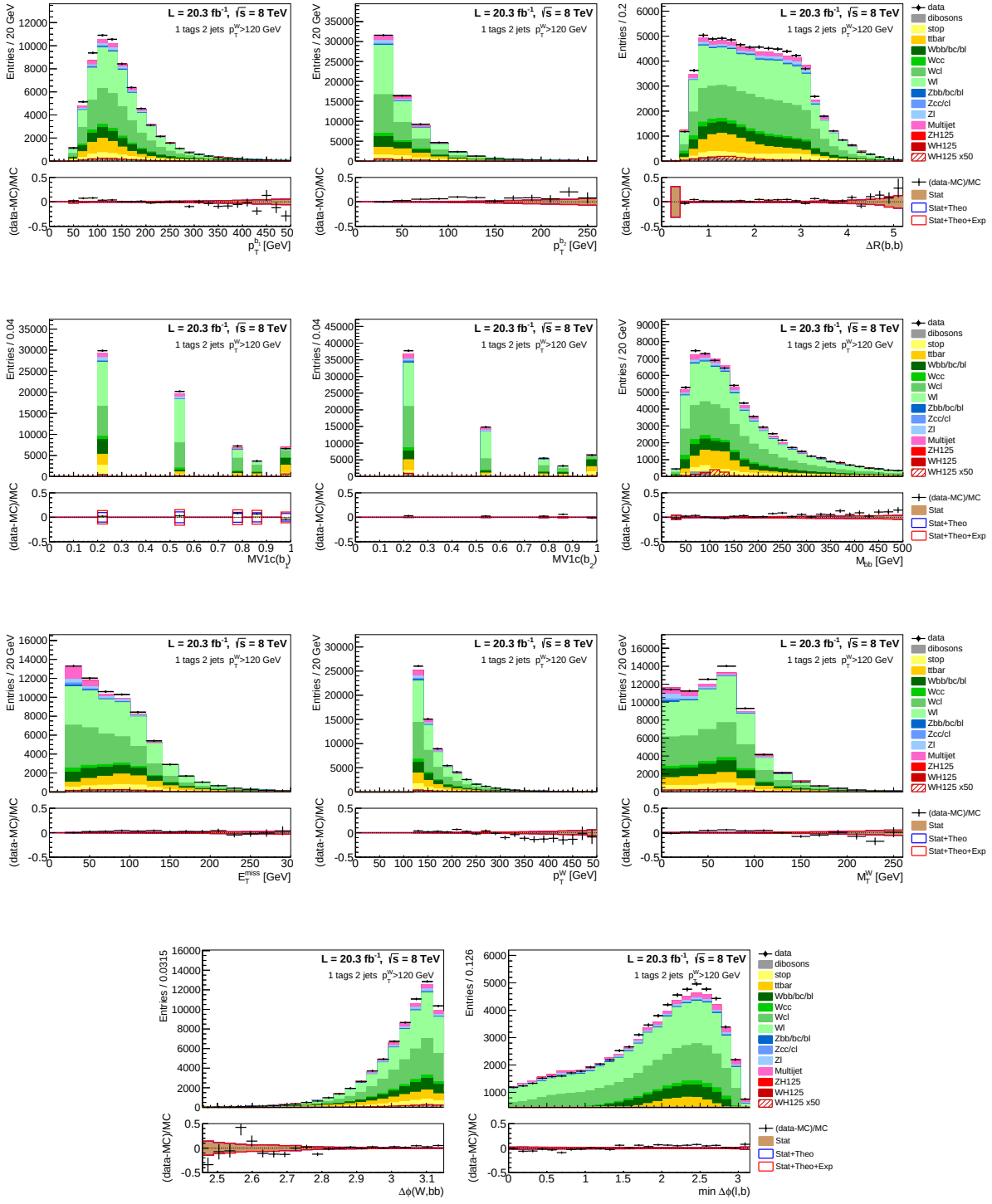


Figure C.2: Distribution of the BDT training variables for data and prediction for events with 2 signal jets, 1 $b-$tagged and $p_{\mathrm{T}}^{W} > 120$ GeV. The uncertainty bands are exhibited in the bottom plot as detailed in Section 7.1.

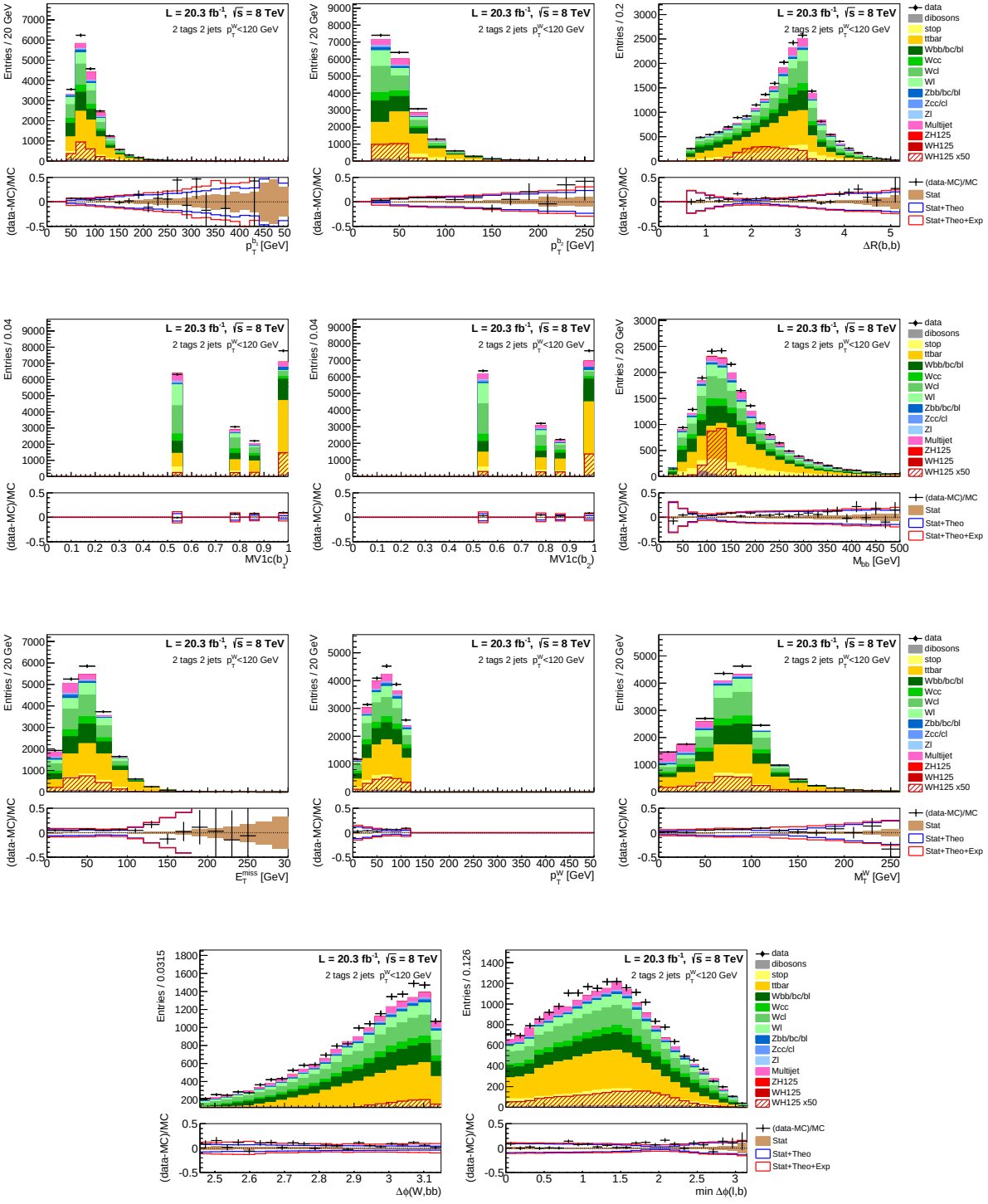


Figure C.3: Distribution of the BDT training variables for data and prediction for events with 2 signal jets, both $b-$tagged, and $p_{\mathrm{T}}^{W} < 120$ GeV. The uncertainty bands are exhibited in the bottom plot as detailed in Section 7.1.

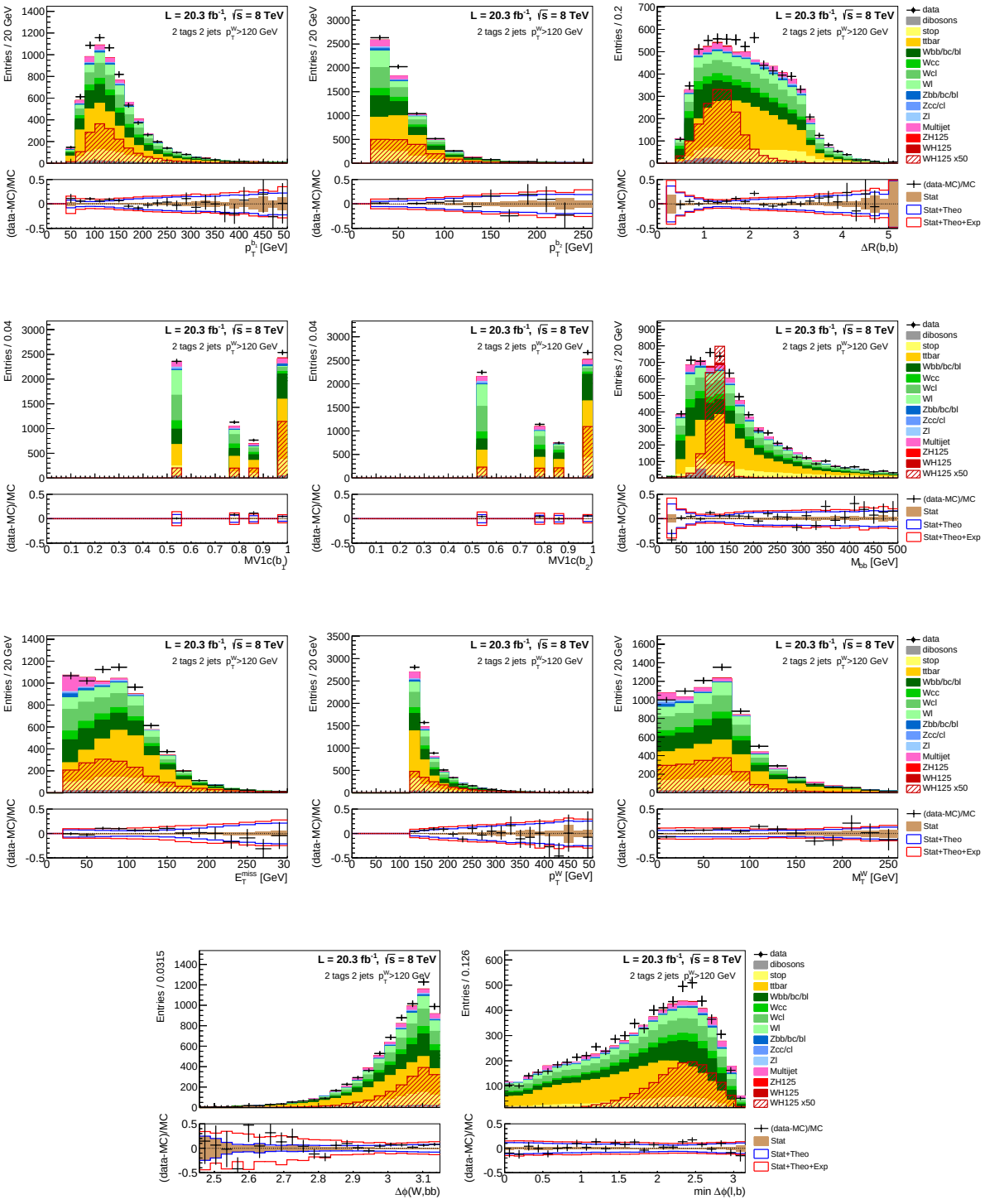


Figure C.4: Distribution of the BDT training variables for data and prediction for events with 2 signal jets, both $b-$tagged, and $p_{\mathrm{T}}^{W} > 120$ GeV. The uncertainty bands are exhibited in the bottom plot as detailed in Section 7.1.

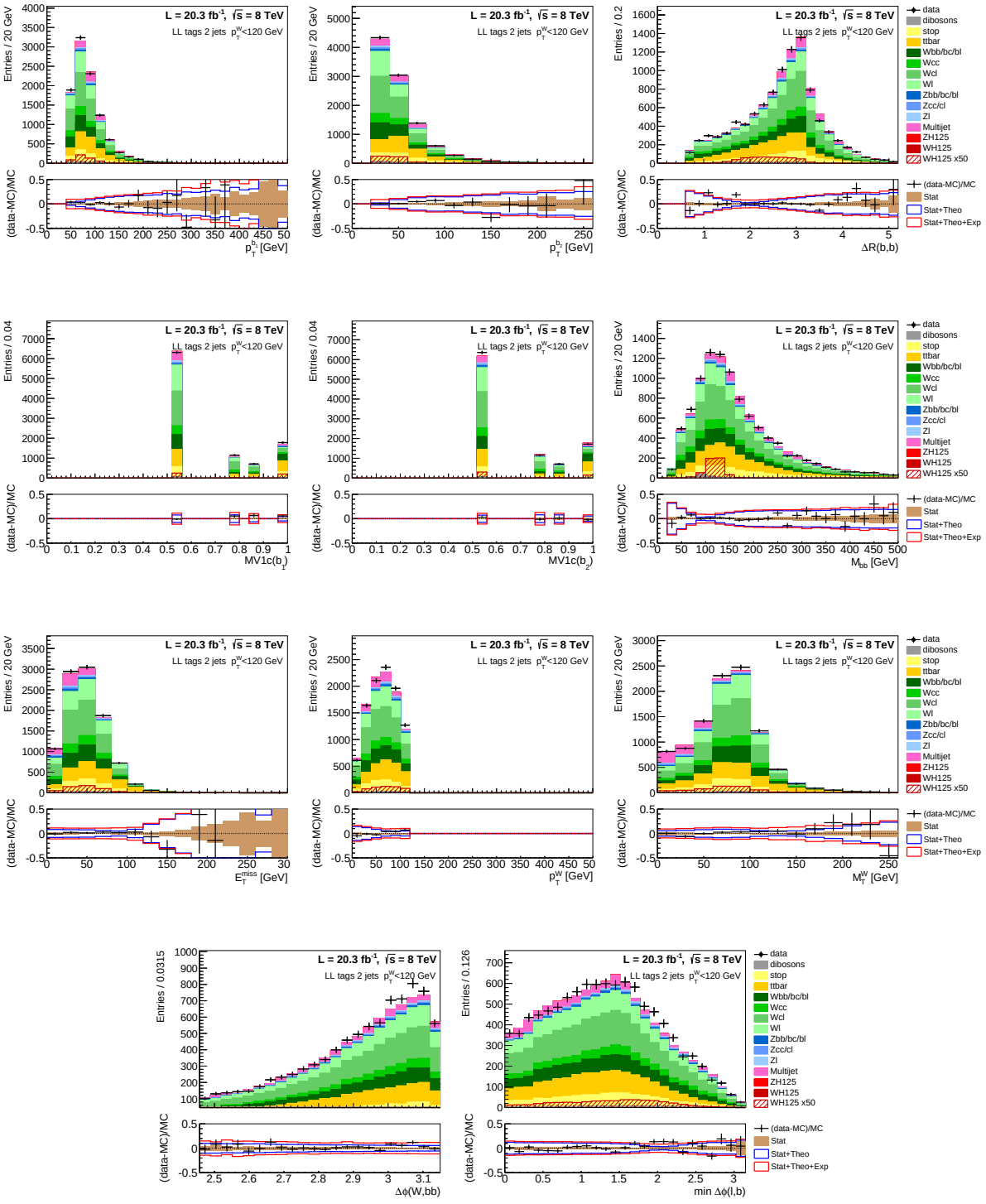


Figure C.5: Distribution of the BDT training variables for data and prediction for events with 2 signal jets, 2 LL $b-$tags and $p_{\mathrm{T}}^{W} < 120$ GeV. The uncertainty bands are exhibited in the bottom plot as detailed in Section 7.1.

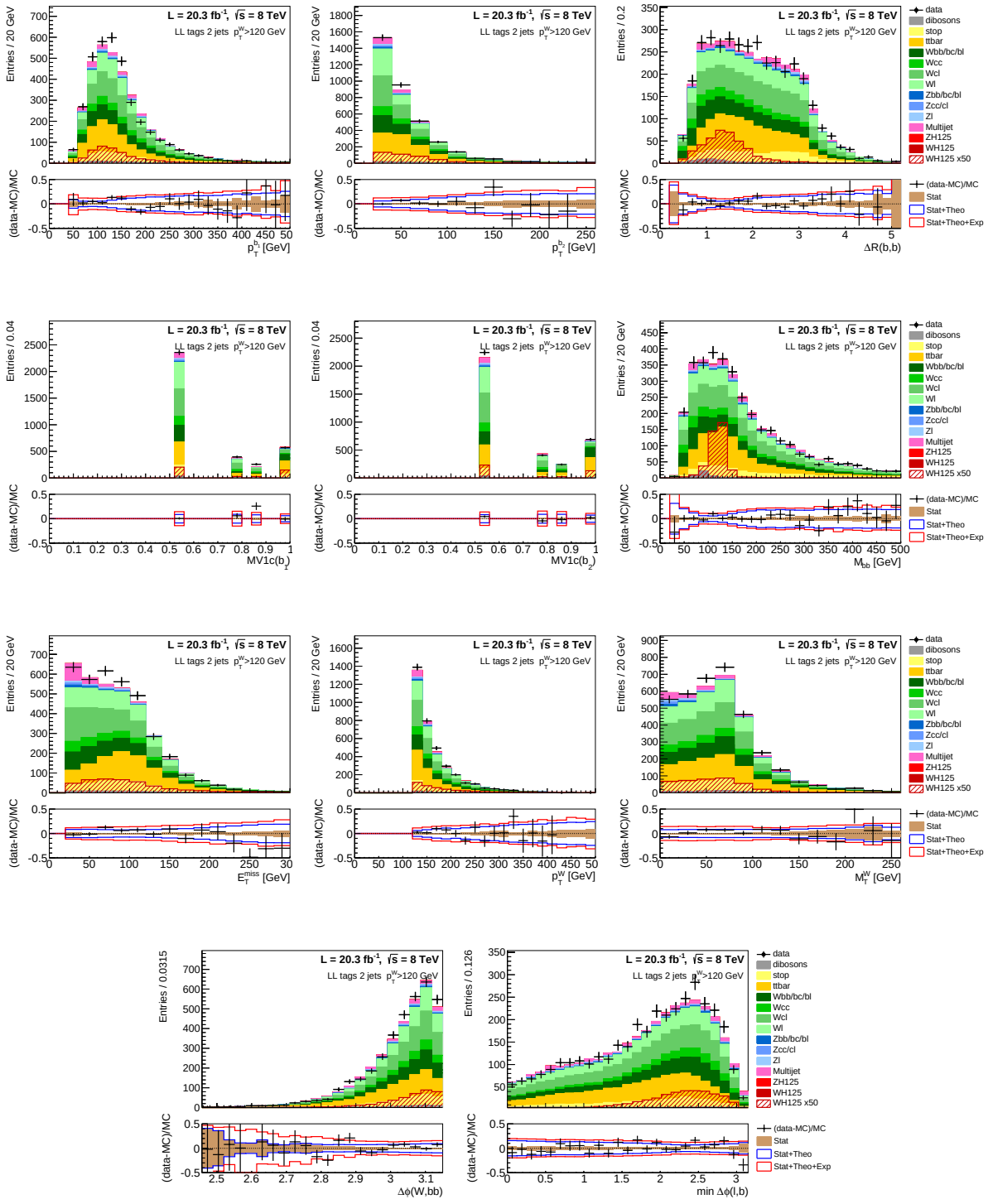


Figure C.6: Distribution of the BDT training variables for data and prediction for events with 2 signal jets, 2 LL $b-$tags and $p_{\mathrm{T}}^{W} > 120$ GeV. The uncertainty bands are exhibited in the bottom plot as detailed in Section 7.1.

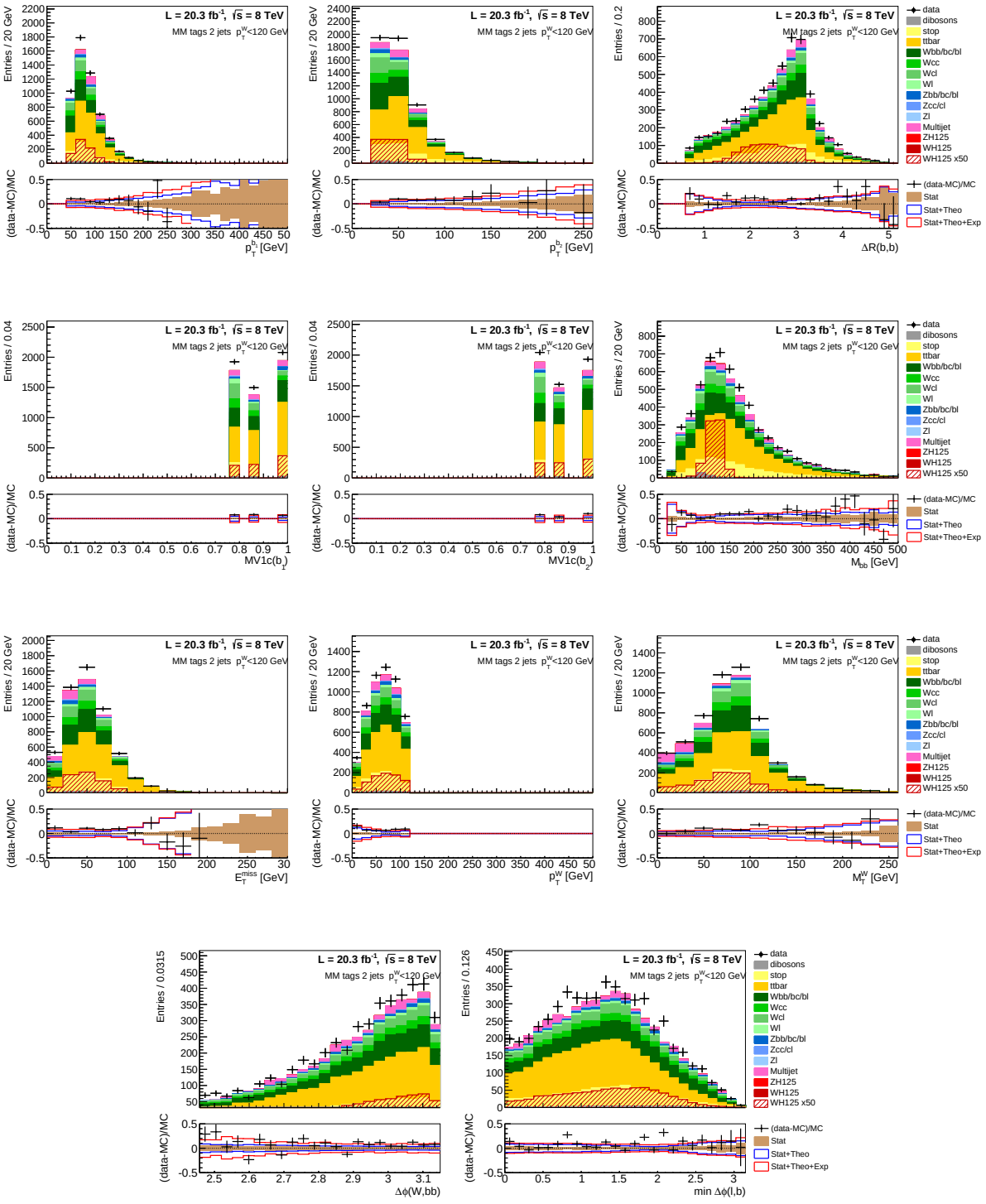


Figure C.7: Distribution of the BDT training variables for data and prediction for events with 2 signal jets, 2 MM $b-$tags and $p_{\mathrm{T}}^{W} < 120$ GeV. The uncertainty bands are exhibited in the bottom plot as detailed in Section 7.1.

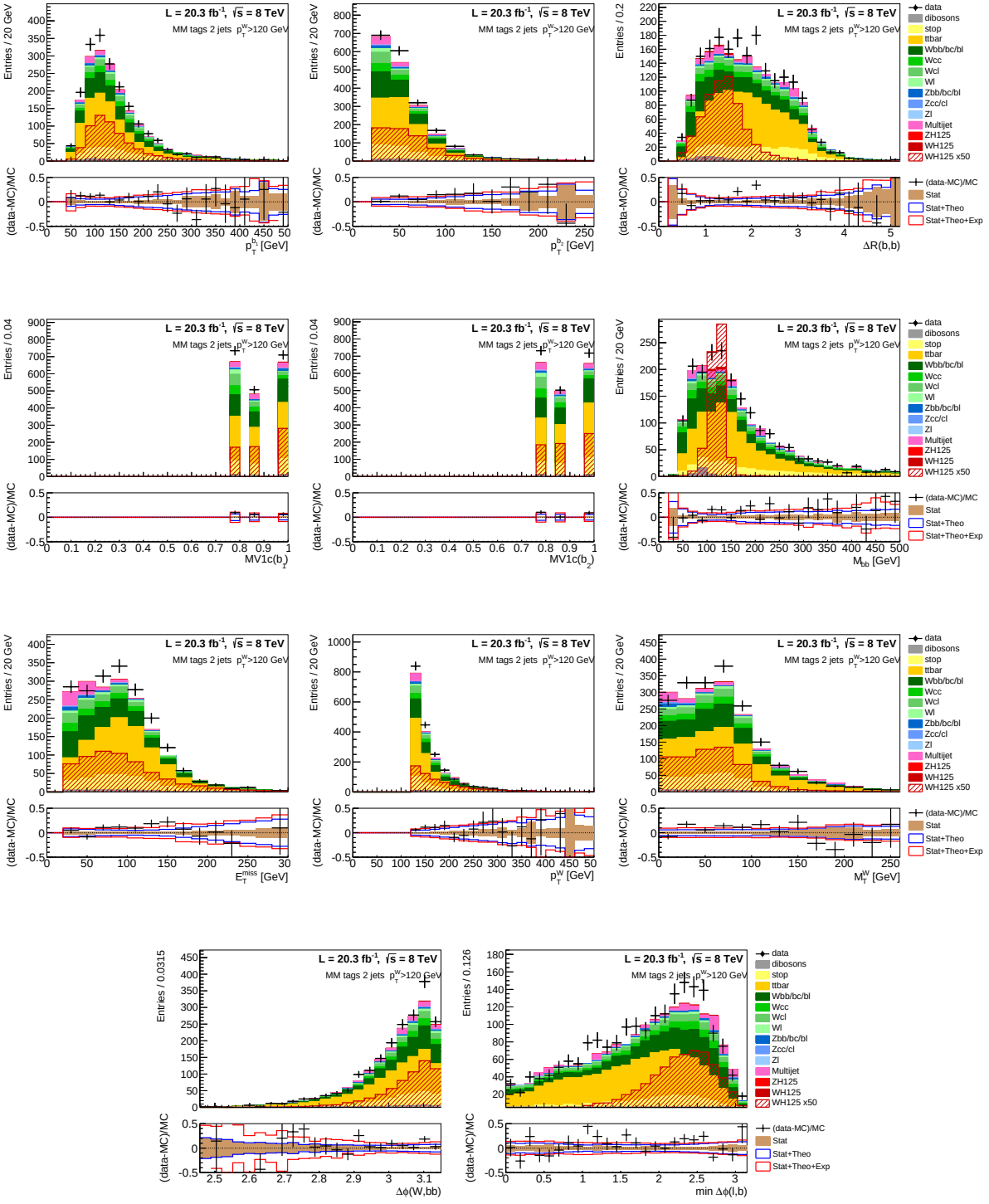


Figure C.8: Distribution of the BDT training variables for data and prediction for events with 2 signal jets, 2 MM $b-$tags and $p_{\mathrm{T}}^{W} > 120$ GeV. The uncertainty bands are exhibited in the bottom plot as detailed in Section 7.1.

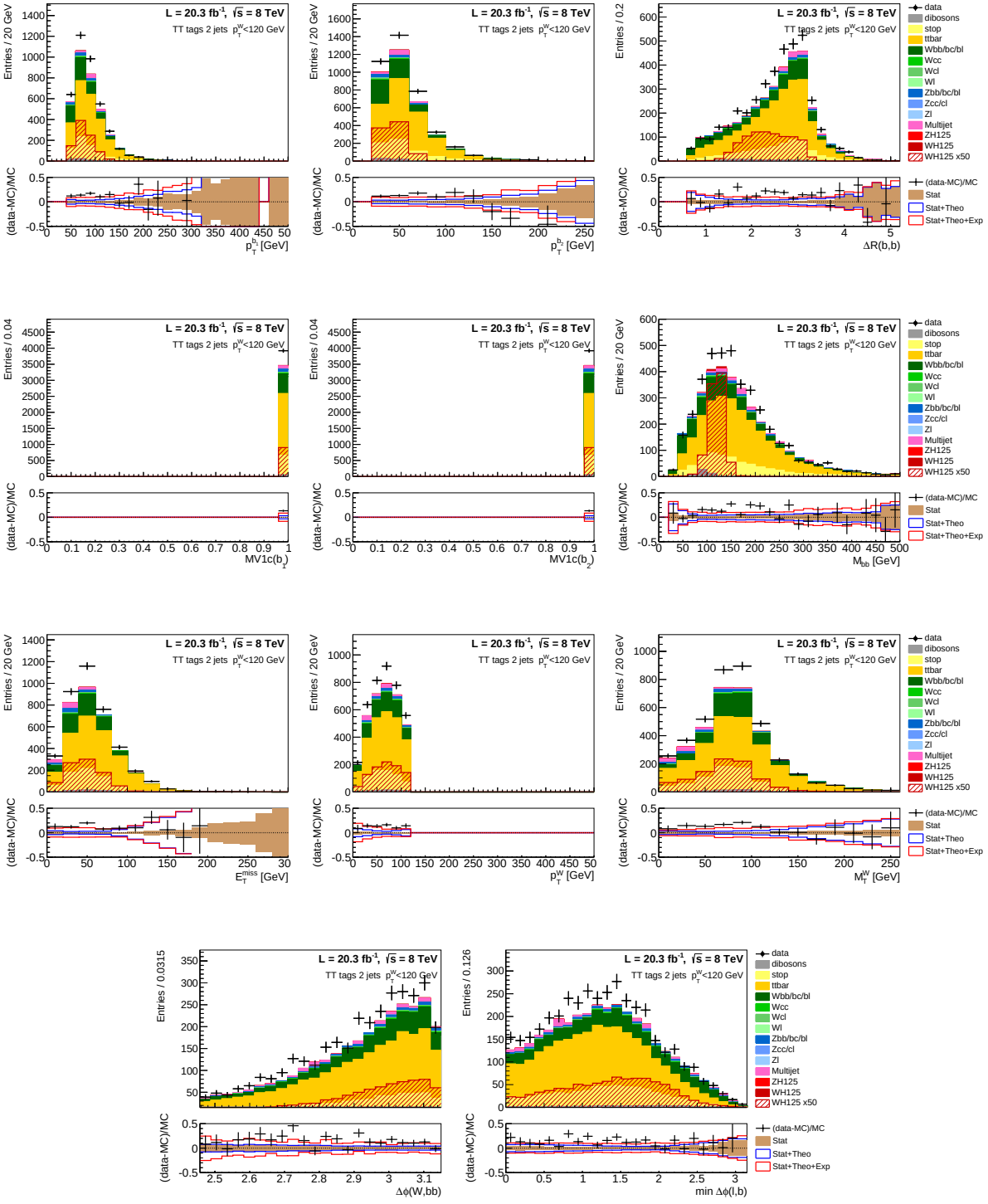


Figure C.9: Distribution of the BDT training variables for data and prediction for events with 2 signal jets, 2 TT $b-$tags and $p_{\mathrm{T}}^{W} < 120$ GeV. The uncertainty bands are exhibited in the bottom plot as detailed in Section 7.1.

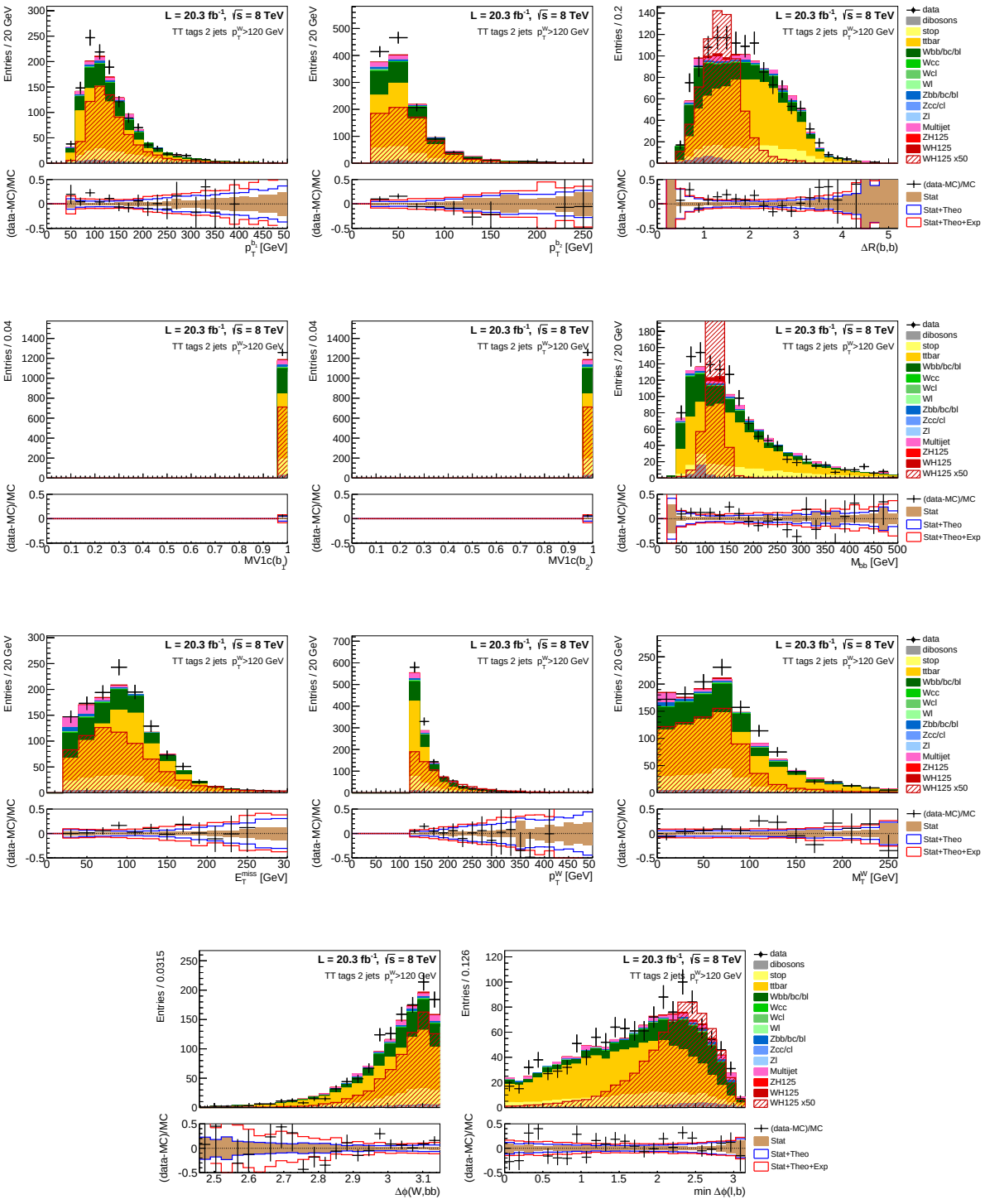


Figure C.10: Distribution of the BDT training variables for data and prediction for events with 2 signal jets, 2 TT $b-$tags and $p_{\mathrm{T}}^{W} > 120$ GeV. The uncertainty bands are exhibited in the bottom plot as detailed in Section 7.1.

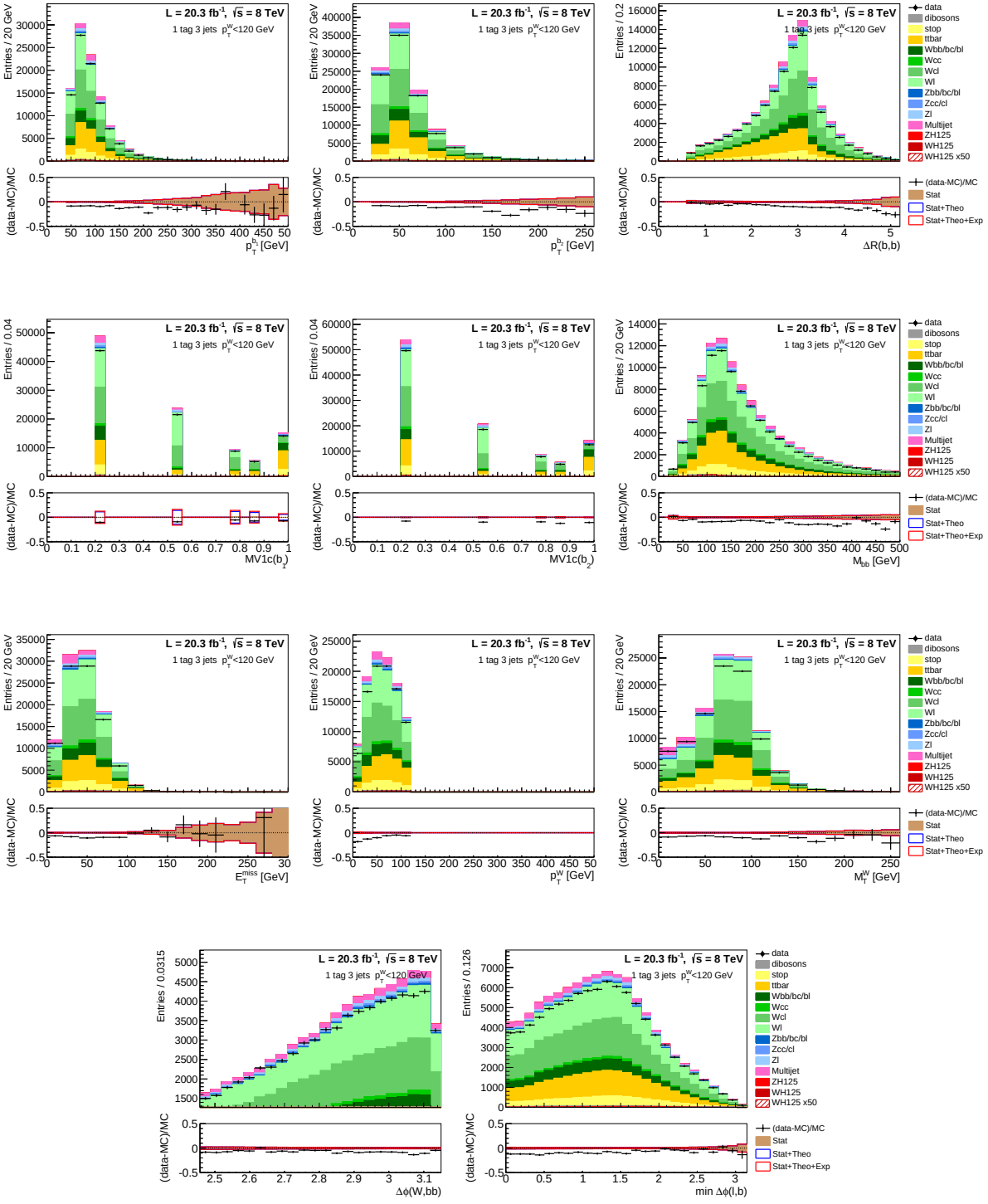


Figure C.11: Distribution of the BDT training variables for data and prediction for events with 3 signal jets, 1 $b-$tagged and $p_{\mathrm{T}}^{W} < 120$ GeV. The uncertainty bands are exhibited in the bottom plot as detailed in Section 7.1.

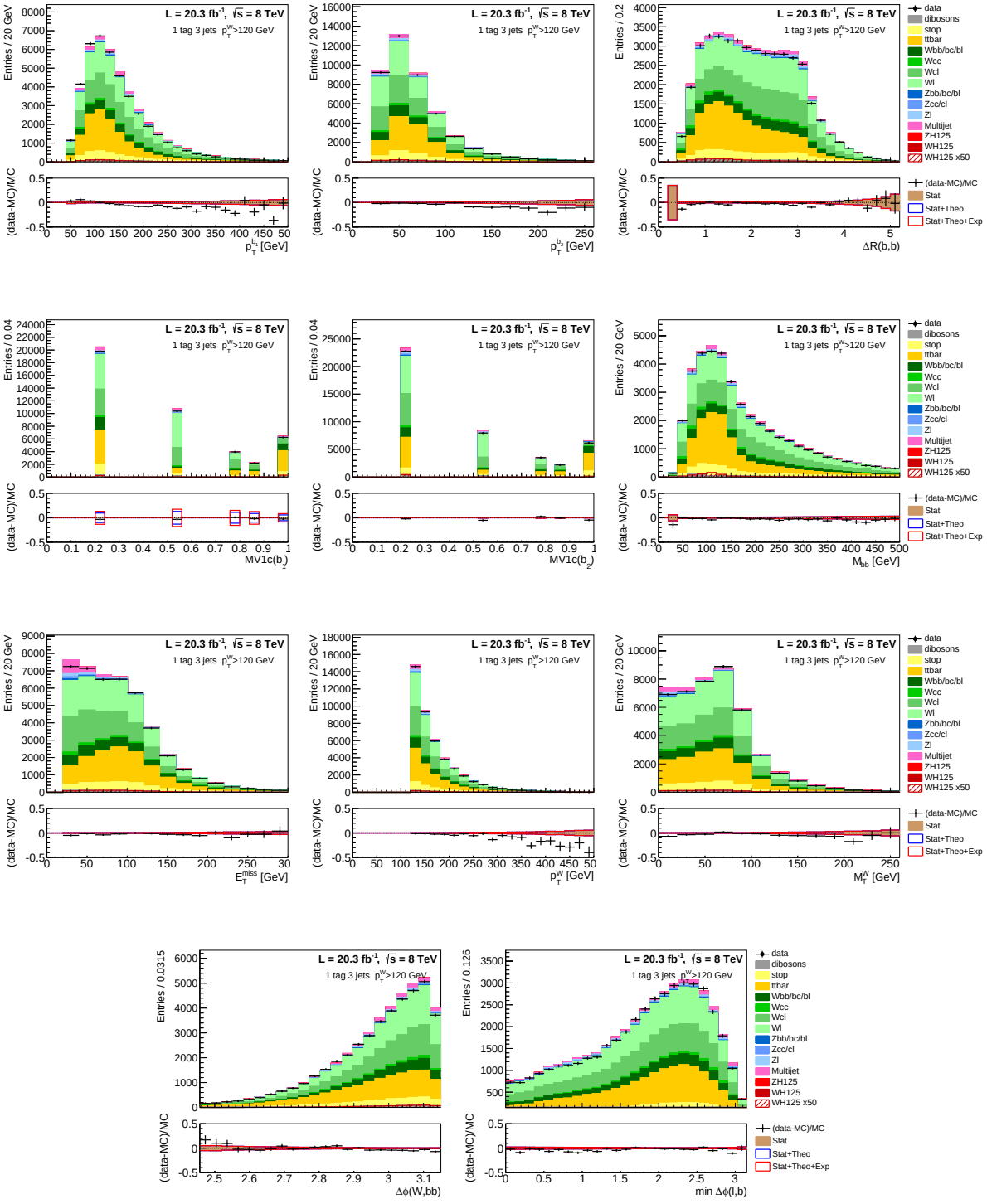


Figure C.12: Distribution of the BDT training variables for data and prediction for events with 3 signal jets, 1 $b-$tagged and $p_{\mathrm{T}}^{W} > 120$ GeV. The uncertainty bands are exhibited in the bottom plot as detailed in Section 7.1.

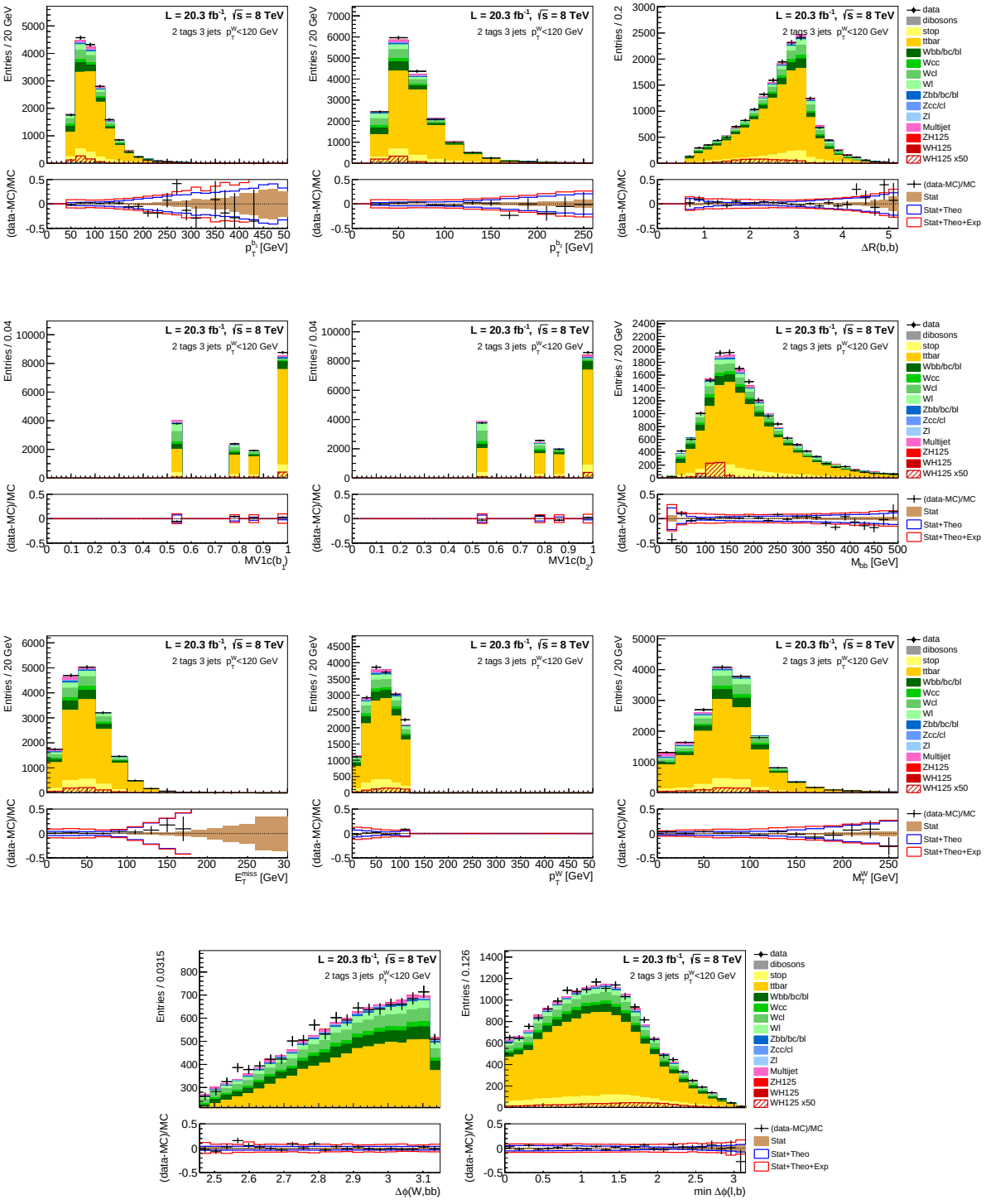


Figure C.13: Distribution of the BDT training variables for data and prediction for events with 3 signal jets, both $b-$tagged, and $p_{\mathrm{T}}^{W} < 120$ GeV. The uncertainty bands are exhibited in the bottom plot as detailed in Section 7.1.

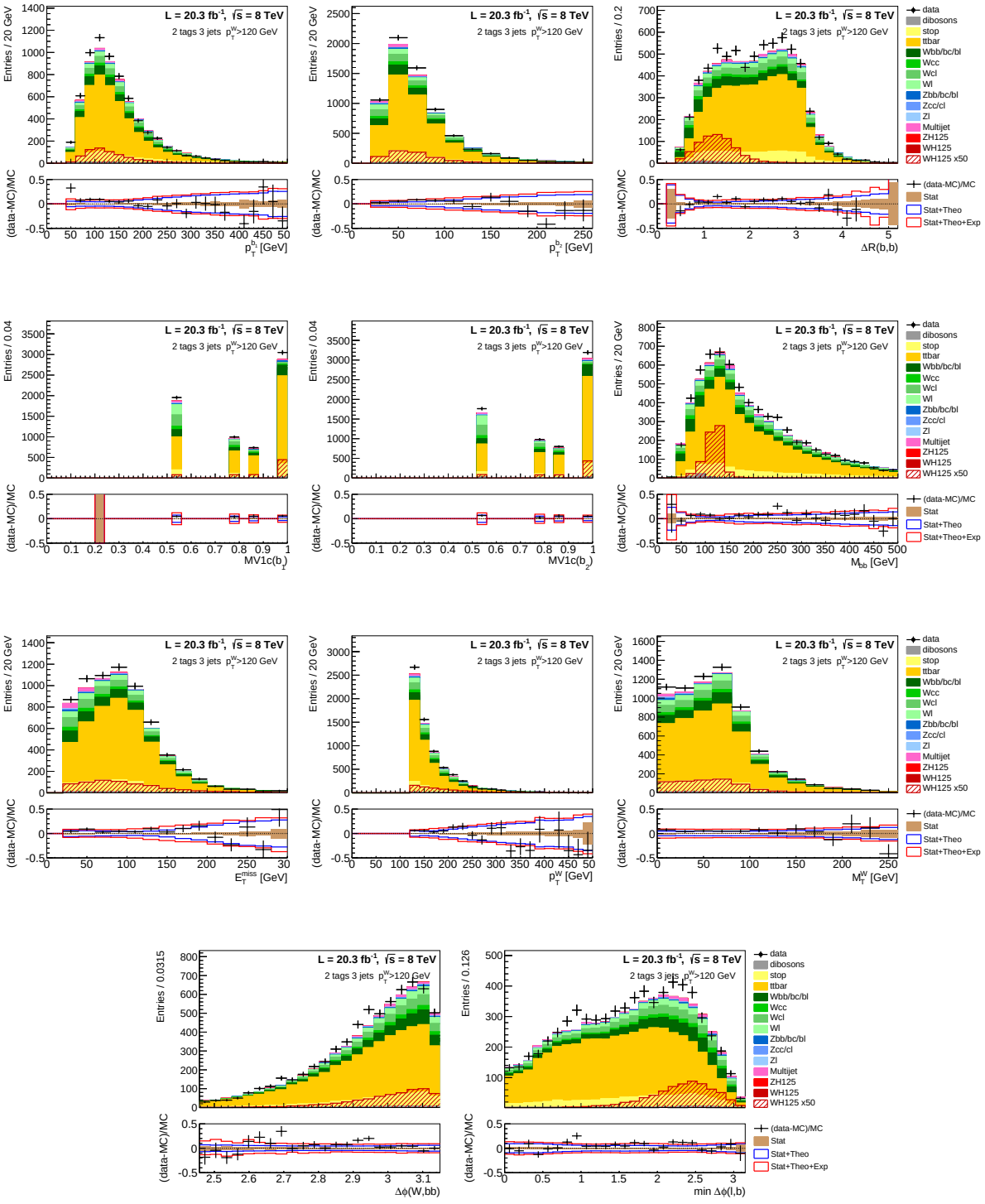


Figure C.14: Distribution of the BDT training variables for data and prediction for events with 3 signal jets, both $b-$tagged, and $p_{\mathrm{T}}^{W} > 120$ GeV. The uncertainty bands are exhibited in the bottom plot as detailed in Section 7.1.

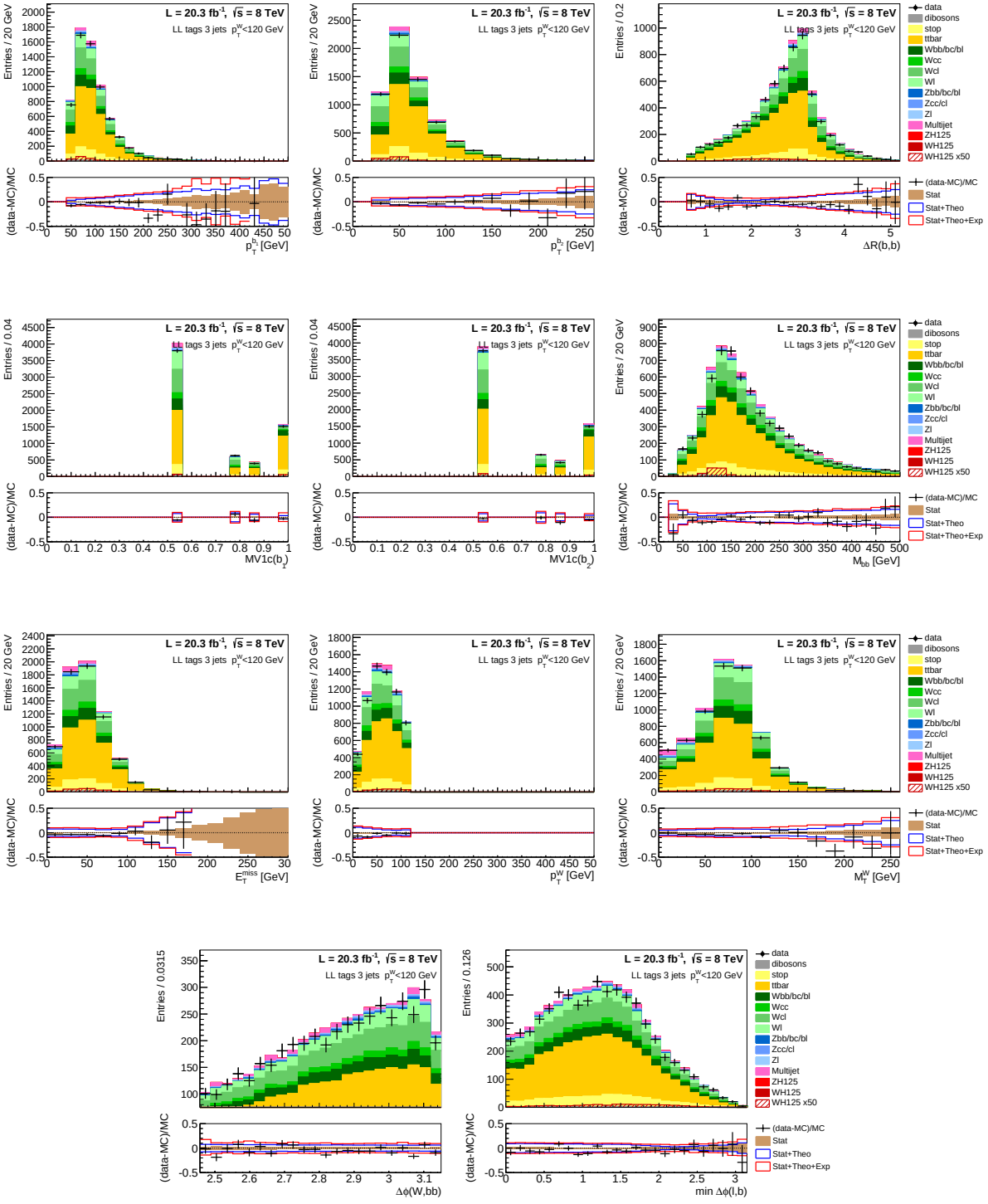


Figure C.15: Distribution of the BDT training variables for data and prediction for events with 3 signal jets, 2 LL $b-$tags and $p_{\mathrm{T}}^{W} < 120$ GeV. The uncertainty bands are exhibited in the bottom plot as detailed in Section 7.1.

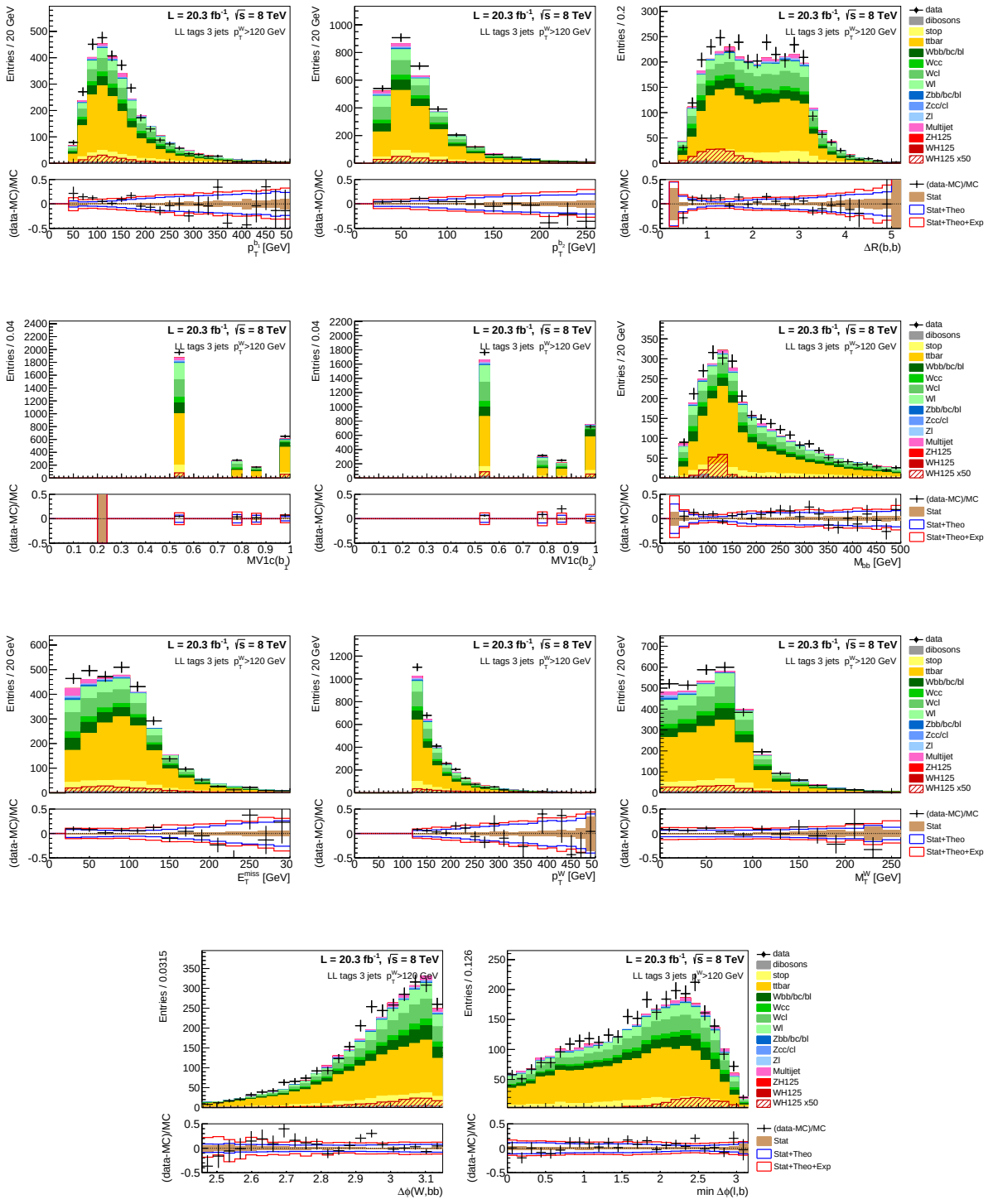


Figure C.16: Distribution of the BDT training variables for data and prediction for events with 3 signal jets, 2 LL $b-$tags and $p_{\mathrm{T}}^{W} > 120$ GeV. The uncertainty bands are exhibited in the bottom plot as detailed in Section 7.1.

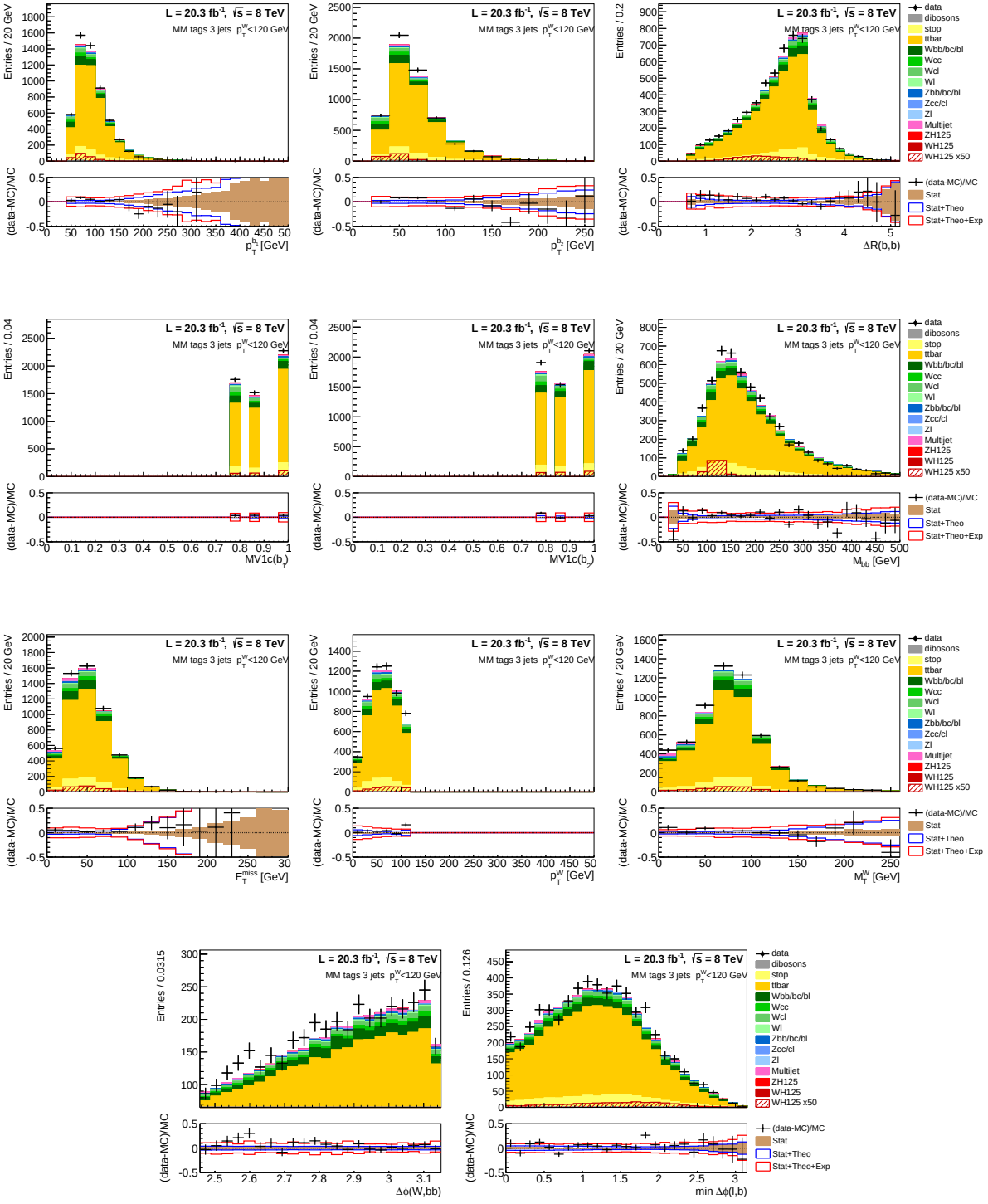


Figure C.17: Distribution of the BDT training variables for data and prediction for events with 3 signal jets, 2 MM $b-$tags and $p_{\mathrm{T}}^{W} < 120$ GeV. The uncertainty bands are exhibited in the bottom plot as detailed in Section 7.1.

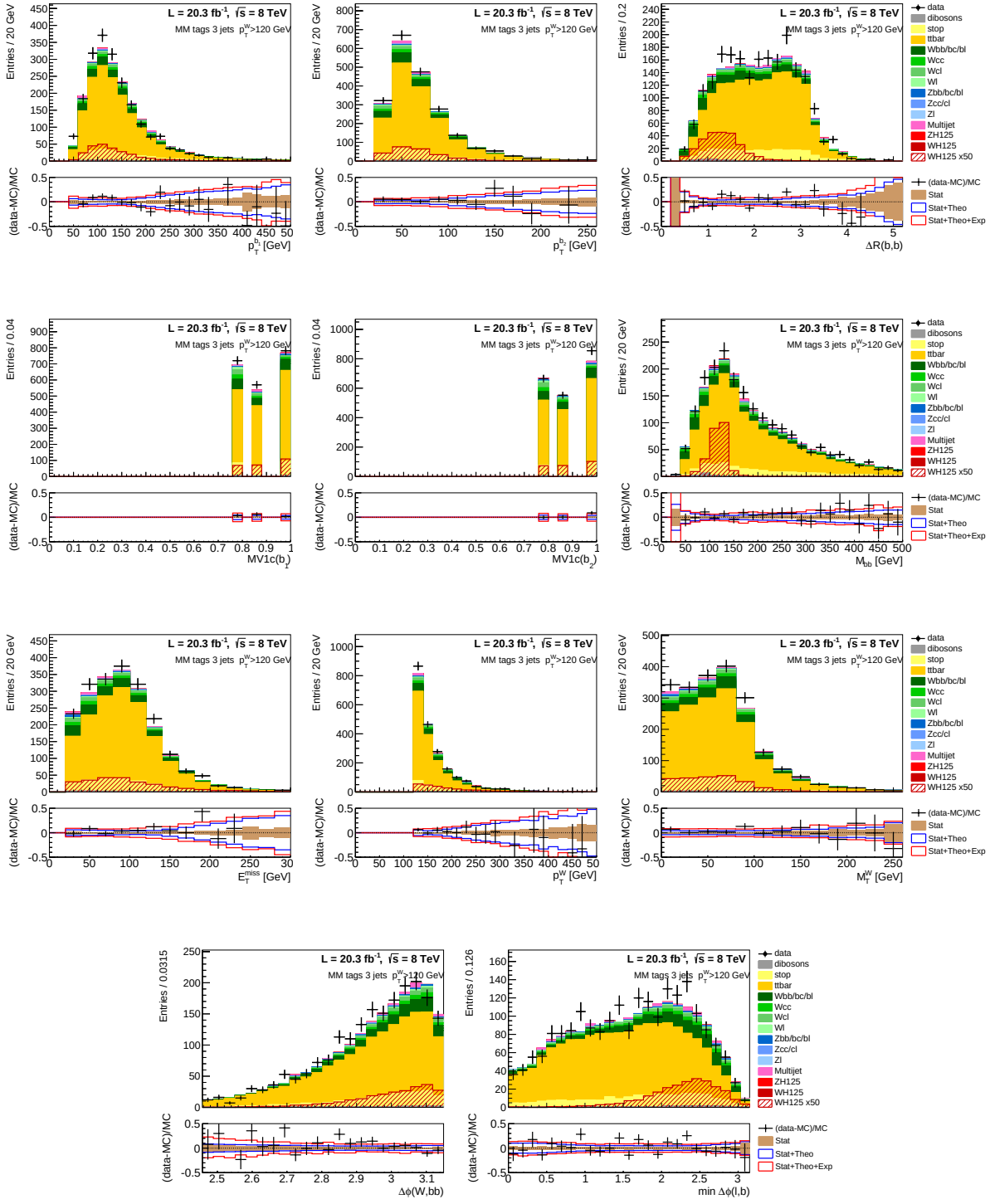


Figure C.18: Distribution of the BDT training variables for data and prediction for events with 3 signal jets, 2 MM $b-$tags and $p_{\mathrm{T}}^{W} > 120$ GeV. The uncertainty bands are exhibited in the bottom plot as detailed in Section 7.1.

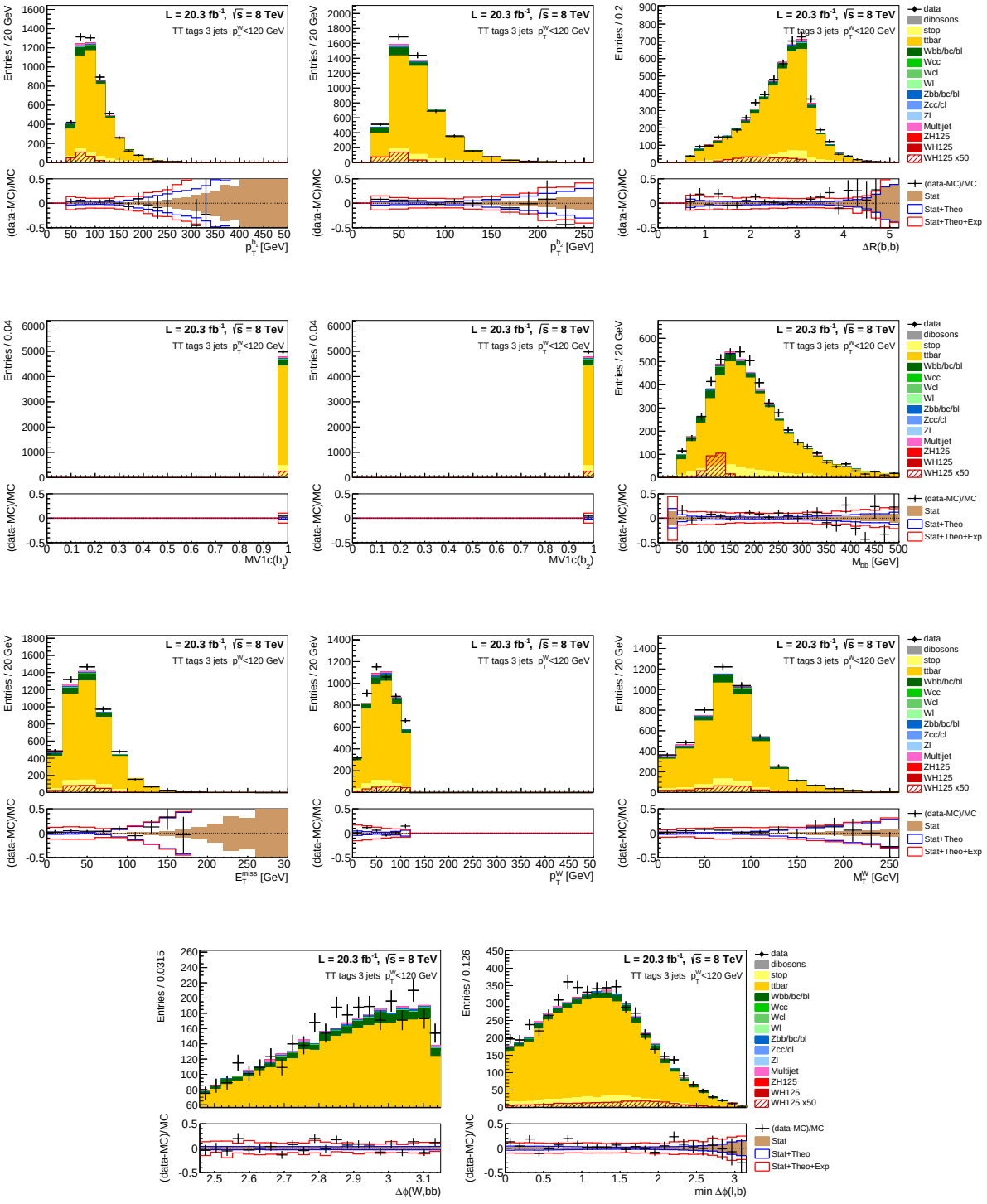


Figure C.19: Distribution of the BDT training variables for data and prediction for events with 3 signal jets, 2 TT $b-$tags and $p_{\mathrm{T}}^{W} < 120$ GeV. The uncertainty bands are exhibited in the bottom plot as detailed in Section 7.1.

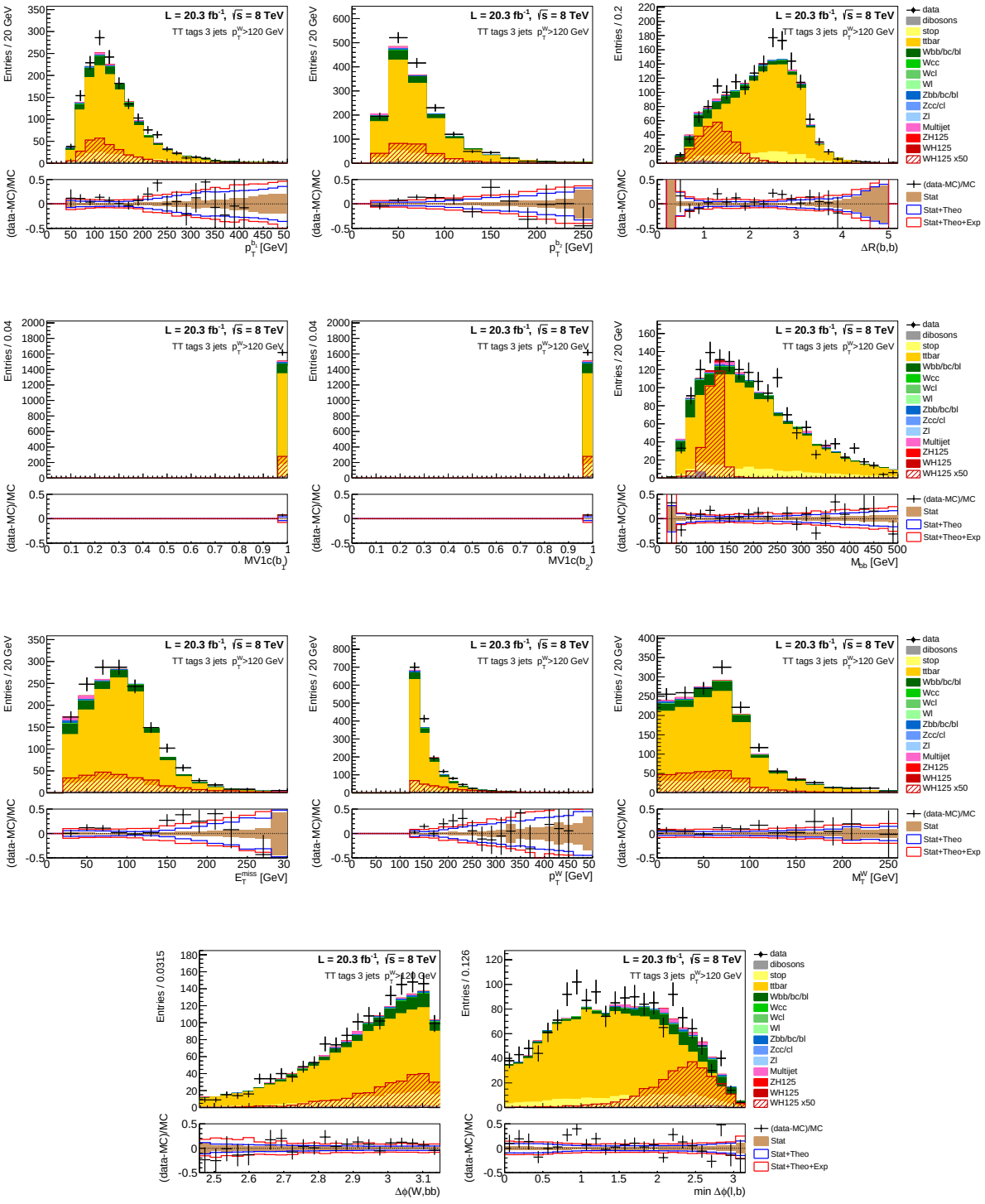


Figure C.20: Distribution of the BDT training variables for data and prediction for events with 3 signal jets, 2 TT $b-$tags and $p_{\mathrm{T}}^{W} > 120$ GeV. The uncertainty bands are exhibited in the bottom plot as detailed in Section 7.1.

Appendix D

$\Delta Y(W,H)$ and m_{Wb_1} Distributions

The distributions of the $\Delta Y(W,H)$ and m_{Wb_1} used in the 1 lepton MVA analysis are exhibited for data and prediction for all the analysis categories in Figures D.1, D.2, D.3 and D.4 for events with 2 and 3 jets respectively.

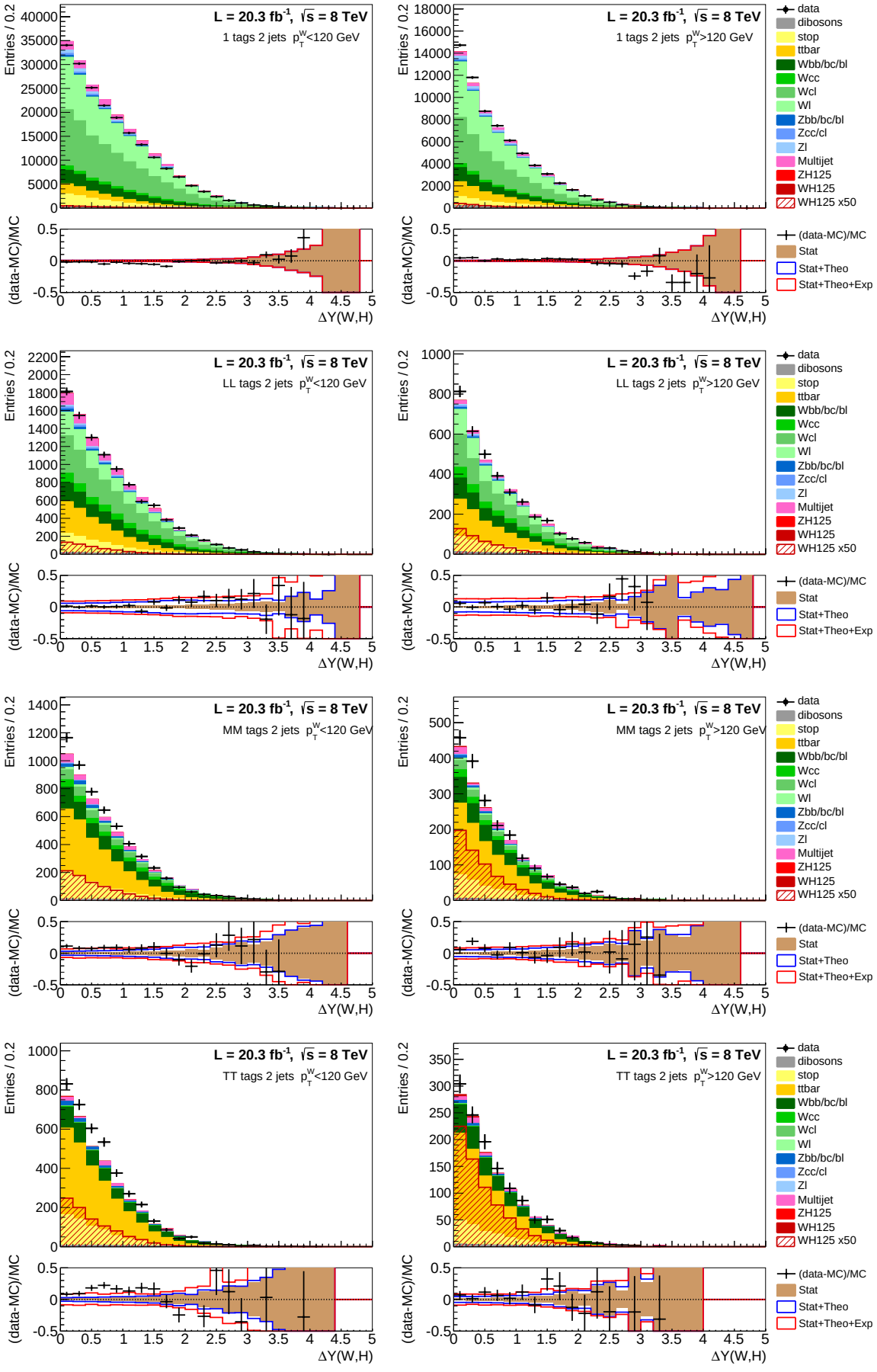


Figure D.1: $\Delta Y(W,H)$ distributions used as input to the WH BDT in the 1 lepton channel, for the 2 jets category.

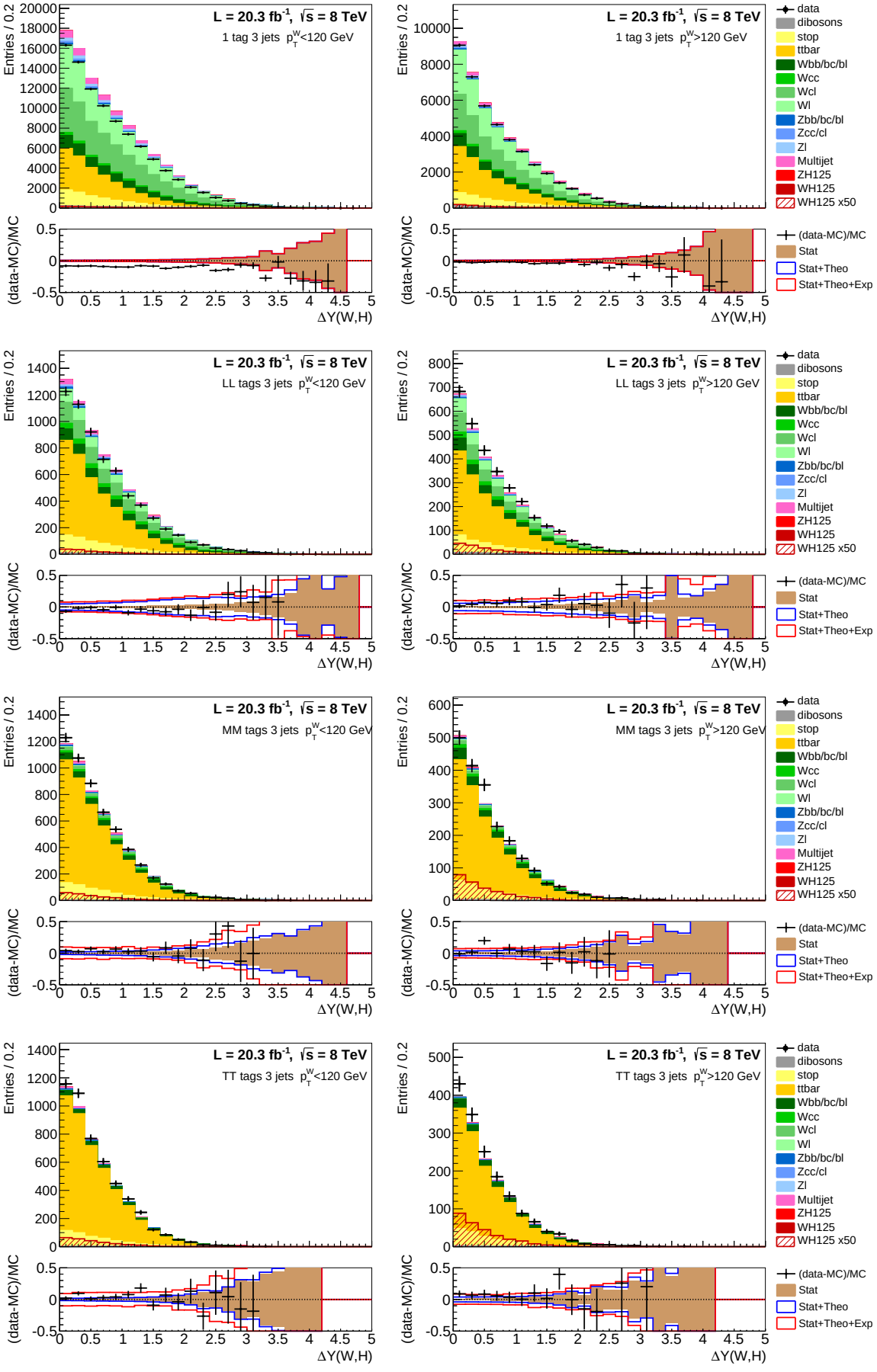


Figure D.2: $\Delta Y(W,H)$ distributions used as input to the WH BDT in the 1 lepton channel, for the 3 jets category.

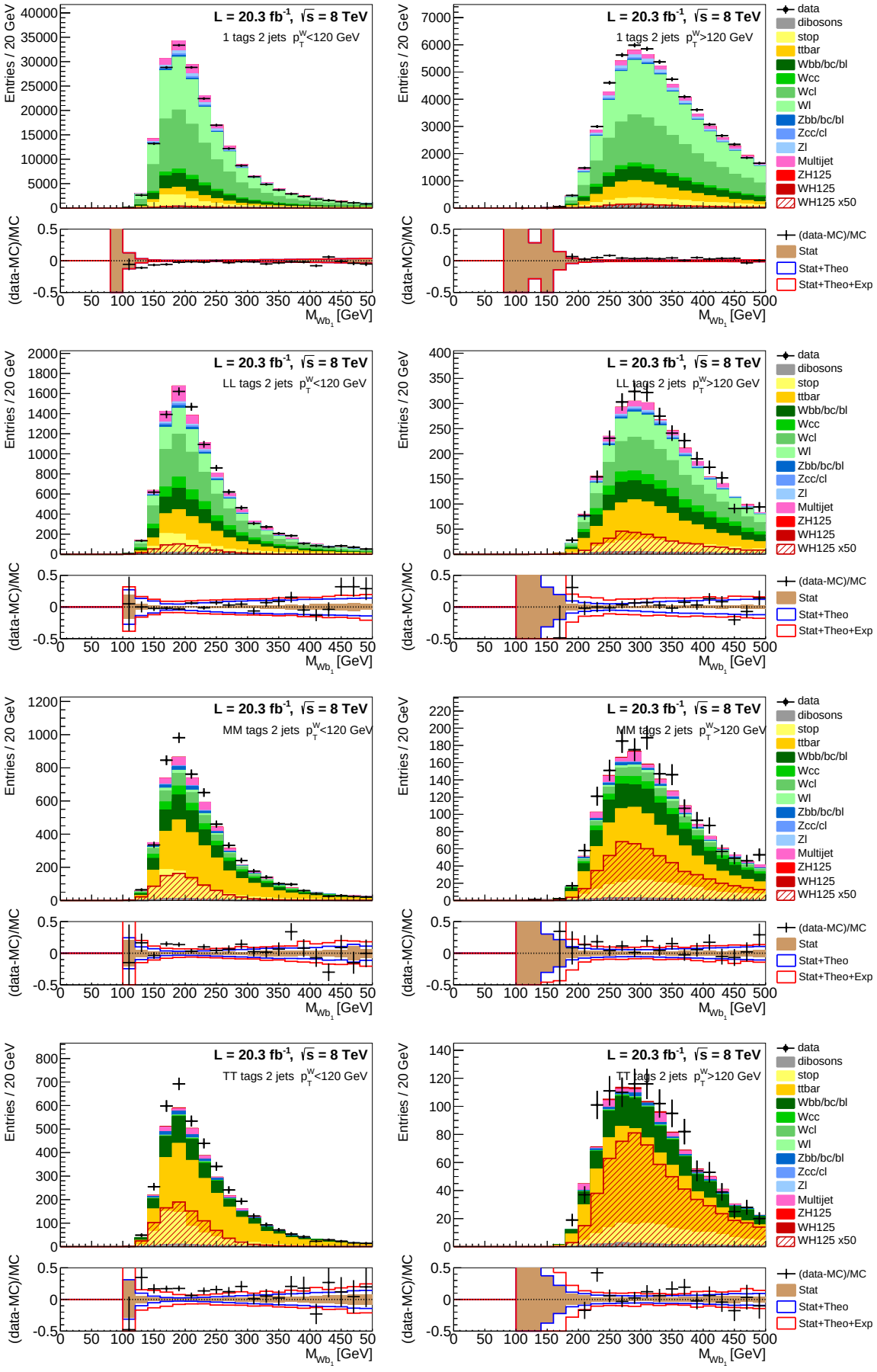


Figure D.3: m_{Wb_1} distributions used as input to the WH BDT in the 1 lepton channel, for the 2 jets category.

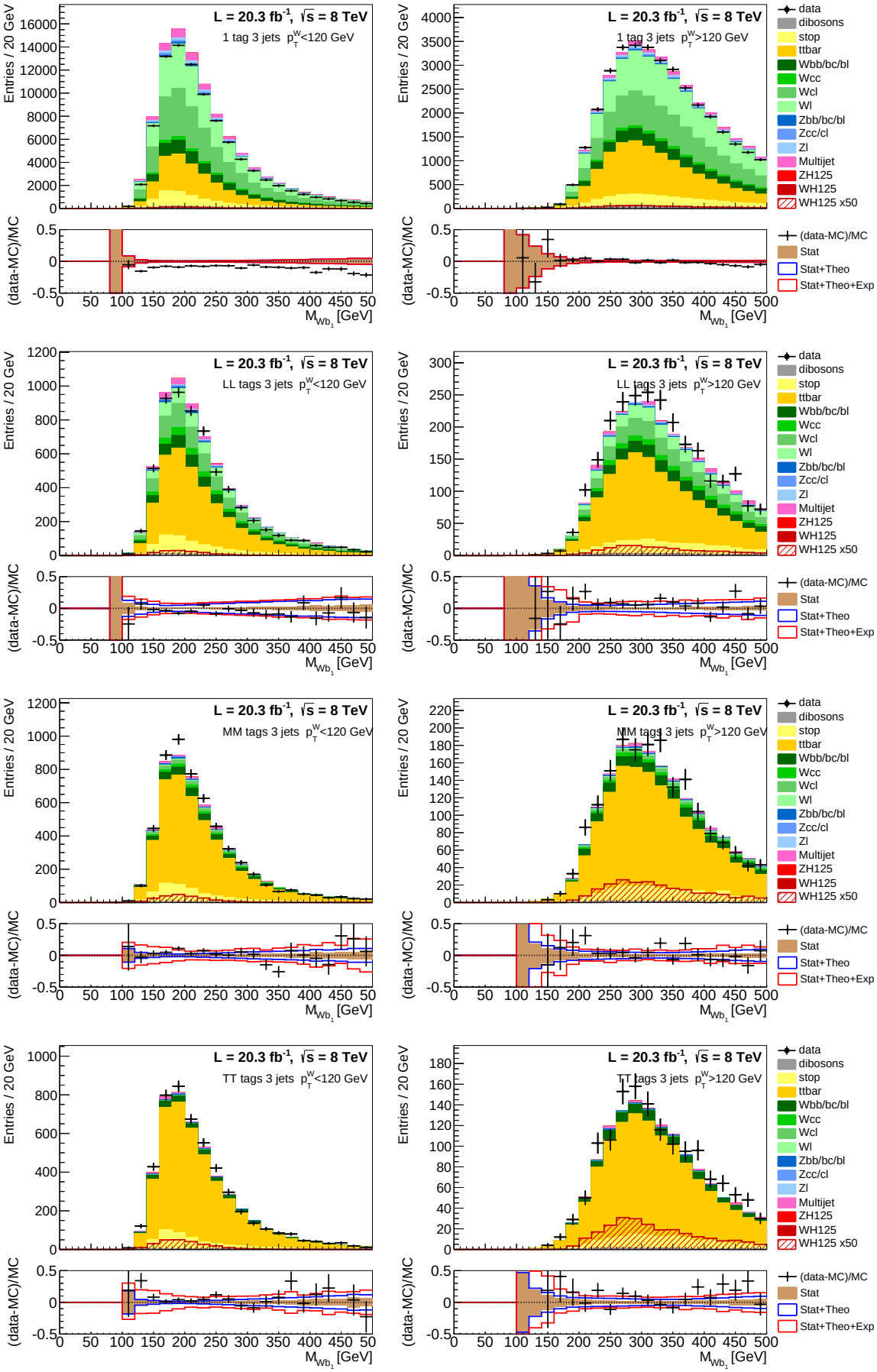


Figure D.4: m_{Wb_1} distributions used as input to the WH BDT in the 1 lepton channel, for the 3 jets category.

Appendix E

Validation of the implementation of the Systematic Uncertainties

Table E.1 shows the outcome of the validation of the analysis code with respect to the systematic uncertainties evaluation. For the first type of systematics, the systematic variation outcome must be compared with the nominal number of selected events, no weights considered, in order to roughly evaluate the impact on the analysis. For the second type, the weighted number of events must be compared: the number of events with the weight shift and with the corresponding nominal weight alone.

Systematic Variation	LIP	CERN/Edinburgh	CPPM	Tsukuba
Nominal (no weights)	11589	0,00	0,00	0,00
SysElecEUp	11586	0,08	-0,10	-0,09
SysElecEDo	11594	-0,09	0,16	0,09
SysElecEResolUp	11581	0,06	0,00	0,00
SysElecEResolDo	11601	-0,15	0,00	-0,02
Nominal Electron Efficiency Weights	11561,1	0,54	0,46	0,45
SysElecEfficUp	11626,9	0,56	-	0,71
SysElecEfficDo	11495,3	0,53	0,49	0,39
SysMuonEResolIDUp	11590	0,00	0,00	0,02
SysMuonEResolIDDo	11581	0,00	0,00	0,02
SysMuonEResolMSUp	11596	0,00	0,00	-0,28
SysMuonEResolMSDo	11586	-0,35	0,00	-0,28
Nominal Muon Efficiency Weights	11569	0,00	0,05	0,04
SysMuonEfficUp	11678,1	0,01	-0,70	-0,01
SysMuonEfficDo	11459,8	-0,01	0,81	-0,02
SysMETScaleSoftTermsUp	11568	0,01	-0,09	0,01
SysMETScaleSoftTermsDo	11621	0,00	-0,37	0,00
SysMETResoSoftTermsUp	11591	-0,16	-0,19	-0,28
SysMETResoSoftTermsDo	11573	0,26	0,41	0,04
SysJetNP1Up	11624	0,03	0,33	0,18
SysJetNP1Do	11542	0,03	-0,06	-0,06
SysJetNP2Up	11509	-0,04	-0,03	-0,03
SysJetNP2Do	11679	0,08	0,56	0,14

Systematic Variation	LIP	CERN/Edinburgh	CPPM	Tsukuba
SysJetNP3Up	11654	0,06	0,40	0,00
SysJetNP3Do	11538	-0,03	0,00	0,03
SysJetNP4Up	11575	0,00	0,04	0,04
SysJetNP4Do	11598	0,03	0,12	-0,07
SysJetNP5Up	11590	0,01	0,01	-0,06
SysJetNP5Do	11591	-0,01	0,05	0,05
SysJetNP6_restUp	11594	0,03	0,15	0,00
SysJetNP6_restDo	11576	0,00	0,02	0,02
SysJetNonClosUp	11651	0,09	0,61	-0,12
SysJetNonClosDo	11528	-0,05	0,08	0,12
SysJetEtaModelUp	11623	0,03	0,40	0,02
SysJetEtaModelDo	11531	-0,09	-0,03	-0,02
SysJetEtaStatUp	11595	0,03	0,21	0,03
SysJetEtaStatDo	11574	0,00	0,03	0,02
SysJetNPVUp	11646	0,05	0,30	0,01
SysJetNPVDo	11570	-0,07	0,09	-0,06
SysJetMuUp	11559	0,02	0,25	0,02
SysJetMuDo	11607	0,00	0,04	0,00
SysJetPilePtUp	11618	-0,17	-0,11	-0,17
SysJetPilePtDo	11555	0,20	0,22	0,20
SysJetPileRhoUp	11627	0,12	0,49	0,03
SysJetPileRhoDo	11581	-0,06	-0,03	-0,03
SysJetFlavCompUp	11370	-0,11	1,39	-0,32
SysJetFlavCompDo	11813	0,01	0,12	0,11
SysJetFlavRespUp	11465	0,03	0,84	-0,11
SysJetFlavRespDo	11677	0,11	0,21	0,19
SysJetFlavBUp	11827	-0,10	-0,40	-0,08
SysJetFlavBDo	11400	-0,06	-0,05	-0,05
Nominal $b-$Tagging Weights	11777,4	-	0,00	-
SysBTagL0EfficUp	11442	-0,02	35,44	-0,02
SysBTagL0EfficDo	11466,7	-0,02	35,31	-0,02
SysBTagL1EfficUp	11449,7	-0,02	35,40	-0,02
SysBTagL1EfficDo	11459,1	-0,02	35,35	-0,02
SysBTagL2EfficUp	11451,4	-0,02	35,38	-0,02
SysBTagL2EfficDo	11457,3	-0,02	35,37	-0,02
SysBTagL3EfficUp	11452,5	-0,02	35,38	-0,02
SysBTagL3EfficDo	11456,3	-0,02	35,36	-0,02
SysBTagL4EfficUp	11452,4	-0,02	35,38	-0,02
SysBTagL4EfficDo	11456,4	-0,02	35,37	-0,02
SysBTagL5EfficUp	11454,9	-0,02	35,37	-0,02
SysBTagL5EfficDo	11453,8	-0,02	35,38	-0,02
SysBTagL6EfficUp	11452	-0,02	35,39	-0,02
SysBTagL6EfficDo	11456,7	-0,02	35,36	-0,02
SysBTagL7EfficUp	11455,4	-0,02	35,36	-0,02
SysBTagL7EfficDo	11453,4	-0,02	35,38	-0,02
SysBTagL8EfficUp	11453,6	-0,02	35,38	-0,02
SysBTagL8EfficDo	11455,1	-0,02	35,37	-0,02

Systematic Variation	LIP	CERN/Edinburgh	CPPM	Tsukuba
SysBTagL9EfficUp	11453,6	-0,02	35,38	-0,02
SysBTagL9EfficDo	11455,2	-0,02	35,37	-0,02
SysBTagC0EfficUp	11454,6	-0,02	35,37	-0,02
SysBTagC0EfficDo	11454,2	-0,02	35,38	-0,02
SysBTagC1EfficUp	11455,7	-0,01	35,37	-0,01
SysBTagC1EfficDo	11453,1	-0,02	35,38	-0,02
SysBTagC2EfficUp	11453	-0,01	35,39	-0,01
SysBTagC2EfficDo	11455,8	-0,03	35,35	-0,03
SysBTagC3EfficUp	11455,2	-0,01	35,36	-0,01
SysBTagC3EfficDo	11453,5	-0,02	35,38	-0,02
SysBTagC4EfficUp	11455,1	-0,02	35,37	-0,02
SysBTagC4EfficDo	11453,7	-0,01	35,38	-0,01
SysBTagC5EfficUp	11455	-0,01	35,38	-0,01
SysBTagC5EfficDo	11453,8	-0,03	35,37	-0,03
SysBTagC6EfficUp	11452,9	-0,01	35,38	-0,01
SysBTagC6EfficDo	11455,8	-0,03	35,37	-0,03
SysBTagC7EfficUp	11454,5	-0,02	35,37	-0,02
SysBTagC7EfficDo	11454,2	-0,01	35,38	-0,01
SysBTagC8EfficUp	11454	-0,01	35,38	-0,01
SysBTagC8EfficDo	11454,8	-0,03	35,37	-0,03
SysBTagC9EfficUp	11454,7	-0,03	35,38	-0,03
SysBTagC9EfficDo	11454,1	-0,01	35,36	-0,01
SysBTagC10EfficUp	11453,9	-0,01	35,38	-0,01
SysBTagC10EfficDo	11454,9	-0,02	35,37	-0,02
SysBTagC11EfficUp	11454,2	-0,02	35,37	-0,02
SysBTagC11EfficDo	11454,6	-0,01	35,38	-0,01
SysBTagC12EfficUp	11454,6	-0,03	35,37	-0,03
SysBTagC12EfficDo	11454,2	-0,01	35,38	-0,01
SysBTagC13EfficUp	11454,3	-0,02	35,37	-0,02
SysBTagC13EfficDo	11454,5	-0,02	35,37	-0,02
SysBTagC14EfficUp	11454,6	-0,02	35,37	-0,02
SysBTagC14EfficDo	11454,2	-0,02	35,37	-0,02
SysBTagB0EfficUp	11663,1	-0,01	35,34	-0,01
SysBTagB0EfficDo	11247,5	-0,02	35,41	-0,02
SysBTagB1EfficUp	11671	-0,02	35,47	-0,03
SysBTagB1EfficDo	11239,8	-0,02	35,27	-0,02
SysBTagB2EfficUp	11494,6	-0,02	35,31	-0,02
SysBTagB2EfficDo	11414,2	-0,01	35,44	-0,01
SysBTagB3EfficUp	11509,7	-0,02	35,26	-0,02
SysBTagB3EfficDo	11399,2	-0,01	35,49	-0,01
SysBTagB4EfficUp	11590,4	-0,02	35,22	-0,02
SysBTagB4EfficDo	11318,7	-0,02	35,54	-0,02
SysBTagB5EfficUp	11568,7	-0,01	35,44	-0,01
SysBTagB5EfficDo	11340,5	-0,02	35,30	-0,02
SysBTagB6EfficUp	11527,5	-0,02	35,35	-0,02
SysBTagB6EfficDo	11381,4	-0,01	35,40	-0,01
SysBTagB7EfficUp	11468,8	-0,01	35,35	-0,01

Systematic Variation	LIP	CERN/Edinburgh	CPPM	Tsukuba
SysBTagB7EfficDo	11439,9	-0,02	35,40	-0,02
SysBTagB8EfficUp	11474,8	-0,01	35,36	-0,01
SysBTagB8EfficDo	11433,9	-0,02	35,39	-0,02
SysBTagB9EfficUp	11358,2	-0,02	35,37	-0,02
SysBTagB9EfficDo	11551	-0,02	35,38	-0,02

Table E.1: The number of events passing the 2 $b-$tagged signal jets selection is shown for a reference samples of simulated WH signal events for each systematic variation using the LIP analysis code. In case of systematic variations related with the event weight, the weighted number of events is shown istead, together withe their nominal weight. Percentage deviation of the number events as obtained by the other groups codes with respect to LIP, defined as $\Delta = (N_{group} - N_{LIP})/N_{LIP} \times 100\%$.

Bibliography

[1] Francis Halzen and Alan D. Martin. *Quarks and Leptons: An Introductory Course in Modern Particle Physics*. John Wiley & Sons, Inc., 1984.

[2] David Griffiths. *Introduction to Elementary Particles*. Wiley-VCH, 2008.

[3] Michele Maggiore. *A Modern Introduction to Quantum Field Theory*. Oxford University Press, 2005.

[4] F. Englert and R. Brout. Broken Symmetry and the Mass of Gauge Vector Mesons. *Phys. Rev. Lett.*, 13:321–323, 1964.

[5] Peter W. Higgs. Broken Symmetries and the Masses of Gauge Bosons. *Phys. Rev. Lett.*, 13:508–509, 1964.

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

[7] ATLAS Collaboration. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett.*, B716:1–29, 2012.

[8] CMS Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys. Lett.*, B716:30–61, 2012.

[9] M. Lindner. Implications of Triviality for the Standard Model. *Z. Phys.*, C31:295, 1986.

[10] Henning et al. Flacher. Revisiting the Global Electroweak Fit of the Standard Model and Beyond with Gfitter. *Eur. Phys. J.*, C60:543–583, 2009.

[11] LHC Higgs Cross Section Working Group. Handbook of LHC Higgs Cross Sections: 3. Higgs Properties. *CERN-2013-004*, CERN, Geneva, 2013.

[12] John C. Collins, Davison E. Soper, and George F. Sterman. Factorization of Hard Processes in QCD. *Adv. Ser. Direct. High Energy Phys.*, 5:1–91, 1989.

[13] Bo Andersson, G. Gustafson, G. Ingelman, and T. Sjostrand. Parton Fragmentation and String Dynamics. *Phys. Rept.*, 97:31–145, 1983.

[14] Thomas D. Gottschalk. An Improved Description of Hadronization in the QCD Cluster Model for e^+e^- Annihilation. *Nucl. Phys.*, B239:349–381, 1984.

[15] S. Dittmaier et al. Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables. (arXiv:1101.0593), 2011.

[16] R. Barate et al. Search for the standard model Higgs boson at LEP. *Phys. Lett.*, B565:61–75, 2003.

[17] TEVNPH Working Group. Combined CDF and D0 Upper Limits on Standard Model Higgs Boson Production with up to 8.6 fb^{-1} of Data. 2011.

[18] T. Aaltonen et al. Evidence for a particle produced in association with weak bosons and decaying to a bottom-antibottom quark pair in Higgs boson searches at the Tevatron. *Phys. Rev. Lett.*, 109:071804, 2012.

[19] J. Stirling. "http://mstwpdf.hepforge.org", 2015.

[20] ATLAS Collaboration. Measurements of the Higgs boson production and decay rates and coupling strengths using pp collision data at $\sqrt{s} = 7$ and 8 TeV in the ATLAS experiment. *Eur. Phys. J.*, C76(1):6, 2016.

[21] CMS Collaboration. Precise determination of the mass of the Higgs boson and tests of compatibility of its couplings with the standard model predictions using proton collisions at 7 and 8 TeV. *Eur. Phys. J.*, C75(5):212, 2015.

[22] 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. *Phys. Rev. Lett.*, 114:191803, 2015.

[23] ATLAS and CMS Collaborations. Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV. *JHEP*, 08:045, 2016.

[24] ATLAS Collaboration. Study of the spin and parity of the Higgs boson in diboson decays with the ATLAS detector. *Eur. Phys. J.*, C75(10):476, 2015.

[25] CMS Collaboration. Constraints on the spin-parity and anomalous HVV couplings of the Higgs boson in proton collisions at 7 and 8 TeV. *Phys. Rev.*, D92(1):012004, 2015.

[26] European Organization for Nuclear Research. LHC Machine. *JINST*, 3, 2008.

[27] CERN. CERN Webpage. `https://home.cern`.

[28] ATLAS Collaboration. Luminosity public results for Run I. `https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResults`.

[29] ATLAS Collaboration. Luminosity public results for Run II. `https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2`.

[30] ATLAS Collaboration. The ATLAS Experiment at the CERN Large Hadron Collider. *JINST*, 3, 2008.

[31] ATLAS Collaboration. Performance of the ATLAS Inner Detector Track and Vertex Reconstruction in the High Pile-Up LHC Environment. Technical Report ATLAS-CONF-2012-042, CERN, Mar 2012.

[32] ATLAS Collaboration. Performance of primary vertex reconstruction in proton-proton collisions at $\sqrt{s}$ =7 TeV in the ATLAS experiment. Technical Report ATLAS-CONF-2010-069, CERN, Jul 2010.

[33] ATLAS Collaboration. Electron reconstruction and identification efficiency measurements with the ATLAS detector using the 2011 LHC proton-proton collision data. *Eur. Phys. J. C*, page 74. 38 p, Apr 2014.

[34] ATLAS Collaboration. Electron efficiency measurements with the ATLAS detector using 2012 LHC proton-proton collision data. *Eur. Phys. J.*, C77(3):195, 2017.

[35] ATLAS Collaboration. Electron trigger public results. "https://twiki.cern.ch/twiki/bin/view/AtlasPublic/EgammaTriggerPublicResults", 2015.

[36] ATLAS Collaboration. Electron and photon energy calibration with the ATLAS detector using LHC Run 1 data. *Eur. Phys. J. C*, page 74. 51 p, Jul 2014.

[37] K. A. Olive et al. Review of Particle Physics. *Chin. Phys.*, C38, 2014.

[38] ATLAS Collaboration. Measurement of the muon reconstruction performance of the ATLAS detector using 2011 and 2012 LHC proton–proton collision data. *Eur. Phys. J.*, C74(11):3130, 2014.

[39] ATLAS Collaboration. Performance of the ATLAS muon trigger in pp collisions at $\sqrt{s} = 8$ TeV. *Eur. Phys. J.*, C75:120, 2015.

[40] ATLAS Collaboration. Muon reconstruction efficiency and momentum resolution of the ATLAS experiment in proton-proton collisions at $\sqrt{s}$ = 7 TeV in 2010. *Eur. Phys. J.*, C74(9):3034, 2014.

[41] ATLAS Collaboration. Jet energy measurement and its systematic uncertainty in proton-proton collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector. *Eur. Phys. J.*, C75:17, 2015.

[42] ATLAS Collaboration. Jet energy measurement with the ATLAS detector in proton-proton collisions at $\sqrt{s} = 7$ TeV. *Eur. Phys. J.*, C73(3):2304, 2013.

[43] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The Anti-k(t) jet clustering algorithm. *JHEP*, 04:063, 2008.

[44] ATLAS Collaboration. Jet energy resolution in proton-proton collisions at $\sqrt{s} = 7$ TeV recorded in 2010 with the ATLAS detector. *Eur. Phys. J.*, C73(3):2306, 2013.

[45] ATLAS Collaboration. Performance of pile-up mitigation techniques for jets in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector. Technical Report CERN-PH-EP-2015-206, CERN, 2015.

[46] ATLAS Collaboration. Jet global sequential corrections with the ATLAS detector in proton-proton collisions at sqrt(s) = 8 TeV. Technical Report ATLAS-CONF-2015-002, CERN, Mar 2015.

[47] M. J. Sousa. *Search for the Higgs Boson at ATLAS/LHC, in associated production with a Z boson*. PhD thesis, Universidade de Lisboa, 2017.

[48] ATLAS Collaboration. Performance of b-Jet Identification in the ATLAS Experiment. Technical Report CERN-PH-EP-2015-216, CERN, 2015.

[49] ATLAS Collaboration. Commissioning of the ATLAS high-performance b-tagging algorithms in the 7 TeV collision data. Technical Report ATLAS-CONF-2011-102, CERN, 2011.

[50] ATLAS Collaboration. Search for the bb decay of the Standard Model Higgs boson in associated (W/Z)H production with the ATLAS detector. *JHEP*, 01:069, 2015.

[51] ATLAS Collaboration. Performance of Missing Transverse Momentum Reconstruction in ATLAS studied in Proton-Proton Collisions recorded in 2012 at 8TeV. Technical Report ATLAS-CONF-2013-082, CERN, Aug 2013.

[52] ATLAS Collaboration. Readiness of the ATLAS Tile Calorimeter for LHC collisions. *Eur. Phys. J. C*, 70(arXiv:1007.5423. CERN-PH-EP-2010-024):1193–1236, Jul 2010.

[53] ATLAS Collaboration. ATLAS Calorimeter Response to Single Isolated Hadrons and Estimation of the Calorimeter Jet Scale Uncertainty. Technical Report ATLAS-CONF-2011-028, CERN, Mar 2011.

[54] ATLAS Collaboration. Update on the jet energy scale systematic uncertainty for jets produced in proton-proton collisions at $\sqrt{s} = 7$ TeV measured with the ATLAS detector. Technical Report ATLAS-CONF-2011-007, CERN, Feb 2011.

[55] D Boumediene, E Dubreuil, and D Pallin. Calibration of the atlas tile calorimeter channels using the laser system. Technical Report ATL-TILECAL-INT-2014-002, CERN, Feb 2014.

[56] J. Abdallah et al. The Laser calibration of the Atlas Tile Calorimeter during the LHC run 1. *JINST*, 11(10):T10005, 2016.

[57] V. Giangiobbe. The tilecal laser calibration system. *Physics Procedia*, 37(0):287–292, 2012.

[58] ATLAS Collaboration. The ATLAS Simulation Infrastructure. *Eur. Phys. J.*, C70(arXiv:1005.4568), 2010.

[59] S. Agostinelli et al. GEANT4: A Simulation toolkit. *Nucl. Instrum. Meth.*, A506, 2003.

[60] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands. PYTHIA 6.4 Physics and Manual. *JHEP*, 05(arXiv:hep-ph/0603175), 2006.

[61] T. Gleisberg, Stefan. Hoeche, F. Krauss, M. Schonherr, S. Schumann, F. Siegert, and J. Winter. Event generation with SHERPA 1.1. *JHEP*, 02:0–07, 2009.

[62] T. Sjostrand, S. Mrenna, and Peter Z. Skands. A Brief Introduction to PYTHIA 8.1. *Comput. Phys. Commun.*, 178, 2008.

[63] G. Corcella et al. HERWIG 6: An Event generator for hadron emission reactions with interfering gluons (including supersymmetric processes). *JHEP*, 01:0–10, 2001.

[64] S. Frixione, P. Nason, and C. Oleari. Matching NLO QCD computations with Parton Shower simulations: the POWHEG method. *JHEP*, 11:070, 2007.

[65] Borut Paul Kersevan and Elzbieta Richter-Was. The Monte Carlo event generator AcerMC versions 2.0 to 3.8 with interfaces to PYTHIA 6.4, HERWIG 6.5 and ARIADNE 4.1. *Comput. Phys. Commun.*, 184, 2013.

[66] Stefano Frixione and Bryan R. Webber. Matching NLO QCD computations and parton shower simulations. *JHEP*, 06:0–29, 2002.

[67] Piotr Golonka and Zbigniew Was. PHOTOS Monte Carlo: A Precision tool for QED corrections in Z and W decays. *Eur. Phys. J.*, C45:97–107, 2006.

[68] A. Sherstnev and R. S. Thorne. Parton Distributions for LO Generators. *Eur. Phys. J.*, C55:553–575, 2008.

[69] ATLAS Collaboration. Further ATLAS tunes of PYTHIA6 and Pythia8. Technical Report ATL-PHYS-PUB-2011-014, CERN, 2011.

[70] ATLAS Collaboration. General Production and Performance Public Results of the ATLAS Data Preparation group. `https://twiki.cern.ch/twiki/bin/view/AtlasPublic/DataPrepGenPublicResults`.

[71] J. Pumplin, D. R. Stump, J. Huston, H. L. Lai, Pavel M. Nadolsky, and W. K. Tung. New generation of parton distributions with uncertainties from global QCD analysis. *JHEP*, 07:0–12, 2002.

[72] Hung-Liang Lai, Marco Guzzi, Joey Huston, Zhao Li, Pavel M. Nadolsky, Jon Pumplin, and C.-P. Yuan. New parton distributions for collider physics. *Phys. Rev.*, D82, 2010.

[73] W H Bell, L Bellagamba, C Bernard, K Black, J T Childers, R Di Sipio, C Gabaldon, J Kvita, M Romano, J Sjolin, F Spano, I J Watson, and M Yamada. Measurements of top-quark pair differential cross-sections in the l+jets channel in pp collisions at $\sqrt{s} = 7$ TeV using the ATLAS detector. Technical Report ATL-COM-PHYS-2013-1661, CERN, Dec 2013.

[74] Michał Czakon, Paul Fiedler, and Alexander Mitov. Total Top-Quark Pair-Production Cross Section at Hadron Colliders Through $O(\alpha_S^4)$. *Phys. Rev. Lett.*, 110:252004, 2013.

[75] Nikolaos Kidonakis. NNLL resummation for s-channel single top quark production. *Phys. Rev.*, D81:054028, 2010.

[76] Nikolaos Kidonakis. Next-to-next-to-leading-order collinear and soft gluon corrections for t-channel single top quark production. *Phys. Rev.*, D83:091503, 2011.

[77] Nikolaos Kidonakis. Two-loop soft anomalous dimensions for single top quark associated production with a W- or H-. *Phys. Rev.*, D82:054018, 2010.

[78] Kirill Melnikov and Frank Petriello. Electroweak gauge boson production at hadron colliders through $O(\alpha_s^2)$. *Phys. Rev.*, D74:114017, 2006.

[79] ATLAS Collaboration. Expected performance of the ATLAS *b*-tagging algorithms in Run-2. Technical Report ATL-PHYS-PUB-2015-022, CERN, Jul 2015.

[80] P. Berta, F. Filthaut, V. Dao, E. Le Menedeu, F. Parodi, G. Piacquadio, T. Scanlon, M. Ughetto, and L. Zhang. Continuous *b*-tagging for the ATLAS experiment. Technical Report ATL-COM-PHYS-2014-035, CERN, Jan 2014.

[81] R. Barlow and C. Beeston. Fitting using finite monte carlo samples. *Computer Physics Communications*, 1993.

[82] Andreas Hoecker, Peter Speckmayer, Joerg Stelzer, Jan Therhaag, Eckhard von Toerne, and Helge Voss. TMVA: Toolkit for Multivariate Data Analysis. *PoS*, ACAT:040, 2007.

[83] F. Ahmadov et al. Supporting Document for the Search for the bb decay of the Standard Model Higgs boson in associated (W/Z)H production with the ATLAS detector. Technical report, CERN.

[84] J. Gallicchio, J. Huth, M. Kagan, M. D. Schwartz, K. Black, and B. Tweedie. Multivariate discrimination and the Higgs + W/Z search. *JHEP*, 04:069, 2011.

[85] ATLAS Collaboration. Search for the Standard Model Higgs boson produced in association with a vector boson and decaying to a $b\bar{b}$ pair in pp collisions at 13 TeV using the ATLAS detector. Technical Report ATLAS-CONF-2016-091, CERN, Aug 2016.

[86] Nikolaos Kidonakis. NNLL resummation for s-channel single top quark production. *Phys. Rev.*, D81:054028, Jan 2010.

[87] Nikolaos Kidonakis. Next-to-next-to-leading-order collinear and soft gluon corrections for t-channel single top quark production. *Phys. Rev.*, D83:091503, Mar 2011.

[88] Nikolaos Kidonakis. Two-loop soft anomalous dimensions for single top quark associated production with a W- or H-. *Phys. Rev.*, D82:054018, May 2010.

[89] Nikolaos Kidonakis. Differential and total cross sections for top pair and single top production. In *Proceedings, 20th International Workshop on Deep-Inelastic Scattering and Related Subjects (DIS 2012): Bonn, Germany, March 26-30, 2012*, May 2012.

[90] Emanuele Re. Single-top Wt-channel production matched with parton showers using the POWHEG method. *Eur. Phys. J.*, C71:1547, 2011.

[91] CMS Collaboration. Search for the standard model Higgs boson produced in association with a W or a Z boson and decaying to bottom quarks. *Phys. Rev.*, D89(1):012003, 2014.

[92] ATLAS Collaboration. Evidence for the $H \to b\bar{b}$ decay with the ATLAS detector. 2017.

[93] CMS Collaboration. Evidence for the Higgs boson decay to a bottom quark-antiquark pair. 2017.

List of Figures

List of Tables