

中国科学技术大学

博士学位论文



ATLAS 实验上的希格斯粒子与矢量玻色子协产生且希格斯粒子衰变为底夸克对过程的测量

作者姓名： 黄奕聪

学科专业： 粒子物理

导师姓名： 刘衍文 教授

完成时间： 二〇二二年七月二十五日



University of Science and Technology of China
A dissertation for doctor's degree



**Measurement of the Higgs boson produced
in association with a vector boson and
decaying to a pair of b -quarks with the
ATLAS detector**

Author: Yicong Huang

Speciality: Particle Physics

Supervisor: Prof. Yanwen Liu

Finished time: July 25, 2022

中国科学技术大学学位论文原创性声明

本人声明所呈交的学位论文，是本人在导师指导下进行研究工作所取得的成果。除已特别加以标注和致谢的地方外，论文中不包含任何他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的贡献均已在论文中作了明确的说明。

作者签名：_____

签字日期：_____

中国科学技术大学学位论文授权使用声明

作为申请学位的条件之一，学位论文著作权拥有者授权中国科学技术大学拥有学位论文的部分使用权，即：学校有权按有关规定向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅，可以将学位论文编入《中国学位论文全文数据库》等有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。本人提交的电子文档的内容和纸质论文的内容相一致。

保密的学位论文在解密后也遵守此规定。

☒ 公开 ☐ 保密（____年）

作者签名：_____

导师签名：_____

签字日期：_____

签字日期：_____

摘要

希格斯机制是标准模型中电弱对称性自发破缺的根源，解释了基本粒子的质量。所预言的希格斯粒子是标准模型中唯一的标量粒子。2012 年，ATLAS 和 CMS 实验分别在大型强子对撞机上发现了希格斯粒子。在过去的十年中，ATLAS 和 CMS 实验对希格斯粒子的性质的测量精度不断提高，检验了标准模型，同时探测超出标准模型的新物理。希格斯玻色子衰变为一对底夸克在所有希格斯衰变模式中具有最大的分支比，这使得它成为研究希格斯粒子与费米子耦合的最灵敏的衰变道之一。

本论文介绍了与 W 或 Z 玻色子协同产生并衰变为底夸克对的希格斯玻色子的产生过程的测量。分析的数据样本是 ATLAS 探测器于 2015 年至 2018 年在质心能量为 13 TeV 的质子-质子对撞中收集的，对应于 139 fb^{-1} 的积分亮度。测量主要关注高横动量的相空间，其中希格斯玻色子的衰变产物可以有效地被重建为大半径喷注 ($R = 1.0$)。这篇论文还介绍了大半径喷注的重建算法及其性能，特别是喷注能量和质量的校准。

ATLAS 实验已经开发了多种大半径喷注的重建算法。这些大半径喷注的性能是根据它们的重建，校准和标记的性能来评估的。综合性能最好的喷注被选为 ATLAS 实验的标准，并且推荐给所有的物理分析。对于 $VH(H \rightarrow b\bar{b})$ 分析所运用的大半径喷注，校准后的喷注能量达到 1% 的精度，校准后的喷注质量达到 10% 的精度。最先进的重建算法进一步提高了大半径喷注的校准精度，校准后的喷注能量达到小于 0.5% 的精度。

在大型强子对撞机的第一阶段运行 (2011 — 2012) 收集的数据中没有发现 $H \rightarrow b\bar{b}$ 衰变的证据。因为该衰变通道在希格斯粒子通过胶子-胶子融合产生时的纯强子末态的本底太高，其统计涨落淹没了信号。观测 $H \rightarrow b\bar{b}$ 衰变最灵敏的希格斯粒子产生模式是和 W 或 Z 玻色子的协同产生，其中矢量玻色子衰变到轻子。大型强子对撞机的第二阶段运行提高了对撞能量，也增加了亮度，使得 VH 产生模式有了足够的统计量。在 2015-2016 年期间收集的数据中首次观察到 $H \rightarrow b\bar{b}$ 衰变的证据，对应 36.1 fb^{-1} 的积分亮度。ATLAS 实验和 CMS 实验观测到的信号显著性分别是 3.5 倍标准差和 3.8 倍标准差。在这个分析中希格斯粒子的衰变产物被重建为两个标记为 b 的小半径喷注 ($R = 0.4$)。

进一步的测量运用 2015 年至 2018 年收集的数据继续提高了分析的灵敏度。该测量采用了两种互相独立的分析方法。其中一种方法主要关注中低横动量的相空间，希格斯粒子的衰变产物可以被有效重建为两个小半径喷注，以下称之为 $VH(b\bar{b})$ resolved 分析。另一种方法主要关注中高横动量的相空间，希格斯粒

子的衰变产物可以被重建为一个大半径喷注，以下称之为 $VH(b\bar{b})$ boosted 分析。 $VH(b\bar{b})$ resolved 分析用了和发现 $H \rightarrow b\bar{b}$ 衰变的分析相同的方法。事例选择要求末态有两个标记为 b 的小半径喷注。观测到的 ZH 信号显著性是 5.3 倍标准差， WH 信号显著性是 4.1 倍标准差。

高横动量区间对潜在的反常耦合更加灵敏。但是“resolved”结构对于重建高横动量的希格斯粒子的衰变产物不是非常有效。当希格斯粒子的横动量增加时，两个 b 喷注的夹角会变小，逐渐合并成一个大半径喷注。 $VH(b\bar{b})$ boosted 分析通过将用径迹重建的喷注和横动量最大的大半径喷注匹配来选择事例。其中两个用径迹重建的喷注需要能够被标记为 b 喷注。 VH 信号的提取是通过对大半径喷注的质量分布进行统计分析实现的。测量信号强度，也就是测量到的截面和标准模型预测的截面之比为 $0.72^{+0.39}_{-0.36}$ 。测量和预期到的信号的显著性分别为 2.1 和 2.7 倍标准差。在 W 或 Z 玻色子的两个横动量区域：250 – 400 GeV, > 400 GeV 中的信号事例数与标准模型预期一致。这些结果在有效场论的框架中限制了反常耦合。

在现有的 $VH(b\bar{b})$ 分析的基础上，对 $H \rightarrow b\bar{b}$ 的测量仍然可以从多方面改进。在 ATLAS 合作组一个综合 $VH(b\bar{b})$ resolved 和 boosted 方案的分析正在进行。该分析的目的是设计一个统一、优化的方法来提取最准确的测量结果。通过在 boosted 结构中引入多元分析，测量精度可以得到显著的提高。同时，联合分析对 b 喷注的标记算法和信号事例的选择方法进行了优化。

此外，以后可以通过大型强子对撞机第三阶段运行增加的亮度 (300 fb^{-1})，具有更好分辨率的大半径喷注，标记 b 喷注算法的优化以及更多重建算法的优化来提高 $H \rightarrow b\bar{b}$ 衰变的测量精度。

关键词：大型强子对撞机，ATLAS 实验，希格斯粒子，底夸克，大半径喷注

ABSTRACT

The Higgs mechanism causes the spontaneous electroweak symmetry breaking and generates masses for the elementary particles. The predicted particle, the Higgs boson, is the only scalar particle in the Standard Model (SM). The Higgs boson was discovered by the ATLAS and the CMS Collaborations at the Large Hadron Collider (LHC) in 2012. Over the past decade, the ATLAS and the CMS experiments continued to improve the precision of the measurements of the Higgs boson properties. The measurements tested the Standard Model predictions while probing new physics beyond the SM. The Higgs boson decaying to a pair of b -quarks has the largest branching ratio among all the Higgs decay modes, which makes it one of the most sensitive channel to study the Yukawa nature of the Higgs coupling to the fermions.

This thesis presents a measurement for a Higgs boson produced in association with a W or Z boson and decaying to a $b\bar{b}$ pair. The analyzed data samples, corresponding to an integrated luminosity of 139 fb^{-1} , were collected in proton-proton collisions at the LHC at a center-of-mass energy of 13 TeV from 2015 to 2018 (Run 2) by the ATLAS detector. The measurement focuses on high vector boson transverse momentum phase spaces where the decay products of the Higgs boson can be effectively reconstructed as a large-radius jet ($R = 1.0$). This thesis also presents the reconstruction of the large-radius jet collections, in particular the jet energy and mass scale calibration.

Several large- R jet definitions have been developed at the ATLAS experiment. These jet collections were evaluated based on the performance of reconstruction, calibration and tagging. The jet collection with the best overall performance becomes the baseline recommended to all the physics analyses. For the large- R jet collection used in the $VH(H \rightarrow b\bar{b})$ measurement, the jet energy scale is calibrated to 1% precision level and the jet mass scale is calibrated to 10% precision level. The calibration performance has been improved with the latest large- R jet collection. The non-closure of the calibrated jet energy scale is smaller than 0.5%.

The $H \rightarrow b\bar{b}$ decay was not observed in Run 1 (2011 - 2012) of the LHC because this channel has an all-hadronic final state through the gluon-gluon fusion production mode. It is very difficult to select signal events from the overwhelming multijet background at the LHC. The most sensitive production mode to observe $H \rightarrow b\bar{b}$ decay is the associated production of a Higgs boson and a W or Z boson, where the vector boson decays leptonically. The increased colliding energy and luminosity in Run 2 enabled

sufficient statistics of VH production. The evidence of the $H \rightarrow b\bar{b}$ decay was first observed in data collected during 2015-2016, corresponding to an integrated luminosity of 36.1 fb^{-1} . The signal was observed with significances of 3.5 and 3.8 standard deviations by the ATLAS and the CMS experiments, respectively. The decay products of the Higgs boson were reconstructed as two b -tagged small- R jets ($R = 0.4$) in the analysis.

Further measurements continued to improve the sensitivity of the analysis with full Run 2 dataset. The measurements have been conducted through two independent approaches. One of the approaches mainly targets low to medium transverse momentum phase spaces where the decay products of the Higgs boson can be reconstructed as two small- R jets effectively, referred to as the $VH(b\bar{b})$ resolved analysis. The other approach mainly targets medium to high transverse momentum phase spaces where the decay products of the Higgs boson can be reconstructed as a large- R jet, referred to as the $VH(b\bar{b})$ boosted analysis. The analysis presented in this thesis follows the second approach. The $VH(b\bar{b})$ resolved analysis had the same strategy as the original measurement that observed $H \rightarrow b\bar{b}$ decay. The events were selected by requiring two b -tagged small- R jets in the final states. The ZH signal was observed with a significance of 5.3 standard deviations and the WH signal was observed with a significance of 4.1 standard deviations.

The high transverse momentum phase spaces can be more sensitive to potential anomalous couplings. Moreover, the resolved topology is not very efficient for reconstruction of the decay products of the Higgs boson in high p_T regions. The angle between the two b -jets becomes smaller with larger Higgs boson p_T , and the two jets merge into a large- R jet. In the $VH(b\bar{b})$ boosted analysis, the events were selected by matching jets reconstructed from tracks (track jets) to the leading large- R jet. Two of the track jets were required to be b -tagged. The VH signal was measured through a maximum profiled likelihood fit to the large- R jet mass distributions. The signal strength, defined as the ratio of the observed cross section and the Standard Model predicted cross section, was measured as $0.72^{+0.39}_{-0.36}$. The signal was observed (expected) with a significance of 2.1 (2.7) standard deviations. The cross sections of the associated production of a Higgs boson and a W or Z boson were measured in two transverse momentum regions, $250 - 400 \text{ GeV}$ and above 400 GeV . All the measurements are compatible with the Standard Model predictions. The results were interpreted in the Effective Field Theory as constraints on anomalous couplings.

After the success of the $VH(b\bar{b})$ resolved and boosted analysis, the measurement

of $H \rightarrow b\bar{b}$ decay can still be improved from many aspects. The $VH(b\bar{b})$ resolved and boosted analyses have been combined into an analysis that aims to design one coherent, harmonized approach to extract $VH(b\bar{b})$ results and extend the physics output. This analysis is in progress at the time of the publish of this thesis. The sensitivity of the measurement has been improved significantly through the multivariate analysis introduced to the boosted topology. The b -tagging algorithms and the Higgs tagging strategy are also optimized together with the multivariate analysis.

Furthermore, the measurement of $H \rightarrow b\bar{b}$ decay will benefit from the increased luminosity in Run 3 of the LHC (expected 300 fb^{-1}), the large- R jet collection with better resolution and future improvements of b -tagging, object reconstruction.

Key Words: LHC; ATLAS experiment; Higgs boson; VH production; b -quark; large- R jet

Contents

Introduction	15
1 The theoretical overview of the Higgs boson	19
1.1 The Standard Model	19
1.1.1 Quantum Electrodynamics	21
1.1.2 Non-Abelian gauge theory	22
1.1.3 The Higgs mechanism	26
1.2 Higgs boson phenomenology at the LHC	28
1.2.1 Higgs boson properties	29
1.2.2 Higgs boson production	30
1.2.3 Higgs boson decay	30
1.2.4 $VH(H \rightarrow b\bar{b})$ channel	36
1.3 Beyond the Standard Model	37
1.3.1 Simplified Template Cross Sections	38
1.3.2 Effective Field theories	40
2 The Large Hadron Collider and the ATLAS experiment	43
2.1 The Large Hadron Collider	43
2.2 The ATLAS experiment	45
2.2.1 The coordinates and magnet system	46
2.2.2 Inner Detector	48
2.2.3 Calorimeters	50
2.2.4 Muon Spectrometers	53
2.2.5 Trigger and data acquisition system	56
3 Object reconstruction	59
3.1 Charged particle track reconstruction	59
3.2 Primary vertex	63
3.3 Electron reconstruction	63
3.3.1 Electron identification	66
3.3.2 Electron isolation	67
3.4 Muon reconstruction	68
3.4.1 Muon identification	69
3.4.2 Muon reconstruction and identification efficiency	70
3.5 Jet reconstruction	71
3.5.1 Reconstruction of clusters	71
3.5.2 Jet clustering	77

Contents

3.5.3	Jet Calibration	78
3.6	Missing transverse energy reconstruction	82
4	Calibration of reconstructed jet energy and mass	85
4.1	Jet energy calibration	85
4.2	Jet mass calibration	87
4.3	Performance of the calibration	89
4.3.1	$R = 1.0$ jets reconstructed from topo-clusters	89
4.3.2	$R = 1.0$ jets reconstructed from Track Calo Clusters	92
4.3.3	$R = 1.0$ jets reconstructed from Unified Flow Objects	93
5	Measurement of boosted $VH(H \rightarrow b\bar{b})$ process	97
5.1	Data samples and simulated event samples	97
5.1.1	Dataset and triggers	97
5.1.2	Simulated events	99
5.2	Event Selection	106
5.2.1	Higgs boson identification	106
5.2.2	Baseline event selection	110
5.2.3	Event categorization	112
5.2.4	Large- R jet mass	114
5.2.5	Jet substructure variables	115
5.3	Systematic uncertainties	119
5.3.1	Experimental systematic uncertainties	119
5.3.2	Modeling uncertainties	121
5.4	Statistical analysis	135
5.4.1	The Likelihood function	135
5.4.2	The likelihood ratio test	136
5.4.3	Fit inputs	137
5.4.4	Treatment of the nuisance parameters	138
5.5	Results of boosted $VH(H \rightarrow b\bar{b})$ production	144
5.5.1	Post-fit m_J distributions	144
5.5.2	Signal strength and significance	144
5.5.3	Cross section measurements	147
5.5.4	Effective Field Theory interpretations	149
5.6	Future improvements of $H \rightarrow b\bar{b}$ measurements	153
5.6.1	Selection of Higgs boson candidate	153
5.6.2	Multivariate discriminants	157
6	Conclusions	161
	Appendix	167
A	MC calibration of $R = 0.4$ jets in low-μ environment	167
B	Nuisance parameter pull plots	171

Introduction

The Standard Model (SM) of particle physics [1] has been the most successful theory since 1960s. It provides an accurate description of all the known particles and three of the four fundamental interactions: strong, weak and electromagnetic forces. The reliability of the SM has been tested in many experiments and the last missing piece, the Higgs boson was discovered by the ATLAS and CMS Collaborations at the Large Hadron Collider (LHC) at CERN in 2012 [2, 3]. The LHC [4] is the largest and the most powerful particle accelerator in the world. The protons are accelerated to very high energy and then collide. The collisions produce an environment similar to 10^{-12} s after the Big Bang. The ATLAS and CMS are two general-purpose detectors that investigate a wide range of physics, including two areas: precision tests of the SM and search for physics Beyond the Standard Model (BSM).

The Higgs boson was discovered in the data collected at proton-proton collisions in Run 1, with center-of-mass energies at 7 and 8 TeV. The major decay channels like $H \rightarrow \gamma\gamma$, $H \rightarrow ZZ^*$ and $H \rightarrow WW^*$ were observed through the most abundant production modes, the gluon-gluon fusion (ggF), the vector boson fusion (VBF), the associated production with a W or Z boson (VH) and the associated production with a pair of top quarks (ttH). The main production modes of the Higgs boson are shown in Figure 0.1. The particle that was observed had a mass around 125 GeV [5]. After the discovery, the properties of this particle have been measured. The observation of the Higgs boson decaying to vector bosons permits measurements of couplings with improved precisions between the Higgs boson and the vector bosons. The results agree with the SM predictions for a Higgs boson with this mass. However, the couplings between the Higgs boson and fermions predicted by the SM need to be studied directly in channels where the Higgs boson decays to fermions, which is crucial to tests of the properties of the Higgs sector.

The dominant decay channel $H \rightarrow b\bar{b}$, with the largest branching ratio of around 58% for Higgs boson mass $m_H = 125$ GeV, was not well studied in Run 1. The ggF production mode and the VBF

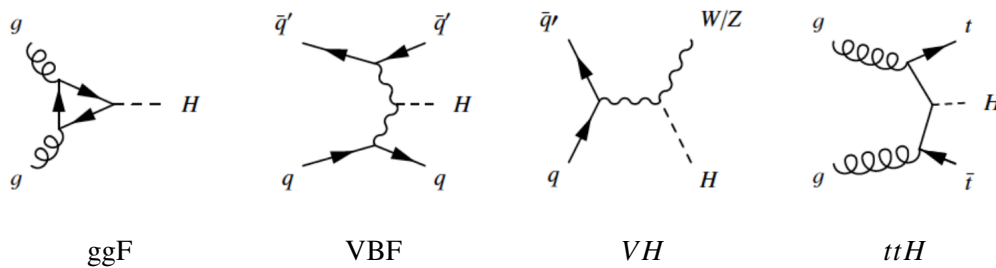


Figure 0.1: Main production modes of the Higgs boson at the LHC.

production mode take up to 94% of the total Higgs production cross section. Both the ggF and the VBF production, followed by $H \rightarrow b\bar{b}$ decay, result in an all-hadronic final state. It is very challenging to select decays with such final states in the overwhelming backgrounds from the multi-jet QCD processes at hadron colliders. The most sensitive production mode to study $H \rightarrow b\bar{b}$ decays is the VH production, where the vector boson decays leptonically. Despite its cross section one order of magnitude lower than the ggF production, the leptonic decay of the vector boson provides clean signature to select signal events from the multi-jet background. The $VH(H \rightarrow b\bar{b})$ channel is not only important to test the SM, it also has large potential to reveal new physics. It allows to study the Higgs boson at high transverse momentum, which provides enhanced sensitivity to BSM physics, and the cross section of VH production can even be larger than the ggF one at high p_T .

The analyses described in this thesis focus on the search for $VH(H \rightarrow b\bar{b})$ process in high transverse momentum regions, with data collected at $\sqrt{s} = 13$ TeV pp collisions in Run 2. The Higgs boson is reconstructed using $R = 1.0$ large radius jet, the so-called large- R jet. This technique has already been used in searches for BSM physics, but not in measurement of the Higgs boson production cross sections before. The use of large- R jet allowed exploration of high p_T phase space which is particularly sensitive to BSM physics. As an important ingredient to the physics analysis, the energy and mass of the large- R jet need to be reconstructed accurately. The derivation of several sets of large- R jet energy and mass calibration factors is also described in this thesis.

Structure of the thesis and personal contribution

This thesis is divided into six chapters. Chapter 1 gives an overview of the SM and the Higgs mechanism, highlighting $VH(H \rightarrow b\bar{b})$ process from the theoretical and experimental viewpoints. Despite the success of the SM in describing the observed particles, there are still limitations and some methods to explore potential BSM phenomena through $VH(H \rightarrow b\bar{b})$ process are introduced.

Chapter 2 describes the design and performance of the LHC and the ATLAS experiment. The trigger and data acquisition system is also introduced in this chapter.

Chapter 3 gives an overview of the reconstruction in ATLAS. Objects like leptons, jets, particle tracks and missing transverse momentum are needed to reconstruct the final state of the physics processes. The methods to reconstructed these objects from "low-level" electrical signals are described.

Calibration of the large- R jet is a necessary procedure after the reconstruction because a fraction of energy in a hadronic shower is invisible to the detector. The jet energy scale (JES) calibration is derived through comparing the reconstructed energy to truth scale energy in MC simulated QCD di-jet samples, and then correct all the reconstructed energy to MC truth scale. I have worked on calibration of several large- R jets and small- R jets. The derivation of the calibrations are presented in Chapter 4.

Chapter 5 describes the $VH(b\bar{b})$ analysis. Section 5.1 reports the data samples and simulated event samples used in the analysis. The triggers used in pre-selection of the data samples are described as well. The signal and background samples are simulated with several Monte Carlo (MC) generators, the simulation of every process is introduced.

Event selection and categorisation of the analysis is described in Section 5.2. I have contributed to the optimization of the b-tagging working point in the event selection, and studied the potential

improvement of several large- R jet substructure variables. The substructure variables are able to improve the event selection but it did not enter the final event selection due to the lack of properly derived systematic uncertainty in time for the publication.

The signal cross section and statistical significance are extracted from a maximum profile likelihood fit to data and a log-likelihood ratio test after event selection. Section 5.3 describes the source of systematic uncertainties. The statistical procedure is described in Section 5.4.

Section 5.5 presents the results of the $VH(b\bar{b})$ analysis in high p_T regions. VH signal strength is measured from the likelihood fit performed with the large- R jet mass m_J as discriminant variable. The background $VZ(Z \rightarrow b\bar{b})$ is a "signal-like" background. Its signal strength is also measured in the fitting as a validation of the analysis method. The cross section of $VH(H \rightarrow b\bar{b})$ is measured within the Simplified Template Cross Section (STXS) framework. I contributed to the statistical analysis, especially in 0-lepton channel.

Measurement of $H \rightarrow b\bar{b}$ process has potential to be improved. Section 5.6 is dedicated to the on-going $VH(b\bar{b})$ legacy analysis that combines the two stand-alone $VH(b\bar{b})$ analyses: resolved and boosted. "Resolved" denotes a $VH(b\bar{b})$ analysis that mainly interested in lower p_T regions, where the two b-jets can be separated clearly. "Boosted" denotes the $VH(b\bar{b})$ analysis in high p_T regions. Small- R jet with better resolution and a better b-tagging algorithm are used compared to the stand-alone $VH(b\bar{b})$ analyses. The boosted analysis is also improved with optimized Higgs tagging strategy and the use of MVA training, which I have contributed to especially in 0-lepton channel.

Finally, Chapter 6 summarizes the work presented in this thesis and provides some personal remarks on the future perspectives.

1 The theoretical overview of the Higgs boson

1.1 The Standard Model

The Standard Model (SM) of particle physics [1, 6] is a quantum field theory that accurately describes all the known elementary particles and three of the four fundamental interactions in the universe: strong, weak and electromagnetic forces. Only the gravitational force does not have a quantum mechanical formulation, even though it is accurately described by the theory of general relativity. The gravitational force is very weak compared to the other three fundamental interactions. Its effect is negligible at the scale of particles. Therefore, the SM has provided a number of precise and successful predictions for strong and electroweak interaction phenomena.

The Lagrangian of the SM is invariant under the local group $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$ transformation, where the $SU(3)_C$ group corresponds to the strong interactions, the $SU(2)_L \otimes U(1)_Y$ group corresponds to the electroweak interaction. In the SM, the elementary constituents of matter are fermions of spin-1/2, the fundamental interactions are mediated by the fields of bosons of integer spin. Especially, the field of the spin-0 scalar Higgs boson causes the spontaneous electroweak symmetry breaking and generates masses for the elementary particles.

There are three generations of quarks with six flavors: up (u), down (d), charm (c), strange (s), top (t), bottom (b) [7]. The quarks have fractional electric charge, which is $+2/3$ for u, c, t , $-1/3$ for d, s, b . Each quark has an anti-particle with the opposite charge: $-2/3$ for $\bar{u}, \bar{c}, \bar{t}$, $+1/3$ for $\bar{d}, \bar{s}, \bar{b}$. The quarks can interact through strong, electromagnetic and also weak forces. The strong interactions between quarks, mediated by the gluons (g), is described by Quantum Chromodynamics (QCD). The quarks carry "color charge" (C) which comes in three colors: red (R), green (G), blue (B). The three colors form the elements of the $SU(3)_C$ group corresponding to the strong interactions. The $SU(3)$ gauge symmetry of the group corresponds to a conserved quantity (Noether's theorem) [8], which is the color charge. The nature of the strong interaction includes "color confinement". Colored objects must be confined to form neutrally colored particles called hadrons and cannot propagate as free quarks, because separating a free quark from a hadron requires infinite amount of energy. The most probable bound states of quarks are the half-integer spin three-quark (baryons) and the integer spin quark-antiquark (mesons) combinations that meet the neutral colored requirement. More complicated states are allowed, like the two quark two anti-quark states (tetraquark) [9] and the three quark plus quark-antiquark states (pentaquark) [10]. A few of them have been observed in experiments recent years.

The leptons are also arranged in three generations in the SM [11]. The charged leptons are electron (e^-), muon (μ^-), tauon (τ^-) with their anti-particles e^+, μ^+, τ^+ . Leptons that do not carry electric charge are the neutrinos, ν_e, ν_μ, ν_τ with their anti-particles $\bar{\nu}_e, \bar{\nu}_\mu, \bar{\nu}_\tau$. Particles that carry electric charge interact through electromagnetic forces, mediated by the photons (γ), which is described by

Standard Model of Elementary Particles

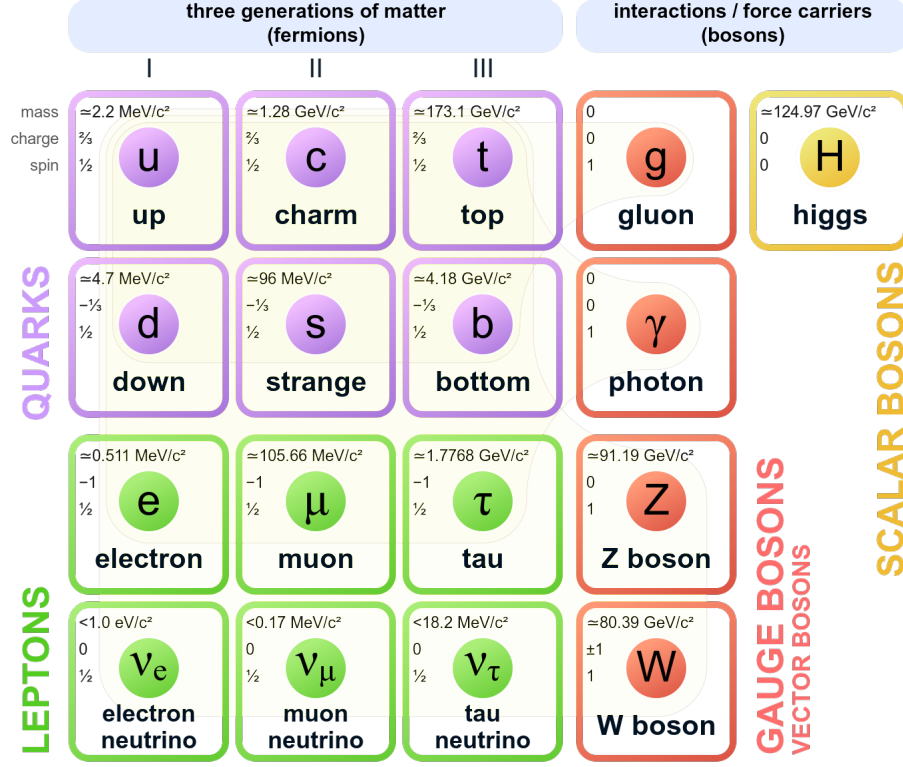


Figure 1.1: Summary of the SM elementary particles and their properties. Three generations of quarks and leptons, vector bosons and the Higgs boson are shown [14].

Quantum Electrodynamics (QED). Besides the electromagnetic interaction, all the leptons interact through weak forces. Three kinds of vector bosons W^+ , W^- and Z mediate the weak interaction. With the $U(1)_Y$ part of the gauge symmetry, the quantity known as "weak hypercharge" (Y_W) relating to the electric charge and the third component of weak isospin, is conserved in electroweak interactions. The $SU(2)_L$ gauge symmetry has a suffix "L", as the weak vector bosons only couple to left-handed fermions and their right-handed antiparticles. The three generations of quarks and leptons have the same gauge structure but with different masses. The $SU(2)_L \otimes U(1)_Y$ symmetry forbids mass terms in the Lagrangian. The Higgs boson (H) provides the mechanism that causes the spontaneous electroweak symmetry breaking and generates masses for other particles, which is the so-called "Higgs mechanism" [12, 13]. It is further described in the following sections. Figure 1.1 summarizes the elementary particles and their properties in the SM.

Quantum field theory [15] is the application of quantum mechanics and special relativity to dynamical systems of fields. From the field viewpoint, the elementary particles are described by relativistic fields. The fundamental law of mechanics is the principle of least action. The action is defined as the time integral of the Lagrangian L . In a local field theory the Lagrangian can be written as the spatial

integral of a Lagrangian density $\mathcal{L}$, which is a function of one or more fields ϕ and their derivatives $\partial_\mu \phi$.

$$S = \int L dt = \int \mathcal{L}(\phi, \partial_\mu \phi) d^4x. \quad (1.1)$$

The Euler-Lagrange equation of motion for a field can be derived from the principle of least action:

$$0 = \delta S = \int d^4x \left\{ \frac{\partial \mathcal{L}}{\partial \phi} \delta \phi - \partial_\mu \left(\frac{\partial \mathcal{L}}{\partial (\partial_\mu \phi)} \right) \delta \phi + \partial_\mu \left(\frac{\partial \mathcal{L}}{\partial (\partial_\mu \phi)} \delta \phi \right) \right\}. \quad (1.2)$$

The last term can be turned into a surface integral over the boundary of the four-dimensional space time region, which is zero as the initial and final field configurations are given, $\delta \phi$ vanishes on the spatial boundary. For the first two terms, to keep the integral vanishes for any $\delta \phi$, the integrand has to be zero every where,

$$\partial_\mu \left(\frac{\partial \mathcal{L}}{\partial (\partial_\mu \phi)} \right) - \frac{\partial \mathcal{L}}{\partial \phi} = 0. \quad (1.3)$$

As the fundamental laws can be put in the form of the principle of least action, this principle should work with relativistic fields. It means for a field ϕ , after a rotation or boost to a different frame of reference, the equation of motion can still be derived from the Euler-Lagrangian equation Eq. 1.3. If a field is relativistically invariant, then the form of its equation of motion should not change in a different frame of reference. This is automatically satisfied if the Lagrangian is a Lorentz scalar. Thus for a field that can exist in nature, its Lagrangian has to be a Lorentz scalar.

1.1.1 Quantum Electrodynamics

The theory of QED [16–19] describes electromagnetic interaction between charged fermions and photons. The QED Lagrangian can be written as:

$$\mathcal{L}_{\text{QED}} = \mathcal{L}_{\text{Dirac}} + \mathcal{L}_{\text{Maxwell}} + \mathcal{L}_{\text{Int}}. \quad (1.4)$$

The first part is the Lagrangian of Dirac field,

$$\mathcal{L}_{\text{Dirac}} = \bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi, \quad (1.5)$$

where ψ and $\bar{\psi} = \psi^\dagger \gamma^0$ are the fields of the fermions, γ^μ are the Dirac matrices from gauge symmetry, m is a scalar corresponds to the mass of the fermions. The equation of motion of the Dirac field can be derived using Eq. 1.3, which is the Dirac equation.

$$(i\gamma^\mu \partial_\mu - m)\psi = 0, \quad -i\partial_\mu \bar{\psi} \gamma^\mu - m\bar{\psi} = 0. \quad (1.6)$$

The solution of the Dirac equation are spinors, composed by two sets of creation and annihilation operators that obey the anticommutation rule, corresponding to fermions and anti-fermions. The angular momentum of fermions is $\pm 1/2$, while for anti-fermions it is $\mp 1/2$.

The second part is the Lagrangian of electromagnetic field,

$$\mathcal{L}_{\text{Maxwell}} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} = -\frac{1}{4} (\partial_\mu A_\nu - \partial_\nu A_\mu)^2, \quad (1.7)$$

1 The theoretical overview of the Higgs boson

where A_μ is the electromagnetic vector potential, $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ is the electromagnetic field tensor.

The third part is the Lagrangian that describes interactions between fermions and photons,

$$\mathcal{L}_{\text{Int}} = -e\bar{\psi}\gamma^\mu\psi A_\mu, \quad (1.8)$$

where e is the electron charge.

A crucial property of the QED Lagrangian is that it needs to be invariant under $U(1)$ transformation:

$$\psi(x) \rightarrow e^{i\alpha(x)}\psi(x), \quad A_\mu \rightarrow A_\mu - \frac{1}{e}\partial_\mu\alpha(x). \quad (1.9)$$

In fact, the QED Lagrangian can be written as a simpler form:

$$\mathcal{L}_{\text{QED}} = \bar{\psi}(i\gamma^\mu D_\mu - m)\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}, \quad (1.10)$$

where D_μ is the gauge covariant derivative,

$$D_\mu = \partial_\mu + ieA_\mu(x). \quad (1.11)$$

The QED Lagrangian is a simple example of gauge symmetry theory that can describe the electromagnetic interactions between charged fermions. It is invariant under $U(1)$ transformation, which corresponds to a conserved quantity, the electric charge. From the modern viewpoint, gauge symmetry is a fundamental principle of the fields, so the form of Lagrangian can be constructed through the gauge symmetry that it follows. Similar to QED, all the interactions are described by gauge theories in the SM, while the gauge symmetry for QCD is $SU(3)$, and for electroweak is $SU(2) \otimes U(1)$.

1.1.2 Non-Abelian gauge theory

The QED Lagrangian can be constructed from terms that are invariant under $U(1)$ transformation. More general field theories can be constructed through any continuous symmetry group, the first of such field theories was constructed by Yang and Mills [20, 21]. Starting with a doublet of Dirac fields,

$$\psi = \begin{pmatrix} \psi_1(x) \\ \psi_2(x) \end{pmatrix}, \quad (1.12)$$

and the $SU(2)$ transformation, which is essential to the symmetry of electroweak interaction,

$$\psi \rightarrow \exp\left(i\alpha^i(x)\frac{\sigma^i}{2}\right)\psi, \quad (1.13)$$

where σ^i are the 2×2 Pauli matrices, $\alpha^i(x)$ are components of a three dimensional arbitrary function of x . This transformation is different from a rotation in physical space. Yang and Mills considered

1.1 The Standard Model

(ψ_1, ψ_2) to be proton-neutron doublet transformed under isotopic spin in their original work. To keep the Lagrangian invariant under the $SU(2)$ transformation, the covariant derivative is found as:

$$D_\mu = \partial_\mu - igA_\mu^i \frac{\sigma^i}{2}, \quad (1.14)$$

here A_μ^i are three vector boson fields, one for each generator of the transformation group, and g is a constant extracted for convenience, similar to the electron charge e in QED.

To write down the complete Lagrangian, gauge-invariant terms that depend only on A_μ^i are also needed, which corresponds to the kinetic energy term of the vector boson field. It is found as:

$$\mathcal{L} = -\frac{1}{4}F_{\mu\nu}^i F^{\nu\mu i}, \quad (1.15)$$

$$F_{\mu\nu}^i = \partial_\mu A_\nu^i - \partial_\nu A_\mu^i + g\epsilon^{ijk} A_\mu^j A_\nu^k, \quad (1.16)$$

where $F_{\mu\nu}^i$ is the field tensor of field A_μ^i , and ϵ^{ijk} is the totally antisymmetric tensor with $\epsilon^{123} = 1$ [22]. The Lagrangian of a Yang-Mills vector field can be constructed by adding the gauge field Lagrangian Eq. 1.15 to the Dirac Lagrangian, with the ordinary derivative ∂_μ replaced by the covariant derivative D_μ ,

$$\mathcal{L}_{\text{Yang-Mills}} = \bar{\psi}(i\gamma^\mu D_\mu)\psi - \frac{1}{4}F_{\mu\nu}^i F^{\mu\nu i} - m\bar{\psi}\psi, \quad (1.17)$$

this is the Yang-Mills Lagrangian, it is the simplest example of a non-Abelian gauge theory.

Chromodynamics

The theory of QCD [7, 23–26] describes the strong interactions between quarks. It is a non-Abelian gauge theory with the color $SU(3)$ gauge symmetry. The color charge is the gauge quantum numbers of quarks. With same steps as the Yang-Mills theory, considering a triplet of Dirac fields, each field operator ψ_X corresponds to the quarks carrying color R , G or B .

$$\psi = \begin{pmatrix} \psi_R \\ \psi_G \\ \psi_B \end{pmatrix}. \quad (1.18)$$

The $SU(3)$ transformation is

$$\psi(x) \rightarrow \exp\left(i\alpha_a(x)\frac{\lambda_a}{2}\right)\psi(x), \quad (1.19)$$

where λ_a are eight 3×3 Gell-Mann matrices corresponding to gluons of color octet, $\alpha_1(x), \dots, \alpha_8(x)$ are eight arbitrary functions of x .

The covariant derivative that can keep the Lagrangian invariant under $SU(3)$ transformation is

$$D_\mu = \partial_\mu - ig_s G_\mu^a \frac{\lambda_a}{2}, \quad (1.20)$$

1 The theoretical overview of the Higgs boson

where G_μ^a $a = 1, \dots, 8$, are eight vector boson fields corresponding to the eight massless gluons, and g_s is the coupling constant of strong interaction.

The kinetic energy term of the gluonic field is

$$\mathcal{L} = -\frac{1}{4}G_{\mu\nu}^a G^{\nu\mu a},$$

$$G_{\mu\nu}^a = \partial_\mu G_\nu^a - \partial_\nu G_\mu^a + g_s f^{abc} G_\mu^b G_\nu^c,$$

here f^{abc} are the structure constants of the SU(3) group.

The complete Lagrangian of QCD is given by:

$$\mathcal{L}_{\text{QCD}} = \sum_{i,j \in R,G,B} \bar{\psi}_i (i\gamma^\mu D_\mu - m\delta_{ij}) \psi_j - \frac{1}{4}G_{\mu\nu}^a G^{\nu\mu a}, \quad (1.21)$$

$$D_\mu = \delta_{ij}\partial_\mu - ig_s G_\mu^a \frac{\lambda_{ij}^a}{2}. \quad (1.22)$$

This is the QCD part of the SM Lagrangian. The sum runs over all the color charge carried by quarks.

Electroweak Interaction

A model with $SU(2)_L \otimes U(1)_Y$ gauge symmetry introduced by Glashow, Weinberg and Salam (GWS) gives a unified description of the electromagnetic interaction and the weak interaction [27–29]. Starting from the $SU(2)$ Yang-Mills theory, the quantum number needed to describe weak interaction is the weak isospin (I_W). It is not real spin, but a quantum number related to the properties of the weak interaction. A very useful tool to study the $SU(2)$ symmetry is the right-handed and left-handed Dirac fields, defined as

$$\psi_L = \frac{1 - \gamma^5}{2}\psi, \quad \psi_R = \frac{1 + \gamma^5}{2}\psi,$$

$$\psi = \psi_L + \psi_R,$$

where $\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3$, $\bar{\psi}_L$ and $\bar{\psi}_R$ can be calculated using the definition $\bar{\psi} = \psi^\dagger\gamma^0$,

$$\bar{\psi}_L = \bar{\psi} \frac{1 + \gamma^5}{2}, \quad \bar{\psi}_R = \bar{\psi} \frac{1 - \gamma^5}{2},$$

$$\bar{\psi} = \bar{\psi}_L + \bar{\psi}_R,$$

notice $(\gamma^5)^\dagger = \gamma^5$, $(\gamma^5)^2 = 1$ and $\{\gamma^0, \gamma^5\} = 0$ are used in the calculation. Then the Lagrangian of free fermions can be expressed with the left-handed and right-handed fields.

$$\mathcal{L}_{\text{Dirac}} = \bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi$$

$$= \bar{\psi}_L i\gamma^\mu \partial_\mu \psi_L + \bar{\psi}_R i\gamma^\mu \partial_\mu \psi_R - m(\bar{\psi}_R \psi_L + \bar{\psi}_L \psi_R).$$

In the weak isospin space, ψ_L corresponds to left-handed fermions, ψ_R corresponds to right-handed fermions.

1.1 The Standard Model

As measured in the experiments, $W^\pm$ bosons only couple to left-handed fermions and right-handed anti-fermions, which means the $SU(2)$ transformation only affect left-handed fermions and their anti-particles. The left-handed fermions have weak isospin of $I_W = 1/2$, the third component of weak isospin for "up" type fermions is $I_3 = +1/2$, and for "down" type fermions is $I_3 = -1/2$. The right-handed anti-fermions have opposite I_3 . Left-handed fermions in the same generation can be put as a weak isospin doublet under $SU(2)$ transformation. The fields of generation i leptons and quarks are arranged as:

$$\Psi_L = \begin{pmatrix} \psi_L(v_i) \\ \psi_L(l_i^-) \end{pmatrix}, \begin{pmatrix} \psi_L(u_i) \\ \psi_L(d_i') \end{pmatrix}, \quad (1.23)$$

where $d_i' = \sum_j V_{ij} d_j$, V is the Cabibbo-Kobayashi-Maskawa (CKM) mixing matrix [30].

Right-handed fermions are weak isospin singlet,

$$\psi_R = \psi_R(u_i), \psi_R(d_i), \psi_R(l_i^-), \quad (1.24)$$

so they are not affected by the $SU(2)_L$ transformation. No right-handed neutrinos and left-handed anti-neutrinos have ever been observed.

However, this $SU(2)$ model forbids mass terms of the gauge bosons in the Lagrangian, while the weak gauge bosons, $W^\pm$ and Z , are massive. An additional scalar field ϕ that transforms as a spinor is proposed to solve the problem of massless particles and spontaneously break the symmetry. This is further described in Section 1.1.3 as the Higgs mechanism. A theory with $SU(2)$ symmetry is found to have two massive gauge bosons and one massless boson, which is not the case for weak interaction. Therefore, introducing an additional $U(1)$ gauge symmetry, the complete transformation becomes:

$$\Psi_L \rightarrow \exp\left(i\alpha^a(x)\frac{\sigma^a}{2}\right) \cdot \exp(i\beta/2)\Psi_L. \quad (1.25)$$

The covariant derivative is

$$D_\mu = \partial_\mu - i\frac{1}{2}gW_\mu^a\tau_L^a - i\frac{1}{2}g'Y_W B_\mu, \quad (1.26)$$

where W_μ^a and B_μ are fields of the $SU(2)$ and $U(1)$ gauge bosons, τ_L^a are the Pauli matrices, with subscript L to indicate they only act on left-handed fermions, g and g' are coupling strength of the $SU(2)$ and $U(1)$ fields, Y_W is the $U(1)$ charge, which is the weak hypercharge defined as:

$$Y_W = 2(Q - I_3). \quad (1.27)$$

The fermion field term can then be expressed with left-handed and right-handed fermions written separately,

$$\mathcal{L} = \sum_{\Psi_L} \bar{\Psi}_L i\gamma^\mu (\partial_\mu - i\frac{1}{2}gW_\mu^a\tau_L^a - i\frac{1}{2}g'Y_W B_\mu) \Psi_L + \sum_{\psi_R} \bar{\psi}_R i\gamma^\mu (\partial_\mu - i\frac{1}{2}g'Y_W B_\mu) \psi_R. \quad (1.28)$$

The kinetic energy term of the vector boson fields is

$$\mathcal{L} = -\frac{1}{4}W_{\mu\nu}^a W^{\nu\mu a} - \frac{1}{4}B_{\mu\nu} B^{\nu\mu}, \quad (1.29)$$

$$W_{\mu\nu}^a = \partial_\mu W_\nu^a - \partial_\nu W_\mu^a + g\epsilon^{abc}W_\mu^b W_\nu^c, \quad B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu.$$

1 The theoretical overview of the Higgs boson

The field tensors $W_{\mu\nu}^a$ and $B_{\mu\nu}$ are constructed from W_μ^a and B_μ for $SU(2)$ and $U(1)$ gauge symmetry.

The vector bosons are linear combinations of the gauge fields,

$$W_\mu^+ = \frac{1}{\sqrt{2}}(W_\mu^1 - iW_\mu^2), \quad W_\mu^- = \frac{1}{\sqrt{2}}(W_\mu^1 + iW_\mu^2),$$

$$Z_\mu = -B_\mu \sin \theta_W + W_\mu^3 \cos \theta_W, \quad A_\mu = B_\mu \cos \theta_W + W_\mu^3 \sin \theta_W,$$

where θ_W is the weak mixing angle defined as:

$$\cos \theta_W = \frac{g}{\sqrt{g^2 + g'^2}}, \quad \sin \theta_W = \frac{g'}{\sqrt{g^2 + g'^2}}. \quad (1.30)$$

There are three massive vector bosons corresponding to the weak vector boson W^+ , W^- and Z , and they acquire masses through the Higgs mechanism. The fourth vector boson γ remain massless, corresponds to photon.

The complete form of electroweak Lagrangian can be written as:

$$\mathcal{L}_{EW} = \sum_{\Psi_L, \Psi_R} \bar{\Psi} i \gamma^\mu D_\mu \Psi - \frac{1}{4} W_{\mu\nu}^a W^{\nu\mu a} - \frac{1}{4} B_{\mu\nu} B^{\nu\mu}. \quad (1.31)$$

There is no fermion mass term in this Lagrangian, because all possible such terms are forbidden by local gauge invariance. As the field ψ_L and ψ_R belong to different $SU(2)$ representations, and they have different $U(1)$ charge, any term that contains $\bar{\psi}_R \psi_L$ and $\bar{\psi}_L \psi_R$ breaks the $SU(2) \otimes U(1)$ gauge symmetry.

1.1.3 The Higgs mechanism

As described in previous section, all possible fermion mass terms are forbidden in the electroweak Lagrangian Eq. 1.31 by the local gauge invariance, and the gauge boson masses can only be generated by an additional scalar field ϕ . This scalar field is an essential part in the structure of GWS theory. It provides a solution to the problem of massless particles, and causes the spontaneous electroweak symmetry breaking. The theory was introduced by Brout, Englert, Higgs and their collaborators in 1964 [12, 13], therefore this mechanism is called as the "Brout-Englert-Higgs" mechanism.

The Higgs field transforms under $SU(2)_L \otimes U(1)_Y$ gauge symmetry as a self-interacting doublet of scalar fields, with the weak hypercharge $Y = 1$. It is parametrized as

$$\Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix}, \quad (1.32)$$

where ϕ^+ and ϕ^0 are two complex scalar fields. The covariant derivative of Φ is

$$D_\mu \Phi = (\partial_\mu + i \frac{1}{2} g W_\mu^a \sigma^a + i \frac{1}{2} g' B_\mu) \Phi, \quad (1.33)$$

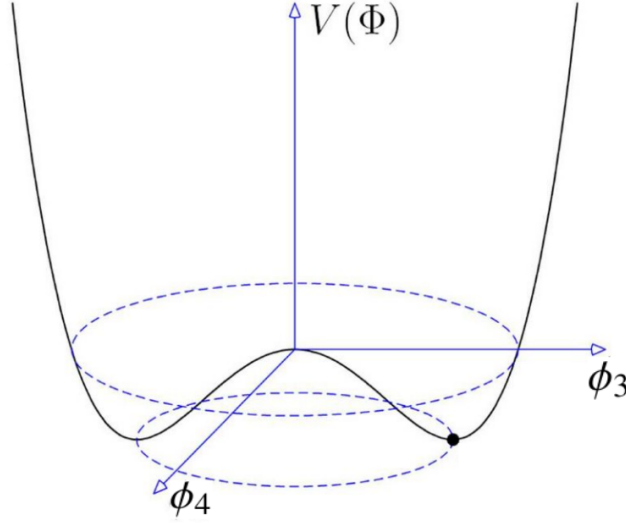


Figure 1.2: The shape of Higgs potential when $\mu^2 < 0$. Spontaneous symmetry breaking occurs when the Higgs field reaches the state of minimum potential.

with the complete form of the Higgs Lagrangian written as:

$$\mathcal{L}_H = (D_\mu \Phi)^\dagger (D^\mu \Phi) - V(\Phi). \quad (1.34)$$

The Higgs field couples to the W_μ^a and B_μ fields through the kinetic term $(D_\mu \Phi)^\dagger (D^\mu \Phi)$, $V(\Phi)$ is potential of the Higgs field. The SM Higgs potential, which has to be invariant under the $SU(2)_L \otimes U(1)_Y$ transformation, is constructed as

$$V(\Phi) = \mu^2 \Phi^\dagger \Phi + \lambda (\Phi^\dagger \Phi)^2, \quad (1.35)$$

where the λ term describes the quartic self-interactions among the scalar fields, μ^2 is a negative constant. Vacuum stability demands λ to be greater than zero. If $\mu^2 > 0$, the state of minimum energy occurs when $\Phi = 0$, and the symmetry is not broken. When $\mu^2 < 0$, the Higgs potential has a "Mexican-hat" shape, and the state of minimum energy is reached when:

$$\Phi^\dagger \Phi = \frac{1}{2}(\phi_1^2 + \phi_2^2 + \phi_3^2 + \phi_4^2) = -\frac{\mu^2}{2\lambda} = \frac{v^2}{2}. \quad (1.36)$$

Where $v = \sqrt{-\mu^2/\lambda}$ is the vacuum expectation value (VEV) of the Higgs field. Spontaneous symmetry breaking occurs when the Higgs field reaches the state of minimum potential. Figure 1.2 shows a schematic drawing of the Higgs potential projected in two dimensions.

With the freedom of $SU(2)$ rotation, the vacuum expectation value of the Higgs field can always be written in the form:

$$\langle \Phi \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}. \quad (1.37)$$

The Higgs field can be written as:

$$\Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H \end{pmatrix}, \quad (1.38)$$

1 The theoretical overview of the Higgs boson

where H is a real scalar field with $\langle H \rangle = 0$, the quanta of this field is the Higgs boson with mass

$$m_H = \sqrt{2\lambda}\nu. \quad (1.39)$$

Then the gauge boson masses arise from $(D_\mu \Phi)^\dagger (D^\mu \Phi)$. The masses can be obtained from the coupling part of the covariant derivative:

$$\Delta \mathcal{L} = \frac{1}{2} \begin{pmatrix} 0 & \nu \end{pmatrix} \left(\frac{1}{2} g W_\mu^a \sigma^a + \frac{1}{2} g' B_\mu \right) \left(\frac{1}{2} g W^{\mu b} \sigma^b + \frac{1}{2} g' B^\mu \right) \begin{pmatrix} 0 \\ \nu \end{pmatrix}, \quad (1.40)$$

evaluate the matrix product explicitly, the masses are found as:

$$m_W = \frac{1}{2} g \nu, \quad m_Z = \frac{1}{2} \nu \sqrt{g^2 + g'^2}, \quad m_A = 0. \quad (1.41)$$

Therefore the spontaneous symmetry breaking caused by the Higgs field generates masses of the W and Z bosons, while the photon remain massless. The fermions acquire mass through renormalizable interaction with the Higgs field. An additional Yukawa Lagrangian is added to describe the interaction between fermions and the Higgs field. It is written as:

$$\mathcal{L}_{\text{Yukawa}} = -y_{ij}^l \bar{L}_{Li} \Phi e_{Rj} - y_{ij}^d \bar{Q}_{Li} \Phi d_{Rj} - y_{ij}^u \bar{Q}_{Li} \tilde{\Phi} u_{Rj} + h.c., \quad (1.42)$$

where y^x are 3×3 Yukawa matrices describing the coupling strength for different generations; $\bar{L}_{Li}$, $\bar{Q}_{Li}$ are the left-handed doublets for three generations of leptons and quarks; e_{Rj} , d_{Rj} , u_{Rj} are the right-handed singlets for three generations of charged leptons, down-type quarks and up-type quarks; i, j refer to the three generations in the quark and lepton sectors; $\tilde{\Phi}$ is defined as $i\sigma_2 \Phi^*$. The mass of fermions can be obtained by inserting Eq. 1.38 to the Lagrangian.

The Lagrangian of the SM is completed at this point, it can be expressed as:

$$\mathcal{L}_{\text{SM}} = \mathcal{L}_{\text{QCD}} + \mathcal{L}_{\text{EW}} + \mathcal{L}_H + \mathcal{L}_{\text{Yukawa}}, \quad (1.43)$$

with each component of the Lagrangian described in this section, expressed as Eq. 1.21, Eq. 1.31, Eq. 1.34 and Eq. 1.42. It is invariant under Lorentz transformation and $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$ transformation.

1.2 Higgs boson phenomenology at the LHC

The SM Higgs boson is the quanta of the Higgs field as described in Section 1.1.3. The SM predicted Higgs boson mass is $m_H = \sqrt{2\lambda}\nu$, where λ is a free parameter not predicted by the SM. The couplings of the Higgs field to other fields also depends on this parameter, so the production rate and decay width of the Higgs boson cannot be determined without knowing the Higgs boson mass. However, once the mass of the Higgs boson is fixed, all the couplings are also fixed. Then the Higgs boson production cross sections and decay width can be computed with the SM. The Higgs boson was discovered in 2012 from several decay channels, the properties of this particle have been measured more and more precisely since the discovery. This section gives an overview of the properties of the SM Higgs boson, the discovery of this particle, and the importance of $VH(b\bar{b})$ channel.

1.2 Higgs boson phenomenology at the LHC

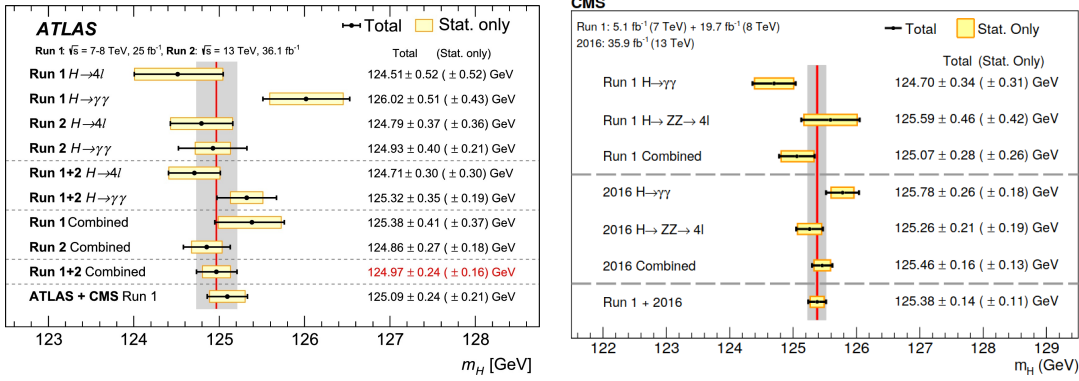


Figure 1.3: Measurements of m_H in $H \rightarrow \gamma\gamma$ and $H \rightarrow ZZ^* \rightarrow 4l$ channels by the ATLAS (left) [31] and the CMS (right) [32] experiment. The measurements are performed with full Run 1 data samples and early Run 2 data samples.

1.2.1 Higgs boson properties

The SM predicted the Higgs boson as a spin-0 and CP-even ($J^P = 0^+$) particle with a free parameter of its mass. One of the first goals of the experiments is to measure the Higgs boson mass. The Higgs boson has two high resolution decay channels: $H \rightarrow \gamma\gamma$ and $H \rightarrow ZZ^* \rightarrow 4l$. The measured mass is obtained from a simultaneous fit to the peak of the invariant mass in these two channels. The measured mass with data collected by the ATLAS detector in Run 1 and early Run 2 is $m_H = 124.97 \pm 0.24$ GeV [31]. The measured mass with data collected by the CMS detector in Run 1 and early Run 2 is $m_H = 125.38 \pm 0.14$ GeV [32]. The measured mass with the combined data samples of the ATLAS and CMS experiments in Run 1 is $m_H = 125.09 \pm 0.24$ GeV [5]. This measurement is used as the most precise result to date. Figure 1.3 is a summary of the Higgs boson mass measurements from several individual and combined analyses by the ATLAS and the CMS experiments.

Measurements of the cross sections in major Higgs decay channels are combined with maximum profile likelihood fit. The signal strength μ , is the parameter of interest in the fitting, representing the measured signal yields in data over the expected signal yields from the SM processes. The significance of the signal is evaluated with a background-only hypothesis test using the profiled likelihood ratio test statistic. More details about the fitting can be found in Section 5.4. The measured signal strength can be used to evaluate the cross section,

$$(\sigma \cdot \mathcal{B}_{H \rightarrow X})_{\text{obs}} = \mu \cdot (\sigma \cdot \mathcal{B}_{H \rightarrow X})_{\text{exp}}, \quad (1.44)$$

where $(\sigma \cdot \mathcal{B}_{H \rightarrow X})_{\text{exp}}$ is the SM predicted cross section of a specific channel. More details about the production cross sections and the decay of the Higgs boson are in the following sections. The measured cross sections in the major Higgs decay channels in Run 1 and Run 2 are in good agreement with the SM predictions.

The spin and parity of the Higgs boson have been studied through bosonic decay channels: $H \rightarrow \gamma\gamma$ and $H \rightarrow ZZ^* \rightarrow 4l$ and $H \rightarrow WW^* \rightarrow l\nu l\nu$. Several alternative hypotheses with $J^P = 0^-, 1^+, 1^-$,

1 The theoretical overview of the Higgs boson

2^+ were studied with the Run 1 data. All the alternative hypotheses are excluded at confidence levels (CL) above 97.8% while the data are compatible with the SM [33].

1.2.2 Higgs boson production

The SM Higgs boson couples to all the quarks, charged leptons and W/Z bosons, which means the Higgs boson can be produced through a variety of production modes. The LHC is a hadron collider that accelerates bunches of protons to very high energy and collide them, so the Higgs boson produced at the LHC mostly comes from hard scatterings between quarks and gluons inside the colliding protons. Since the coupling of the Higgs boson to gluons is null at tree level and the couplings of the Higgs boson to the light quarks (u, d) in protons are very small, the Higgs boson is not produced through direct couplings to gluons or u, d quarks, but through the production of intermediate heavier quarks or vector bosons.

The most abundant Higgs production mode at the LHC is the gluon-gluon fusion (ggF), it is mediated by a top quark loop, the ggF production takes up the majority of the Higgs production modes. The other production modes that have been observed at the LHC are vector boson fusion (VBF), associated production with a W or Z boson (VH), associated production with a $t\bar{t}$ pair ($t\bar{t}H$). The VBF production comes from the coupling of two vector bosons and a Higgs boson, where the vector bosons are emitted by the two colliding quarks. A key feature of this mode is that it has two forward jets with a large angle between each other. The VH production is mediated by a vector boson W or Z , then the vector boson radiates a Higgs boson. The $t\bar{t}H$ production produces the Higgs boson in the fusion of a $t\bar{t}$ quark pair. The Feynman diagrams of these main production modes are shown in Figure 0.1.

The SM predicted cross sections of the Higgs boson production depends on m_H and the colliding energy of the quarks and gluons within protons. The quarks and gluons carry a fraction of the proton energy. The probability distributions of the fraction, called as the parton distribution functions (PDF), are determined from global fits to several observables measured in experiments [34]. In general, the production cross section increase as center-of-mass energy increase, decrease as Higgs boson mass increases. The relations of Higgs production cross section and m_H when $\sqrt{s} = 13$ TeV is shown in Figure 1.4. Figure 1.5 shows the relations of Higgs production cross section and $\sqrt{s}$ when $m_H = 125$ GeV. The SM predicted cross sections have been computed with a center-of-mass energy $\sqrt{s} = 13$ TeV and $m_H = 125.09$ GeV, the results are presented in Table 1.1.

1.2.3 Higgs boson decay

The SM Higgs boson has a very short lifetime so it will decay to final states of fermions or vector bosons. The major decay channels of the Higgs boson are $H \rightarrow b\bar{b}$, $H \rightarrow WW^*$, $H \rightarrow ZZ^*$, $H \rightarrow \gamma\gamma$ and several other decay modes. The total decay width of the Higgs boson is the sum of the partial width of all possible decay channels. The branching ratio (BR) of a channel is defined as the ratio of the partial decay width and the total decay width,

$$\mathcal{B}_{H \rightarrow X} = \frac{\Gamma_{H \rightarrow X}}{\Gamma_H^{\text{tot.}}} \quad (1.45)$$

1.2 Higgs boson phenomenology at the LHC

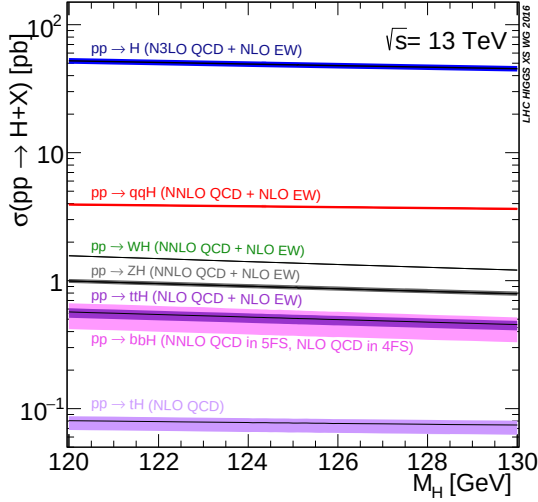


Figure 1.4: SM predicted Higgs boson production cross sections of the main production modes as a function of Higgs boson mass m_H when center-of-mass energy $\sqrt{s} = 13$ TeV [35].

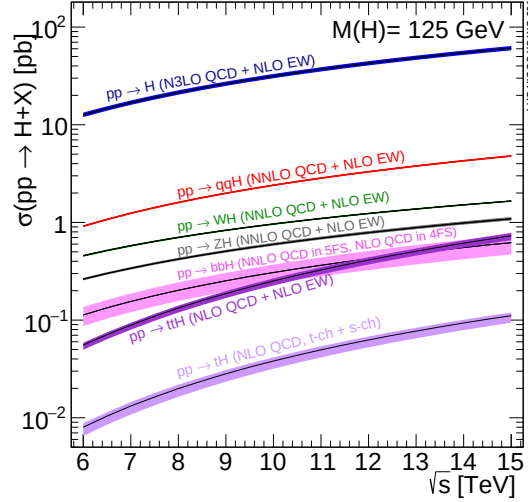


Figure 1.5: SM predicted Higgs boson production cross sections of the main production modes as a function of center-of-mass energy $\sqrt{s}$ when Higgs boson mass $m_H = 125$ GeV [35].

Production mode	σ [pb]	percentage
ggF ($pp \rightarrow H$)	48.61	88%
VBF ($pp \rightarrow qqH$)	3.766	6.8%
WH ($pp \rightarrow WH$)	1.358	2.4%
ZH ($pp \rightarrow ZH$)	0.880	1.6%
ttH ($pp \rightarrow t\bar{t}H$)	0.507	0.9%

Table 1.1: SM predicted Higgs boson production cross sections when center-of-mass energy $\sqrt{s} = 13$ TeV, Higgs boson mass $m_H = 125$ GeV for the main production modes [36]. The values are computed at next-to-leading order (NLO) precision level for VBF, VH, $t\bar{t}H$ and next-to-next-leading order (NNLO) precision level for ggF.

The SM predicted branching ratio depends on the Higgs boson mass, Figure 1.6 shows the branching ratio of several major Higgs decay channels as a function of Higgs boson mass. Table 1.2 gives a summary of the branching ratio for the major Higgs decay channels computed with $m_H = 125.09$ GeV.

The decays to massive weak vector bosons W , Z are suppressed as twice of the masses of these particles exceed the Higgs boson mass, and one of the vector boson has to be off-shell. The dominant decay channel is the decay to $b\bar{b}$ quark pairs, with its branching ratio at 58.2%. However, this channel was not observed in Run 1 because $H \rightarrow b\bar{b}$ through ggF production has an all-hadronic final state, which makes it very difficult to be observed in the overwhelming backgrounds from QCD processes at hadron colliders. VH production is the best option for the observation of $b\bar{b}$ decay channel at the LHC. This thesis focus on the analysis of $VH(H \rightarrow b\bar{b})$ channel in Run 2.

1 The theoretical overview of the Higgs boson

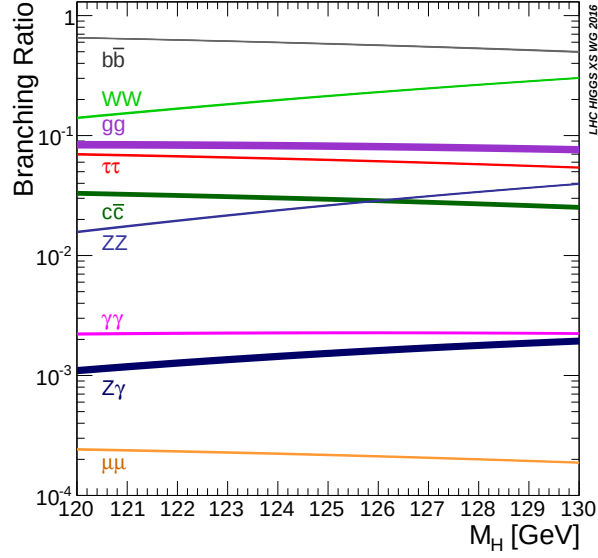


Figure 1.6: BR of the major Higgs decay channels as a function of m_H [35].

Decay channel	BR
$b\bar{b}$	58.24%
WW^*	21.37%
gg	8.19%
$\tau^+\tau^-$	6.27%
$c\bar{c}$	2.89%
ZZ^*	2.62%
$\gamma\gamma$	0.23%
$Z\gamma$	0.15%
$\mu^+\mu^-$	0.02%

Table 1.2: Branching ratios of the major Higgs decay channels computed with $m_H = 125.09$ GeV [36]. The complete NLO electroweak corrections have been implemented.

1.2 Higgs boson phenomenology at the LHC

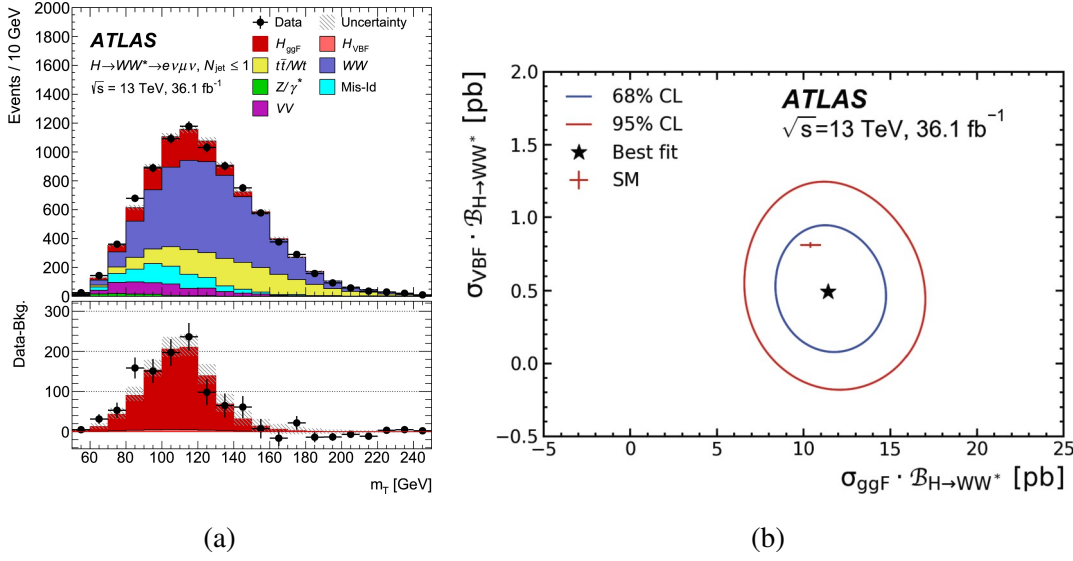


Figure 1.7: Results of $H \rightarrow WW^* \rightarrow e\nu\mu\nu$ analysis with data collected in early Run 2 by the ATLAS detector. (a) Post-fit transverse mass distribution. (b) 68% and 95% confidence level two-dimensional likelihood contours of $\sigma_{ggF} \cdot \mathcal{B}_{H \rightarrow WW^*}$ vs. $\sigma_{VBF} \cdot \mathcal{B}_{H \rightarrow WW^*}$, compared to the SM prediction shown by the red marker [39].

The decay channel with the second largest branching ratio is $H \rightarrow WW^*$, where one of the W bosons is off-shell due to the kinematic threshold. This channel was observed in Run 1 through leptonic decay of W bosons, where the W boson decays to a charged lepton and a neutrino, with the final states of $e\mu$, ee or $\mu\mu$ plus neutrinos. The most challenging part of $H \rightarrow WW^*$ channel is the neutrinos in the final state that cannot be detected. Thus the decay products cannot be completely reconstructed, which makes it impossible to reconstruct the invariant mass of the final state. The HWW analysis used a "transverse mass" m_T calculated from transverse momentum and energy as the final discriminant. The $H \rightarrow WW^*$ signal was observed in data collected during 2011 and 2012 in Run 1, corresponding to integrated luminosities of 25 fb^{-1} at $\sqrt{s} = 7 \text{ TeV}$ and 8 TeV . The signal was observed by the ATLAS experiment with a significance of 5.8 standard deviations [37]. Similar signal was observed by the CMS experiment with a significance of 4.3 standard deviations [38]. The cross sections have been measured more precisely with data collected in Run 2. Figure 1.7 shows the results of the published ATLAS HWW analysis with data collected during 2015 and 2016 in early Run 2, corresponding to integrated luminosities of 36.1 fb^{-1} at $\sqrt{s} = 13 \text{ TeV}$. The fiducial inclusive and differential cross sections have been measured with full Run 2 data in the CMS HWW analysis. Figure 1.8 shows the differential cross sections in bins of the Higgs boson transverse momentum p_T^H and jet multiplicity N_{jet} . All the measurements agree with the SM predictions.

The other two decay channels with relatively large branching ratios are $H \rightarrow gg$ and $H \rightarrow \tau\tau$. The decay to gluon pairs also has an all-hadronic final state, it cannot be distinguished from the QCD backgrounds at the LHC, so this decay mode has not been studied. For the case of $H \rightarrow \tau\tau$, the τ -lepton can decay hadronically or leptonically. The final state of the leptonic decay contains two neutrinos and an electron or a muon, where the neutrinos cannot be detected. Evidence for $H \rightarrow \tau\tau$ was found in data collected in Run 1, then this channel was observed in Run 2 by the ATLAS and the CMS experiments [41, 42].

1 The theoretical overview of the Higgs boson

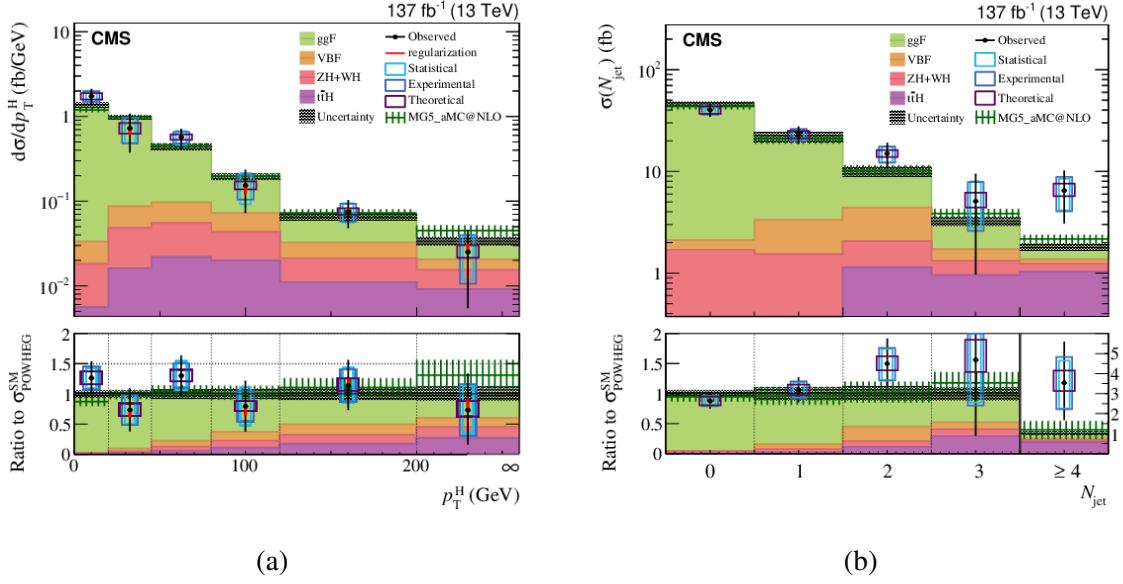


Figure 1.8: Observed and expected fiducial cross sections of $H \rightarrow WW^*$ analysis with full Run 2 data by the CMS detector, presented in bins of p_T^H (left) and N_{jet} (right) [40].

The two decay channels with smaller branching ratio but clean final states are $H \rightarrow \gamma\gamma$ and $H \rightarrow ZZ^*$. The Higgs boson has no direct coupling to photons, so the decay to photons is a loop-induced decay mainly mediated by top quark or W boson. The main backgrounds in $H \rightarrow \gamma\gamma$ channel are non-resonant background from continuum $\gamma\gamma$ production and $V\gamma\gamma$ production. The Z boson in $H \rightarrow ZZ^*$ channel can decay hadronically or leptonically, the leptonic decays of Z bosons, $ZZ \rightarrow 4l$, where $l = e$ or μ , provide a very clean signature in the final state. The main background in this channel is the continuum $(Z^*/\gamma^*)(Z^*/\gamma^*)$ production. These decay channels were observed in Run 1 and were used to perform measurements of Higgs boson properties due to the high resolution. The $H \rightarrow ZZ^* \rightarrow 4l$ signal was observed with a significance of 8.1 standard deviations [43] and the $H \rightarrow \gamma\gamma$ signal was observed with a significance of 5.2 standard deviations [44] by the ATLAS experiment. Similar results were confirmed by the CMS experiment. $H \rightarrow ZZ^*$ was observed via the inclusive $H \rightarrow 4l$ cross section measurement [45]. $H \rightarrow \gamma\gamma$ was observed with a significance of 5.7 standard deviations [46]. Properties of the Higgs boson have been measured more precisely with data collected in Run 2. The invariant mass distributions of the ATLAS $H\gamma\gamma$ and HZZ analyses with Run 2 data are shown in Figure 1.9. Similar studies have been done by the CMS experiment, Figure 1.10 shows the invariant mass distributions from the CMS $H\gamma\gamma$ and HZZ analyses with full Run 2 data. The measured cross sections are all in agreement with the SM predictions.

Another decay channel $H \rightarrow c\bar{c}$ has similar branching ratio compared to $H \rightarrow ZZ^*$, but this channel has an all-hadronic final state, the low branching ratio and the difficulty of tagging jets originated from charm quarks make it very challenging to study this channel. Searches for Higgs boson decaying to a pair of charm quarks have been performed with full Run 2 data collected by the ATLAS detector [51] and the CMS detector [52]. This decay mode has not been observed yet.

The two channels with very low branching ratios are $H \rightarrow Z\gamma$ and $H \rightarrow \mu\mu$. It is very challenging to

1.2 Higgs boson phenomenology at the LHC

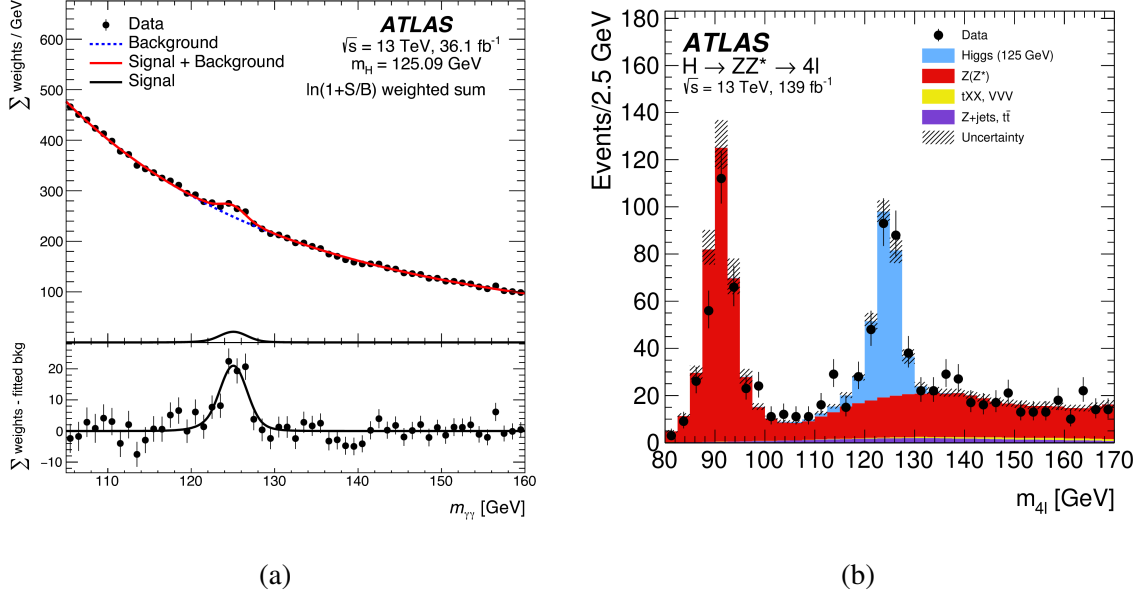


Figure 1.9: (a) Diphoton invariant mass $m_{\gamma\gamma}$ spectrum observed in data collected in 2015 and 2016 by the ATLAS detector with an integrated luminosity of 36.1 fb^{-1} at $\sqrt{s} = 13 \text{ TeV}$ [47]. (b) The observed and expected four-lepton invariant mass distribution for the selected Higgs boson candidates with a constrained Z boson mass, obtained from full Run 2 data collected by the ATLAS detector with an integrated luminosity of 139 fb^{-1} at $\sqrt{s} = 13 \text{ TeV}$ [48].

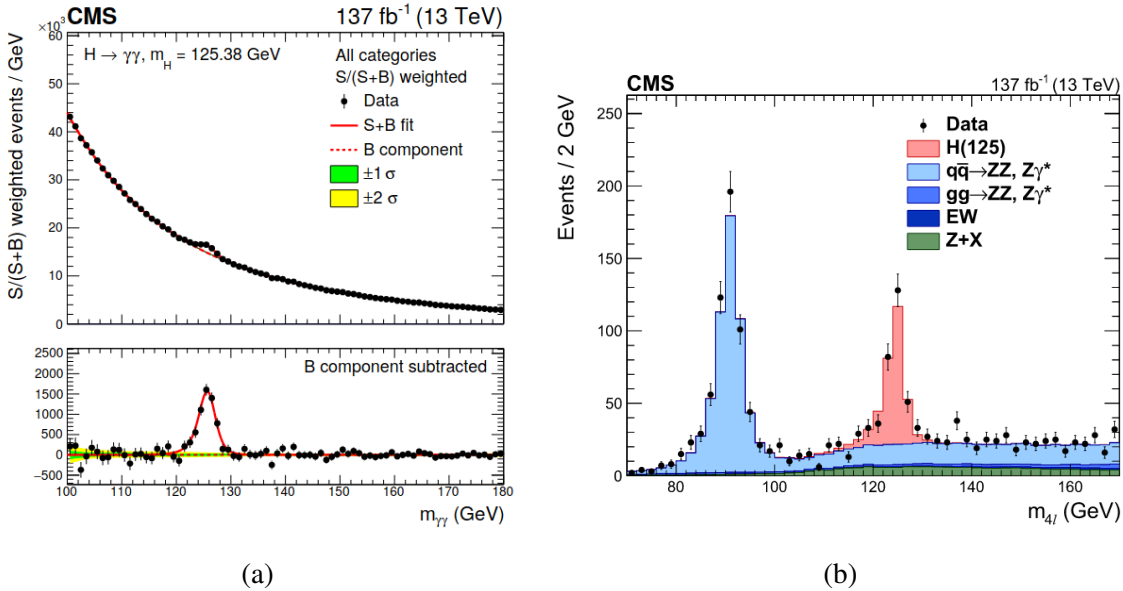


Figure 1.10: (a) Diphoton invariant mass distribution extracted from full Run 2 data collected by the CMS detector [49]. (b) Four-lepton invariant mass distribution of selected Higgs boson candidates with a constrained Z boson mass, obtained from full Run 2 data collected by the CMS detector [50].

1 The theoretical overview of the Higgs boson

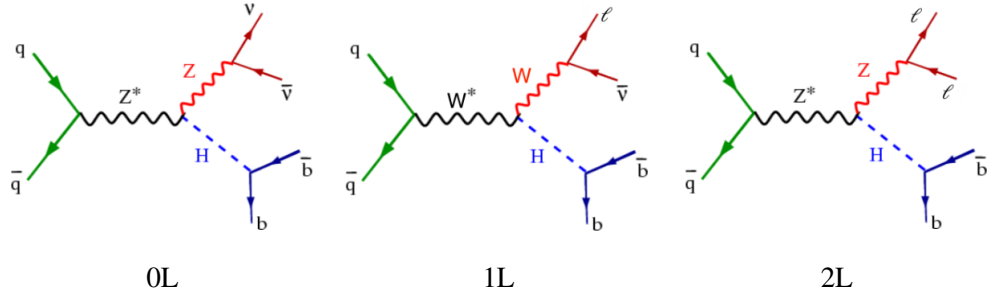


Figure 1.11: Feynman diagrams of $VH(H \rightarrow b\bar{b})$ processes with zero, one or two charged leptons in the final state.

study these two channels due to the low branching ratio. Evidences for these two channels are found in full Run 2 data collected by the ATLAS detector and the CMS detector [53–56].

1.2.4 $VH(H \rightarrow b\bar{b})$ channel

The $H \rightarrow b\bar{b}$ decay channel is very important because the couplings between the Higgs boson and fermions have to be studied in such channels, and the largest branching ratio of the $b\bar{b}$ decay channel makes it the best option. The $Hb\bar{b}$ analysis in Run 1 was not able to confirm the existence of this decay mode as described in Section 1.2.3. $H \rightarrow b\bar{b}$ decay through ggF production has all-hadronic final states. The preferred production mode to study this decay channel is the VH production, where the vector boson decays leptonically. Despite the cross section of VH production one order of magnitude lower than the ggF production cross section, the leptons in the final states enable efficient rejections of the multi-jet backgrounds. The first evidence for the $H \rightarrow b\bar{b}$ decay was found in data collected during 2015 and 2016 corresponding to an integrated luminosity of 36.1 fb^{-1} . The signal was found with an observed significance of 3.5 standard deviations by the ATLAS collaboration [57], and similar result was confirmed by the CMS collaboration with an observed significance of 3.8 standard deviations [58]. The targeted final states contain zero, one or two charged leptons, and two b -tagged small- R jets. Figure 1.11 shows the Feynman diagrams of the three sub-channels in $VH(b\bar{b})$ analysis. This analysis strategy is referred to as the "resolved". It mainly targets low to medium Higgs boson transverse momentum (p_T) phase space, where the two b -jets from the decay of the Higgs boson can be separated clearly. With the high quality data collected in Run 2 from 2015 to 2018 corresponding to an integrated luminosity of 139 fb^{-1} , the $VH(b\bar{b})$ resolved analysis is updated. The production of $VH(b\bar{b})$ signal is observed with a significance of over 5 standard deviations [59].

The other analysis strategy that mainly targets medium to high Higgs p_T phase space is developed in parallel of the $VH(b\bar{b})$ resolved analysis. The angle between the two b -jets from the decay of the Higgs boson can be approximate as:

$$\Delta R \approx \frac{2m_H}{p_T}, \quad (1.46)$$

where m_H is the Higgs boson mass, p_T is the Higgs boson transverse momentum. For $m_H = 125 \text{ GeV}$, the angle between the two b -jets decreases as the p_T of the Higgs boson increases. Eventually the two b -jets are reconstructed as a large- R jet. This analysis strategy is referred to as the "boosted". The

1.3 Beyond the Standard Model

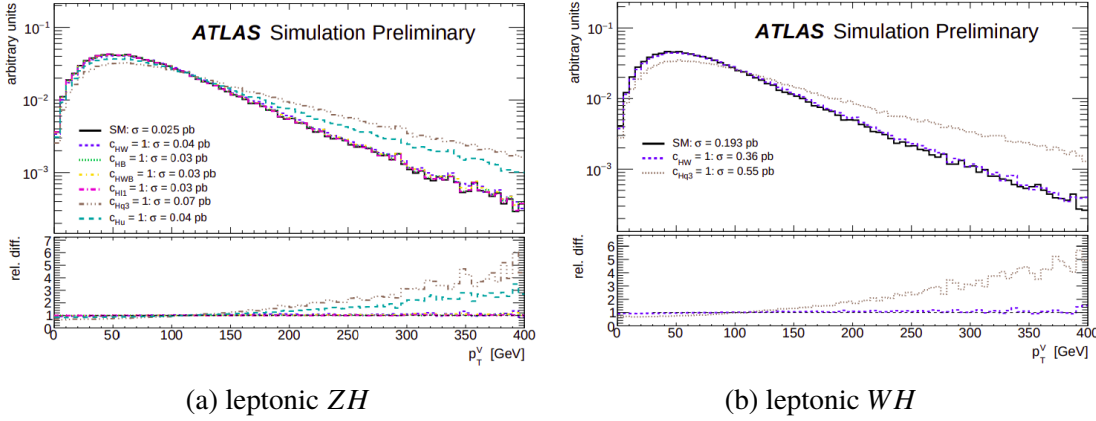


Figure 1.12: Normalized distributions of p_T^V for ZH and WH production with the leptonic decay of the vector boson [60], impacted by the most relevant EFT operators described in Section 1.3.2. High p_T regions are more sensitive to potential anomalous couplings.

$VH(b\bar{b})$ boosted analysis is able to explore very high Higgs boson p_T phase space that has not been studied in the resolved analysis. The high p_T regions are more sensitive to anomalous couplings as shown in Figure 1.12, which means it has greater potential to reveal new physics.

The $VH(b\bar{b})$ analysis is categorized based on the vector boson transverse momentum p_T^V , this variable should be roughly the same in value as the Higgs boson p_T because the vector sum of transverse momentum of all the decay products should be very small. The resolved strategy is applicable in regions below 400 GeV, and the boosted strategy is applicable in regions above 250 GeV. The medium p_T^V region (250-400 GeV) can be studied with both the resolved and the boosted approaches. Combinations of the two analyses have been studied to design one coherent, harmonized approach to extract $VH(b\bar{b})$ results, referred to as the legacy analysis.

1.3 Beyond the Standard Model

The SM has been very successfully tested in experiments. The results from LHC experiments agree very well with the SM predictions. However, the SM still has limitations. Several phenomena observed in other experiments cannot be explained by the SM. First, the SM only includes three of the four fundamental interactions. The gravity cannot be explained by the SM. Second, many lines of evidence from astronomy suggest the possible existence of dark matter and dark energy, which makes up 95% of the universe. The dark matter and dark energy are not included in the SM. Third, several neutrino oscillation experiments have shown three generations of neutrinos do have masses, which is not compatible with the SM. Several other phenomena also indicate the SM is not sufficient. The efforts to study these beyond the Standard Model (BSM) phenomena by extending the SM Lagrangian come in two different ways. One is to construct a Lagrangian based on a specific model. Many models (like SUSY) constructed through extending the SM with new particles and symmetries are introduced to explain some of the BSM phenomena. The other is an effective field theory (EFT) [61]. It constructs a Lagrangian by adding higher order terms to the SM Lagrangian. This approach is model

1 The theoretical overview of the Higgs boson

independent. The measured cross sections can be presented with a common framework adopted by the LHC experiments, the Simplified Template Cross Section (STXS) framework, to test the EFT.

1.3.1 Simplified Template Cross Sections

Many measurements of the Higgs boson in Run 1 resulted in the measured signal strength, which is the parameter of interest (POI) of the maximum profile likelihood fit. Given the increased statistics of data collected in Run 2, the signal strength can be measured more precisely, to probe the production and decay rate of the Higgs boson in more details. The STXS framework [62] was developed for such measurements. In this framework, the cross sections of given channels are measured in various kinematic regions, called as the STXS bins. The STXS framework allows simple combination between measurements from different channels, and also reduces the theoretical uncertainties as the uncertainties are directly folded into the measurements. Thus, this framework has become the common framework adopted by LHC experiments for Higgs measurements, available in all major Higgs decay channels. The STXS measurements can then be used as inputs to subsequent interpretations in and beyond the Standard Model, including

- the determination of overall signal strengths or coupling scale factors,
- fitting to a set of EFT coefficients,
- tests of a specific BSM model.

The STXS bins are defined in mutually exclusive phase space. The kinematic cuts used to define the bins are abstracted and simplified compared to the actual cuts used in the individual analysis of a Higgs decay channel. The definitions of STXS bins are common for all the Higgs decay channels with a specific production mode, as the SM Higgs production process serves as the kinematic templates. The simplified kinematic cuts allow the combination of measurements from different decay channels, which is one of the key features of the STXS measurements. The definitions of STXS bins are set to be close to typical event selections in the analyses, or more generally, the kinematic regions that dominate the sensitivity. The goals of the definition of a set of STXS bins for a specific production mode are to

- minimize the dependence on theoretical uncertainties that are folded into the measurements,
- maximize the experimental sensitivity by making the definitions as close as possible to the event selection in the specific analysis,
- isolate possible BSM effects to be interpreted using EFT model or a specific BSM model,
- limit the number of bins at a manageable level for matching with the experimental sensitivity and practical implementation.

The measurements of cross sections are unfolded into the STXS bins to be compared with the theoretical predictions. The theoretical predictions in the STXS bins are from either calculations or MC simulations to predict the number of entries (cross sections) in every STXS bin at truth level, which means using event information at generator level to decide the STXS bin assigned to this event. Several kinematic cuts common to all the bins are applied. For the Higgs boson, a cut on rapidity is

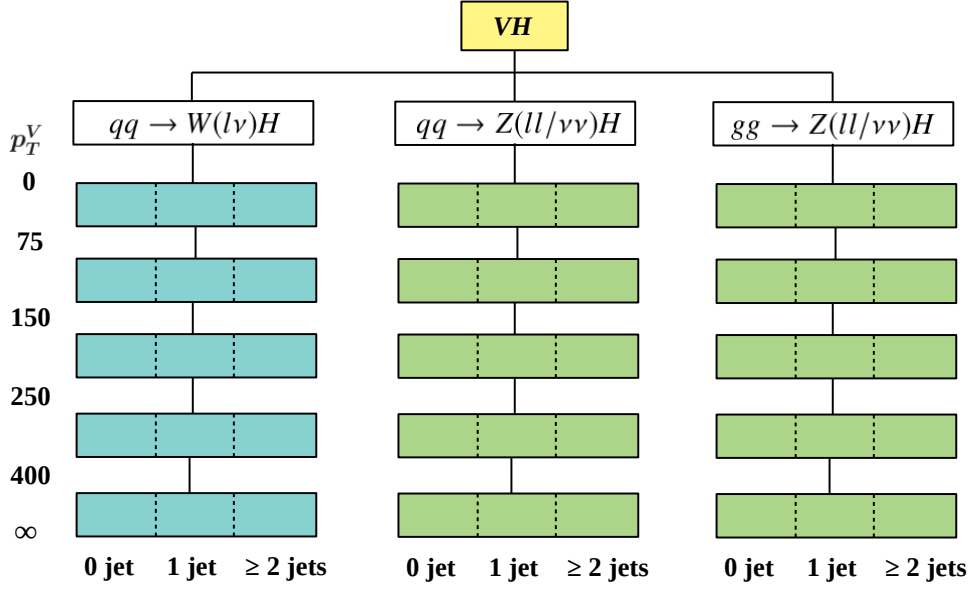


Figure 1.13: Definition of the STXS bins with the maximum splitting, cuts on truth p_T^V are in GeV. The main splitting merged all the ZH categories and no split on the number of additional small- R jets is applied.

applied, $|Y_H| < 2.5$. The vector bosons are defined from the sum of all their decay products including neutrinos. The truth jets are small- R jets reconstructed from truth particles. A p_T threshold requiring $p_T > 30$ GeV is applied to all the truth small- R jets. For VH production mode, it is first separated to three categories:

$$qq \rightarrow W(l\nu)H, \quad qq \rightarrow Z(ll/\nu\nu)H, \quad gg \rightarrow Z(ll/\nu\nu)H. \quad (1.47)$$

Each category is then split into five bins based on transverse momentum of the vector boson (p_T^V), with the boundaries at 75, 150, 250 and 400 GeV. Each p_T^V bin can then be divided based on number of additional jets. "Additional" means small- R jets with $p_T > 30$ GeV that are not from the decay of the Higgs boson. Three bins with zero additional small- R jet, one additional small- R jet and at least two additional small- R jets are defined. The splitting of the STXS bins for VH production is shown in Figure 1.13. This way of splitting is to provide a high flexibility in the choice of the bins to measure. The splitting actually used in STXS measurement for boosted $VH(b\bar{b})$ analysis merged several categories for practical implementation due to the limited statistics. This set of splitting is called as the main splitting, with only WH and ZH categories, and no split on the number of additional small- R jets.

The expected signal events in each STXS bins are estimated using simulated signal samples with truth level information, the estimation is performed in every reconstructed event category for $VH(b\bar{b})$ analysis. The measured cross sections are extracted from multi-POI fits to data, where the cross section in every STXS bin is a STXS-POI.

1 The theoretical overview of the Higgs boson

1.3.2 Effective Field theories

The SM Lagrangian is constructed from dimension-4 operators. From the EFT point of view, the SM is a low energy approximation of a more generic model. The SM itself is an effective theory which is applicable up to a certain energy scale Λ . The field theory above this energy scale should preserve the gauge symmetry of the SM. All the SM elementary particles and interactions should be included, and the theory should reduce to the SM at low energies. Based on principle of least action Eq. 1.1, any Lagrangian have to be dimension-4 because the time-space integral of the Lagrangian is dimensionless. The higher dimension- d operators added to the Lagrangian have to be suppressed by $d - 4$ powers of the scale Λ , to keep the whole term still being dimension-4. The SM EFT Lagrangian can be written as:

$$\mathcal{L}_{\text{EFT}} = \mathcal{L}_{\text{SM}} + \sum_{d \geq 5} \sum_i \frac{c_i^{(d)}}{\Lambda^{d-4}} \mathcal{O}_i^{(d)}, \quad (1.48)$$

where $\mathcal{O}_i^{(d)}$ are higher dimensional operators which are invariant under $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$ local gauge symmetry, and $c_i^{(d)}$ are free parameters called Wilson coefficients that are constrained by experiments. Dimension-5 operators contain operators that violate lepton number and give rise to mass terms for the left-handed neutrinos, so the scale Λ that suppresses the dimension-5 operator would be much larger than that suppresses the dimension-6. The same happens with dimension-7, all the dimension-7 operators violate lepton number, some also violate baryon number. The leading order higher dimension operators are dimension-6. Dimension-8 operators could have comparable impacts through the interference between the SM. Operators with higher dimension ($d \geq 9$) are suppressed with a larger power of Λ , so their contribution is negligible. The EFT Lagrangian most likely to have observable effects then becomes:

$$\mathcal{L}_{\text{EFT}} = \mathcal{L}_{\text{SM}} + \sum_i \frac{c_i^{(6)}}{\Lambda^2} \mathcal{O}_i^{(6)} + \sum_i \frac{c_i^{(8)}}{\Lambda^4} \mathcal{O}_i^{(8)}. \quad (1.49)$$

Only dimension-6 operators are considered for simplicity. The operator $\mathcal{O}^{(6)}$ can be constructed using a complete, non-redundant set, the so-called basis. Several methods of construction have been developed. The choice of basis does not need to be unique. Any complete basis will lead to the same physics results. The basis used to present the results in the $VH(b\bar{b})$ analysis in this thesis is the Warsaw basis [61].

The measurement of cross sections in the STXS framework can be related to the Wilson coefficients. The cross section of EFT model can be expressed as the following form for convenience,

$$\sigma_{\text{EFT}} = \sigma_{\text{SM}} + \sigma_{\text{int}} + \sigma_{\text{BSM}}, \quad (1.50)$$

where σ_{SM} is the cross section computed with just the SM, σ_{int} arises from the interference between the SM and BSM processes, σ_{BSM} comes from just BSM processes. The ratio of the cross section can be expressed with the Wilson coefficients:

$$\frac{\sigma_{\text{EFT}}}{\sigma_{\text{SM}}} = 1 + \sum_i A_i c_i + \sum_{ij} B_{ij} c_i c_j, \quad (1.51)$$

1.3 Beyond the Standard Model

where A_i and B_{ij} are coefficients derived from the following definitions,

$$\frac{\sigma_{\text{int}}}{\sigma_{\text{SM}}} = \sum_i A_i c_i, \quad \frac{\sigma_{\text{BSM}}}{\sigma_{\text{SM}}} = \sum_{ij} B_{ij} c_i c_j, \quad (1.52)$$

they can be calculated for each region in the STXS measurement using Monte Carlo (MC) simulations [63].

The partial decay width of the Higgs boson to a given final state can be parametrized in the same way,

$$\frac{\Gamma_{\text{EFT}}}{\Gamma_{\text{SM}}} = 1 + \sum_i \alpha_i c_i + \sum_{ij} \beta_{ij} c_i c_j, \quad (1.53)$$

where α_i and β_{ij} are coefficients derived from:

$$\frac{\Gamma_{\text{int}}}{\Gamma_{\text{SM}}} = \sum_i \alpha_i c_i, \quad \frac{\Gamma_{\text{BSM}}}{\Gamma_{\text{SM}}} = \sum_{ij} \beta_{ij} c_i c_j, \quad (1.54)$$

these coefficients are also calculated from MC simulations [63].

The cross section of $VH(b\bar{b})$ process is:

$$\sigma_{VH(H \rightarrow b\bar{b})} = \sigma_{VH} \cdot \mathcal{B}_{H \rightarrow b\bar{b}}, \quad (1.55)$$

with the branching ratio of $H \rightarrow b\bar{b}$ derived from:

$$\mathcal{B}_{H \rightarrow b\bar{b}} = \frac{\Gamma_H^{b\bar{b}}}{\Gamma_H^{\text{tot.}}}. \quad (1.56)$$

Therefore, given the production cross section and the decay width of the Higgs boson parametrized with the Wilson coefficients, the cross section in each region in the STXS measurement can also be parameterized with the Wilson coefficients.

$$(\sigma_{VH} \cdot \mathcal{B}_{H \rightarrow b\bar{b}})_{\text{EFT}} = (\sigma_{VH} \cdot \mathcal{B}_{H \rightarrow b\bar{b}})_{\text{SM}} \cdot \frac{\sigma_{\text{EFT}}}{\sigma_{\text{SM}}} \cdot \frac{\Gamma_{\text{EFT}}^{b\bar{b}}/\Gamma_{\text{SM}}^{b\bar{b}}}{\Gamma_{\text{EFT}}^{\text{tot.}}/\Gamma_{\text{SM}}^{\text{tot.}}}. \quad (1.57)$$

Among the operators in the full Warsaw basis, only 17 operators have effects on $VH(b\bar{b})$ process at leading-order (LO), four of which only affect $H \rightarrow b\bar{b}$ decay. The operators and their corresponding Wilson coefficients are listed in Table 1.3.

1 The theoretical overview of the Higgs boson

Wilson coefficient	Operator	Impact
c_{Hbox}	$(H^\dagger H)\square(H^\dagger H)$	ZH, WH
c_{HDD}	$(H^\dagger D^\mu H)^*(H^\dagger D_\mu H)$	ZH, WH
c_{HW}	$H^\dagger HW_{\mu\nu}^I W^{I\mu\nu}$	ZH, WH
c_{HB}	$H^\dagger HB_{\mu\nu} B^{\mu\nu}$	ZH
c_{HWB}	$H^\dagger \tau^I HW_{\mu\nu}^I B^{\mu\nu}$	ZH
c_{HG}	$H^\dagger HG_{\mu\nu}^A G^{A\mu\nu}$	$\Gamma_H^{\text{tot.}}$
$ c_{eH} $	$(H^\dagger H)(\bar{l}_p e_r H)$	$\Gamma_H^{\text{tot.}}$
$ c_{uH} $	$(H^\dagger H)(\bar{q}_p u_r \tilde{H})$	$\Gamma_H^{\text{tot.}}$
$ c_{dH} $	$(H^\dagger H)(\bar{q}_p d_r \tilde{H})$	$\Gamma_H^{b\bar{b}}$
$c_{Hl}^{(1)}$	$H^\dagger i \overleftrightarrow{D}_\mu H (\bar{l}_p \gamma^\mu l_r)$	ZH
$c_{Hl}^{(3)}$	$H^\dagger i \overleftrightarrow{D}_\mu^I H (\bar{l}_p \tau^I \gamma^\mu l_r)$	ZH, WH
$c_{He}^{(1)}$	$H^\dagger i \overleftrightarrow{D}_\mu H (\bar{e}_p \gamma^\mu e_r)$	ZH
$c_{Hq}^{(1)}$	$H^\dagger i \overleftrightarrow{D}_\mu H (\bar{q}_p \gamma^\mu q_r)$	ZH
$c_{Hq}^{(3)}$	$H^\dagger i \overleftrightarrow{D}_\mu^I H (\bar{q}_p \tau^I \gamma^\mu q_r)$	ZH, WH
$c_{Hu}^{(1)}$	$H^\dagger i \overleftrightarrow{D}_\mu H (\bar{u}_p \gamma^\mu u_r)$	ZH
$c_{Hd}^{(1)}$	$H^\dagger i \overleftrightarrow{D}_\mu H (\bar{d}_p \gamma^\mu d_r)$	ZH
$c_{ll}^{(1)}$	$(\bar{l}_p \gamma_\mu l_r)(\bar{l}_s \gamma^\mu l_t)$	ZH, WH

Table 1.3: Operators that have effects on $qq \rightarrow V(\text{leptons})H(\rightarrow b\bar{b})$ process at LO in the Warsaw basis [60]. The "box" operator is the Laplacian of D_μ . The operators l_p, l_r and q_p, q_r are three generations of left-handed leptons and quarks, with $p, r = 1, 2, 3$ to denote the generations. The operators $e_{p(r)}, u_{p(r)}, d_{p(r)}$ are three generations of right-handed leptons, up-type quarks and down-type quarks, respectively. The derivative $\overleftrightarrow{D}_\mu$ and $\overleftrightarrow{D}_\mu^I$ are defined as: $H^\dagger i \overleftrightarrow{D}_\mu H = iH^\dagger (D_\mu - \overleftarrow{D}_\mu)H$, $H^\dagger i \overleftrightarrow{D}_\mu^I H = iH^\dagger (\tau^I D_\mu - \overleftarrow{D}_\mu \tau^I)H$, where $H^\dagger \overleftarrow{D}_\mu H = (D_\mu H)^\dagger H$, H is the Higgs field, τ^I are the Pauli matrices, and D_μ is the covariant derivative of the $SU(2) \otimes U(1)$ transformation.

2 The Large Hadron Collider and the ATLAS experiment

2.1 The Large Hadron Collider

The Large Hadron Collider (LHC) [4] is a proton and heavy ion accelerator located at CERN. It is built in a tunnel approximately 100 m underground, with a circumference of 27 km. The LHC is the last ring in a complex chain of particle accelerators. The structure of the accelerator complex is shown in Figure 2.1. The process of acceleration starts from a linear accelerator called LINAC 2. The protons are produced from hydrogen gas. An electric field is added to strip the electron from the hydrogen atoms to produce protons. The protons are then injected into LINAC 2, and accelerated to 50 MeV. The beam of proton is then injected into a circular accelerator called Booster, which accelerates the protons to 1.4 GeV. After that, the beam is injected into the Proton Synchrotron (PS), which accelerates the protons to 25 GeV. Then the beam is injected into the Super Proton Synchrotron (SPS), which accelerates the protons to 450 GeV. Finally, the protons are injected into the LHC, and are accelerated to the final energy.

The LHC is operated in an ultra high vacuum environment. It uses 1232 superconducting dipole magnets to keep the beams in their circular paths, the magnets provide fields up to 8.3 T. Additionally, 392 quadrupole magnets are installed to focus the beams. The LHC consists of two separate beam

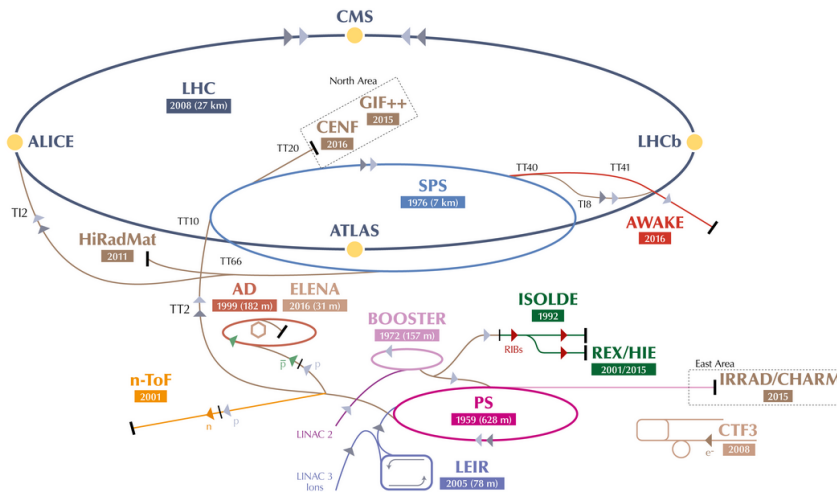


Figure 2.1: Structure of the complex chain of particle accelerators at CERN [64].

2 The Large Hadron Collider and the ATLAS experiment

pipes, each contains a beam of protons. The beams travel in opposite directions and collide at the four interaction points (IP). The detectors are built around these IPs. The major experiments at each IP of the LHC are ATLAS, CMS, ALICE and LHCb. The ATLAS detector and the CMS detector are two general purpose detectors designed with similar scientific goals. They can be used to study a wide range of physics including precision tests of the SM and search for BSM phenomena. The ALICE detector is designed to study heavy ion physics. The LHCb detector is a specialized detector designed to study flavor physics.

The LHC is designed with a center-of-mass collision energy $\sqrt{s}$ up to 14 TeV, corresponding to a magnetic field of 8.33 T. The protons are accelerated to half of the collision energy in bunches. Each bunch contains up to 10^{11} protons. The two beams are separated in different beam pipes. The only common sections are around the interaction points, where the beams share an approximately 38 m long beam pipe in the ATLAS experimental area. The space between bunches are kept at 7.5 m. For protons travel at close to the speed of light, the collisions take place every 25 ns. The separation on time between bunches is called as the bunch spacing. The LHC ring can accommodate 2808 bunches for each proton beam. The instantaneous luminosity of a particle physics experiment is defined as:

$$\Delta N_{\text{event}} [\text{s}^{-1}] = \sigma_{\text{process}} [\text{cm}^2] \cdot L [\text{s}^{-1} \text{cm}^{-2}], \quad (2.1)$$

where ΔN_{event} is the number of events generated in the experiment per second for a process, and σ_{process} is the cross section of this process. The luminosity of the LHC only depends on beam parameters. It can be parameterized as:

$$L = \frac{N_b^2 n_b f_{\text{rev}} \gamma}{4\pi \epsilon_n \beta^*} F, \quad (2.2)$$

where N_b is the number of protons per bunch, n_b is the number of bunches per beam, f_{rev} is the revolution frequency, γ is the relativistic Lorentz boost factor, ϵ_n is the normalized transverse beam emittance, β^* is the beta function at the collision point, and F is the geometrical luminosity reduction factor due to the crossing angle ($2 \times 120 \mu\text{rad}$) at the interaction point. The maximum designed luminosity of the LHC is up to $10^{34} \text{ s}^{-1} \text{ cm}^{-2}$. The side effects of the high luminosity are pile-ups, which means every bunch crossing contains a large number of proton-proton interactions, and the interaction can be affected by previous or later bunch crossings due to the slower detector response. The first one is described by the number of primary vertices, N_{PV} . The second one is described by the mean number of pp interactions per bunch crossing, $\langle \mu \rangle$.

Up to now, the LHC has delivered proton-proton collisions at center-of-mass energies of 7 TeV, 8 TeV and 13 TeV. The first LHC run was from 2010 to 2012, the so-called Run 1. During 2011, the LHC was operated to accelerate the protons to 3.5 TeV, with a center-of-mass energy at 7 TeV. In 2012, the beam energy was increased to 4 TeV, with a center-of-mass energy at 8 TeV. The bunch spacing of LHC Run 1 is 50 ns, and the number of bunches is 1380. The LHC delivered an integrated luminosity of 5.46 fb^{-1} for $\sqrt{s} = 7 \text{ TeV}$, and 22.8 fb^{-1} for $\sqrt{s} = 8 \text{ TeV}$. The ATLAS detector recorded 5.08 fb^{-1} and 21.3 fb^{-1} , among which 4.57 fb^{-1} and 20.3 fb^{-1} are good for physics studies. After Run 1, the LHC was shut down for upgrades to improve the accelerator reliability and detector performance. In 2015, the second LHC run started. The LHC Run 2 was from 2015 to 2018, with a center-of-mass energy at 13 TeV. The bunch spacing is reduced to 25 ns, with almost doubled number of bunches. During Run 2, the LHC delivered an integrated luminosity of 156 fb^{-1} . The ATLAS detector recorded 147 fb^{-1} , among which 139 fb^{-1} are good for physics studies. Figure 2.2 shows the cumulative

Years	center-of-mass energy	delivered luminosity	$\langle\mu\rangle$	bunch spacing
2011	7 TeV	5.46 fb ⁻¹	9.1	50 ns
2012	8 TeV	22.8 fb ⁻¹	20.7	50 ns
2015	13 TeV	4.2 fb ⁻¹	13.7	25 ns
2016	13 TeV	38.5 fb ⁻¹	24.9	25 ns
2017	13 TeV	50.2 fb ⁻¹	37.8	25 ns
2018	13 TeV	63.3 fb ⁻¹	36.1	25 ns

Table 2.1: Summary of center-of mass energy, integrated luminosity delivered by the LHC, out-of-time pile-up $\langle\mu\rangle$ and bunch spacing for the years with stable beams for proton-proton collisions [65].

luminosity versus time delivered by the LHC during the years with stable beams for proton-proton collision. Figure 2.3 shows the cumulative luminosity delivered by the LHC, recorded by the ATLAS detector, and certified to be good quality data in Run 2. Table 2.1 gives a summary of the luminosities, pile-up and bunch spacing during Run 1 and Run 2.

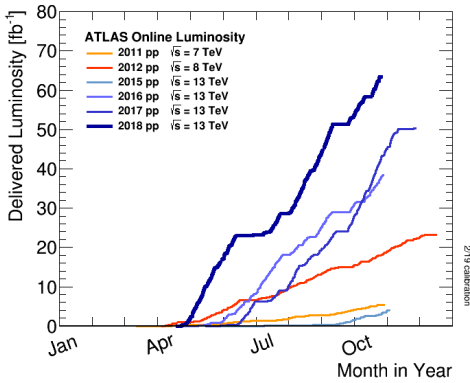


Figure 2.2: Integrated luminosity versus time delivered to ATLAS during the years with stable beams for proton-proton collisions [65].

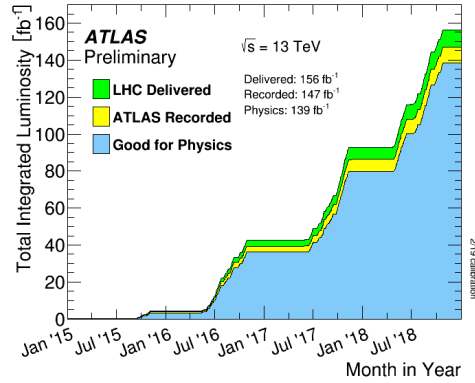


Figure 2.3: Integrated luminosity delivered by the LHC (green), recorded by the ATLAS detector (yellow) and certified to be good quality data (blue) in Run 2 [65].

2.2 The ATLAS experiment

The ATLAS detector is the largest detector at the LHC. It has a cylindrical geometry around the beam pipe. The detector is 44 m long, with a diameter of 25 m. A schematic view of the ATLAS detector is shown in Figure 2.4. The ATLAS detector contains several sub-detectors to provide measurements of different kinds of particles. In general, the sub-detectors are composed of a barrel region and two

2 The Large Hadron Collider and the ATLAS experiment

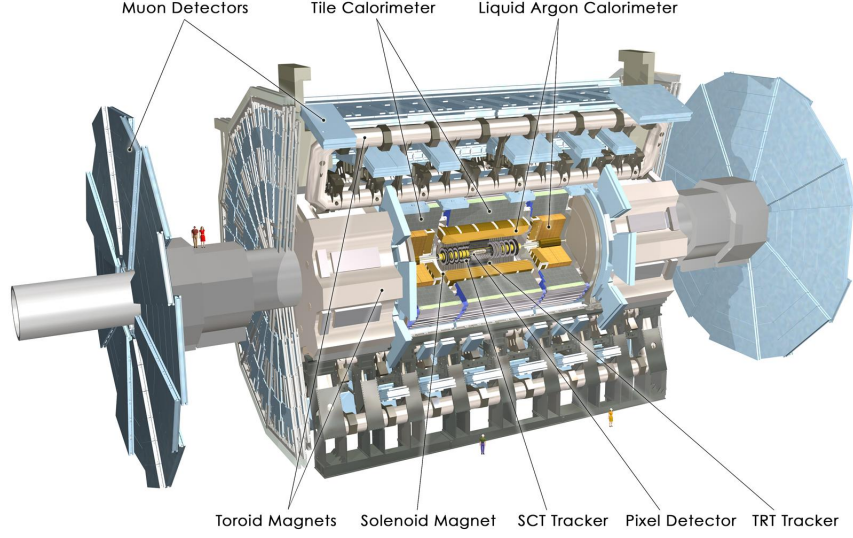


Figure 2.4: Cut-away view of the ATLAS detector.

forward/end-cap regions that are cylindrically symmetric around the beam pipe, covering the full solid angle of almost 4π . The sub-detector systems from the inner-most layer to the outer-most layer are:

- the Inner Detector, measures the trajectories and momenta of charged particles.
- the Calorimeter, including a Liquid Argon Calorimeter and a Tile Calorimeter, measures the energy of electrons, photons and hadrons.
- the Muon Spectrometer, measures tracks and momentum of muons.

The ATLAS detector also has a magnet system to provide magnetic fields for measurements of charged particle tracks. The ATLAS trigger and data acquisition system is developed to select events of interest for physics studies.

2.2.1 The coordinates and magnet system

The ATLAS detector uses a right-handed coordinate system. The interaction point (IP) is defined as the origin of the coordinate system. The z -axis is defined along the direction of the beam at the IP. The x - y plane is perpendicular to the beam. The x -axis points to the center of the LHC ring, and the y -axis points upwards. The coordinates commonly used to describe the position of the objects are the azimuthal angle ϕ and the pseudorapidity η . The pseudorapidity is defined from the polar angle θ ,

$$\eta = -\ln \left(\tan \frac{\theta}{2} \right). \quad (2.3)$$

The absolute value of this variable is mirror-symmetric about the x - y plane. Its value ranges from minus infinite to infinite, with $\eta \rightarrow \pm\infty$ along the beam and $\eta = 0$ on the x - y plane. The real rapidity Y is also used in some cases, the pseudorapidity is more widely used because it is easier to be measured

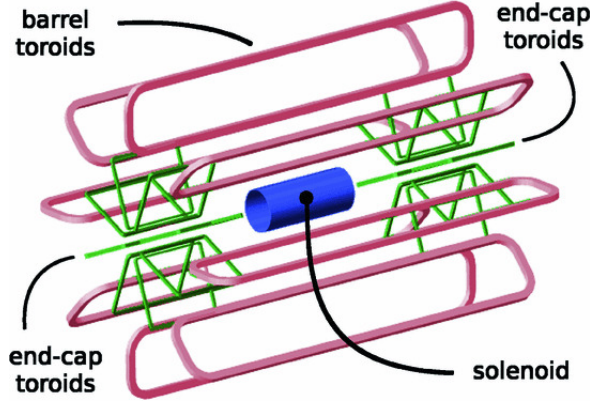


Figure 2.5: Geometry of the ATLAS magnet system: Central Solenoid Magnet (blue), Barrel Toroid (red), End-cap Toroids (green) [69].

and it is a geometric quantity independent from the particle. The angular distance between two objects in the ATLAS coordinate is defined as

$$\Delta R_{ij} = \sqrt{(\phi_i - \phi_j)^2 + (\eta_i - \eta_j)^2}, \quad (2.4)$$

where ϕ_i, ϕ_j are the azimuthal angle of the two objects, η_i, η_j are the pseudorapidity of the two objects.

The transverse momentum p_T is defined as the transverse component of a momentum vector. It is a very useful quantity for physics studies at the LHC. The colliding quarks and gluons only carry a fraction of the center-of-mass colliding energy, so the longitudinal component of the momenta of the original state is unknown. However, the transverse component of the momenta of the original state is very small. It can be approximated as zero compared to the energy of a hard scattering object. That is, the sum of the p_T of the final state objects should be zero. Missing transverse momentum E_T^{miss} is defined as the imbalanced momentum in the transverse plane,

$$E_T^{\text{miss}} = - \sum p_T, \quad (2.5)$$

where the sum runs over all the transverse momentum of detected objects. E_T^{miss} is very useful for studying final states with neutrinos because neutrinos are invisible to the detector. Given the sum of p_T of the final state objects equals to zero, E_T^{miss} can be approximated as p_T of the neutrinos.

The ATLAS magnet system contains two types of superconducting magnets, solenoidal and toroidal. The main sections are Central Solenoid Magnet [66], Barrel Toroid [67] and End-cap Toroids [68]. The solenoid magnet surrounds the inner detector to provide magnetic field for charged particle momentum measurements. The magnet is 5.6 m long, 2.56 m in diameter. It provides a 2 T magnetic field in just 4.5 cm thickness. The toroids use a series of eight coils to provide a 3.5 T magnetic field for momentum measurements of muons. Two toroids are at the end-caps of the muon detector, and one massive toroid surrounds the center of the ATLAS detector. The central toroid is 25.3 m in length, making it the largest toroidal magnet ever constructed. The end-cap toroids are 10.7 m in diameter. Figure 2.5 shows a sketch of the ATLAS magnet system.

2 The Large Hadron Collider and the ATLAS experiment

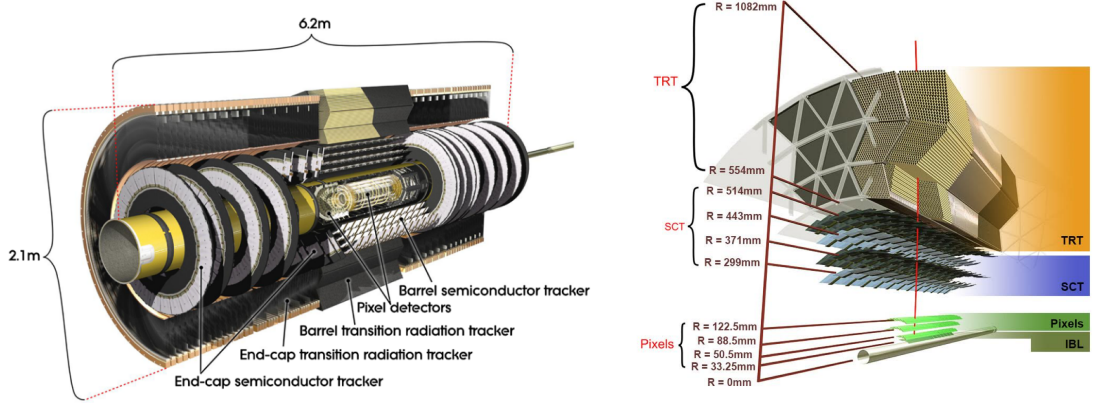


Figure 2.6: Structure of different layers of the ID. From the inside to the outside there are the IBL, the Pixel Detector, the SCT and the TRT [72].

2.2.2 Inner Detector

The Inner detector (ID) [70–72] is the closest sub-detector to the beam pipe. It is designed to measure tracks of the charged particles and vertices in the collisions with high resolution. The whole ID is immersed in the 2 T magnetic field generated by the central solenoid magnet, so the detector can measure the momenta and charge of the particle by bending their trajectories.

Given the very large track density at the LHC, the ID has to be constructed from fine-granularity detectors to achieve the high resolution measurements. The silicon detectors using pixel and microstrip technology offer these features. The ID comprises three types of detectors, moving from the inside to the outside there are the Pixel Detector, the Semi-Conductor Tracker (SCT), and the Transition Radiation Tracker (TRT). During the shutdown after Run 1, the Insertable B-Layer (IBL) was constructed, inserted and commissioned to become an additional (innermost) layer of the existing Pixel Detector. Figure 2.6 shows the structure of the ID.

Pixel Detector

The original pixel detector [73] had three silicon pixel layers, the B-Layer, Layer 1 and Layer 2, covering radial distances between 50.5 mm and 150 mm. During the long shutdown between Run 1 and Run 2, the IBL [74] was installed between the B-Layer of the original pixel detector and the new beam pipe with smaller radius. It becomes the detector closest to the beam, with radial distance at 33.25 mm. The IBL consists of high resolution pixel sensors, providing a coverage of $|\eta| < 3.0$. The pixel sensors have a pixel size of $50 \times 250 \mu\text{m}^2$ in the IBL, the longer side being along the z -axis. The sensors are inclined in azimuth by 14° to achieve an overlap of the active area. The other layers are consisted of pixel sensors with a pixel size of $50 \times 400 \mu\text{m}^2$. The pixel sensors are arranged in three barrel layers and three disks for each end-cap, providing an acceptance range of $|\eta| < 2.5$. The hit position resolution of the pixel sensors is about $10 \mu\text{m}$ in the transverse direction, and $67 \mu\text{m}$ in the longitudinal direction for the IBL [75], $115 \mu\text{m}$ in the longitudinal direction for other layers [76].

The upgraded pixel detector has improved performance. The additional layer with smaller pixel cell length provides high-precision hits closer to the interaction points, particularly important for improving the resolution of low p_T particles and vertex reconstruction. The enhanced primary and secondary vertex reconstruction can also improve the performance of b -tagging, which is important for $VH(b\bar{b})$ analysis.

SCT

The SCT is a silicon microstrip detector surrounding the pixel detector. This detector is designed to measure tracks in the intermediate region within radial distances from 299 mm to 560 mm. The barrel and the end-caps provide a coverage of $|\eta| < 2.5$. The silicon microstrips are arranged in four layers of concentric barrels and two end-caps with nine wheels each. This arrangement provides typical eight strip measurements (four space points) for tracks. In the barrel region, the strips are approximately parallel to the beam axis, and have a size of $80\ \mu\text{m} \times 12.8\ \text{cm}$. Each layer of the concentric barrel is composed of two sets of identical sensors, the second set is glued back-to-back with the first set at a stereo angle of 40 mrad to provide longitudinal space points. The barrel layers are placed at radial distance of 30 cm, 37 cm, 44 cm and 51 cm from the beam axis. The modules on the nine end-cap wheels are all trapezoidal in shape but different in size. The end-cap region requires four types of module size to provide full coverage. The modules are placed on three rings on each disk: the inner ring, the middle ring and the outer ring. The outermost disk only has its outer ring populated with modules. The microstrips have a hit position resolution of $17\ \mu\text{m}$ in the transverse direction, and $580\ \mu\text{m}$ in the longitudinal direction [76].

TRT

The TRT is the outermost sub-detector of the three tracking system of the ID. It is designed for electron identification as well as continuous tracking of charged particles. The detector contains nearly 300000 proportional drift tubes (straws) filled with gas mixture with xenon (70% Xe, 27% CO₂, 3% O₂). The straws in the barrel region are 4 mm in diameter, 150 cm in length, arranged in three cylindrical layers and 32 ϕ sectors. The straws in the end-cap regions are 40 cm long, arranged perpendicularly to the beam line. The sensitive volume covers radius distances between 0.5 m to 1.1 m and pseudorapidity ranges of $|\eta| < 2.0$. The hit position resolution of the straws is about $120\ \mu\text{m}$, not as good as the silicon detectors. However, the large number of tubes provide an average of 30 space points for a track, making the momentum resolution of TRT competitive with that of the silicon detectors.

The space between the straws are filled with polymer fibers or foils, highly relativistic charged particles create transition radiation when they pass through a material boundary. The effects of transition radiation depend on the relativistic factor $\gamma = E/m$, so the effects are much stronger for electrons than other particles like pions. The X-ray photons generated by the electrons are absorbed by the Xe atoms in the tubes, leading to significantly higher readout. Thus the TRT can be used to discriminate electrons and hadrons.

2 The Large Hadron Collider and the ATLAS experiment

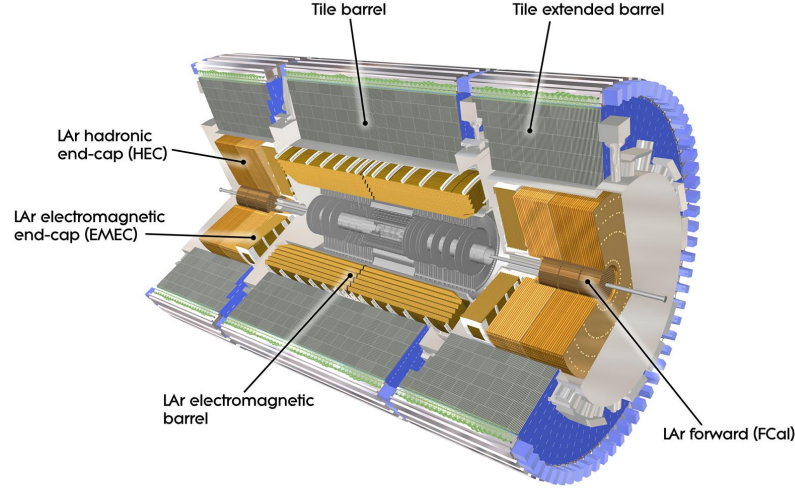


Figure 2.7: Cutaway view of the ATLAS calorimeter system.

With all the modulars combined, the ID is able to measure the tracks of charged particles precisely. The momentum and charge of the particles can be determined from the trajectories. Track reconstruction performance of the ID have been evaluated using Run 2 data and cosmic ray muons [77, 78]. For a particle with transverse momentum p_T , the resolution of p_T is

$$\frac{\sigma_{p_T}}{p_T} = 0.05\% |p_T (\text{GeV})| \oplus 1\%. \quad (2.6)$$

The resolution of the transverse impact parameter d_0 is around $20 \mu\text{m}$, d_0 is defined as the shortest distance between a track and the beam line in the transverse plane. More details about track reconstruction performance are described in Section 3.1.

2.2.3 Calorimeters

The ATLAS calorimeter system is designed to measure the energy of particles except muons and neutrinos. The system is composed of two types of detectors, the Liquid Argon (LAr) calorimeter [79] and the Tile hadronic calorimeter [80]. The LAr calorimeter technology is used to measure the energy of electromagnetic (EM) showers as well as hadronic showers. The Tile calorimeter measures the energy of hadronic particles which do not deposit all the energies in the LAr calorimeter. The calorimeter system is illustrated in Figure 2.7.

Liquid Argon calorimeter

The Liquid Argon (LAr) calorimeter surrounds the Inner detector. It is a sampling detector consisting of layers of metal (tungsten, copper or lead), with liquid argon between the layers as the sensing element. The incoming high energy particles are absorbed by the metal layers and convert into a shower

2.2 The ATLAS experiment

of lower energy particles. These particles ionize the liquid argon, producing an electric current that is measured. The lead-LAr calorimeter is used for EM shower measurements. The copper-LAr and tungsten-LAr calorimeter is used for hadronic shower measurements.

The electromagnetic showers are initiated by high energy electrons through Bremsstrahlung and photons through pair creation. Low energy electrons lose energy mainly through ionization and low energy photons progressively lose energy through Compton scattering. Electrons and photons are the sole components of the showers. The development of the EM shower is parameterized with radiation length (X_0). The electromagnetic part of the calorimeter is consisted of the lead-LAr electromagnetic barrel (EMB) and the two lead-LAr electromagnetic end-caps (EMEC).

The barrel of the LAr calorimeter is 6.8 m long, with an outer radius of 2 m, and an inner radius of 1.4 m, covering $|\eta| < 1.475$. It contains three sampling layers, as shown in Figure 2.8. The first layer is consisted of fine-grained strips with a very fine longitudinal segmentation $\Delta\eta \times \Delta\phi = 0.003 \times 0.1$, which is ideal for precise measurements of electrons and photons. The second layer collects the largest fraction of the electromagnetic shower with its $17 X_0$ depth. The dimension of the cells in this layer is $\Delta\eta \times \Delta\phi = 0.025 \times 0.025$. The third layer collects the tail of the electromagnetic shower, with a cell size of $\Delta\eta \times \Delta\phi = 0.05 \times 0.025$. In addition to the sampling layers, the EMB has a presampler to correct for the energy lost in front of the calorimeter, covering $|\eta| < 1.52$.

The EMEC is 63 cm thick, with a radius of 2.077 m, providing an acceptance range of $1.375 < |\eta| < 3.2$. It also has three sampling layers similar to the barrel region, and a presampler covering $1.5 < |\eta| < 1.8$. Beyond a pseudorapidity of 1.8, the presampler is no longer needed given the limited amount of dead material and the higher energy of particles. The total active thickness of the electromagnetic calorimeter ranges between $22 X_0$ and $33 X_0$ in the barrel region, and between $24 X_0$ and $38 X_0$ in the end-cap regions.

The hadronic showers are initiated by hadrons through strong and electromagnetic interactions. It always has a mixed electromagnetic component. The showers contain a large variety of particles. The development of the shower is described by the nuclear interaction lengths (λ). The hadronic part of the LAr calorimeter consists of two copper-LAr end-caps and the forward calorimeters. Each of the hadronic end-caps (HEC) contains two wheels that are 0.8 m and 1.0 m thick, with outer radius of 2.03 m, providing an acceptance range of $1.5 < |\eta| < 3.2$. Each wheel is composed of two sampling layers, giving a total of four layers per end-cap for the HEC.

The forward calorimeter (FCal) has three layers with thickness of 0.45 m and radius of 0.455 m each. The first layer is instrumented with copper-LAr calorimeters. It is used for electromagnetic energy measurement, the thickness of this module corresponds to $28 X_0$. The other two module, made of tungsten-LAr calorimeters, are used for hadronic energy measurement. The thickness of these two modules corresponds to $180 X_0$ or 2.7λ . The FCal provides an acceptance range of $3.1 < |\eta| < 4.9$.

Tile hadronic calorimeter

The Tile hadronic calorimeter [81] surrounds the LAr calorimeter. It provides precise measurements of jets, hadrons, taus and E_T^{miss} . The calorimeter is divided in three sections, the long central barrel

2 The Large Hadron Collider and the ATLAS experiment

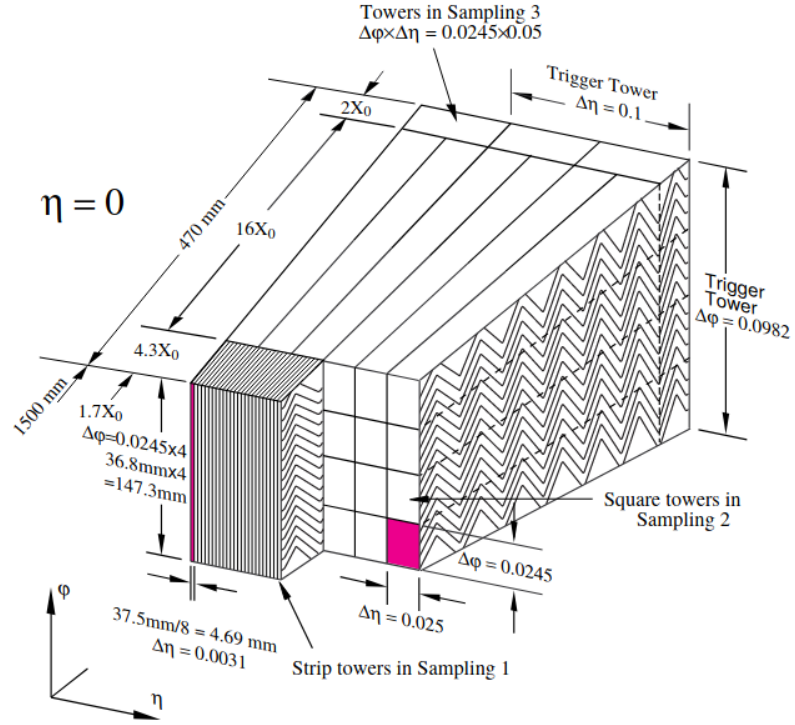


Figure 2.8: Sketch of the calorimeter layers [79].

that covers $|\eta| < 1.0$ and the two extended barrels that cover $0.8 < |\eta| < 1.7$. Each cylinder contains 64 modules, these modules are sampling detectors using steel as the absorber and scintillating tiles as the sensing element. Figure 2.9 is a schematic drawing of the mechanical assembly and the readout of a module.

The central barrel is 5.8 m in length, the two extended barrels are on the sides of the central barrel, each 2.6 m long. The detector has an inner radius of 2.28 m and an outer radius of 4.25 m. The calorimeter consists of three radial layers, with depths of 1.5, 4.1, 1.8 λ for the barrel, and 1.5, 2.6, 3.3 λ for the end-caps. The size of the cells for the first two layers is $\Delta\eta \times \Delta\phi = 0.1 \times 0.1$, for the third layer is $\Delta\eta \times \Delta\phi = 0.2 \times 0.1$. The scintillating tiles are placed perpendicular to the colliding beams and are radially staggered in depth. The signal is read-out by wave-length shifting fibers installed on both sides of a module, and delivered to photo-multipliers (PMTs) located at the outermost side of the calorimeter.

The performance of the calorimeter has direct effects on the reconstruction of physics objects. Good energy resolution is essential for the precise measurements of jets and E_T^{miss} . The energy resolution of a sampling calorimeter can be parameterized as

$$\frac{\sigma_E}{E} = \frac{a}{\sqrt{E \text{ (GeV)}}} \oplus \frac{b}{E \text{ (GeV)}} \oplus c, \quad (2.7)$$

where a is the stochastic term that describes the statistical fluctuations of the shower, b is the noise term accounts for the electronic noise of the detector system, c is a constant that takes into account

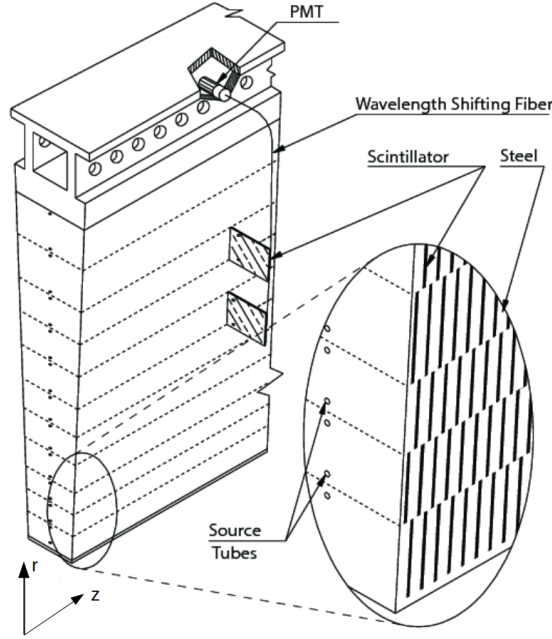


Figure 2.9: Sketch of the mechanical assembly and the readout of Tile Calorimeter modules [82].

the non-uniformity of the calorimeter like non-perfect alignment. The energy resolution of the sub-modules¹ are studied in test beam experiments using electrons and pions. The resolution of EMB and EMEC for electrons is

$$\frac{\sigma_E}{E} = \frac{10\%}{\sqrt{E \text{ (GeV)}}} \oplus \frac{0.17}{E \text{ (GeV)}} \oplus 0.7\%. \quad (2.8)$$

The resolution of Tile calorimeter and HEC for pions is

$$\sigma_E/E = 50\%/\sqrt{E \text{ (GeV)}} \oplus 6\%. \quad (2.9)$$

The resolution of the forward calorimeter for pions is

$$\sigma_E/E = 100\%/\sqrt{E \text{ (GeV)}} \oplus 10\%. \quad (2.10)$$

This level of precision guarantees good sensitivity for measurements of physics processes at the TeV scale.

2.2.4 Muon Spectrometers

The ATLAS Muon Spectrometer (MS) [83] is designed to measure the muon momenta and provide fast signal for muon trigger. The MS surrounds the calorimeter and it is the outermost sub-detector of the ATLAS detector. It measures the muon momenta by bending the muon trajectories with magnetic

¹ Numbers taken from Ref. [79], [80].

2 The Large Hadron Collider and the ATLAS experiment

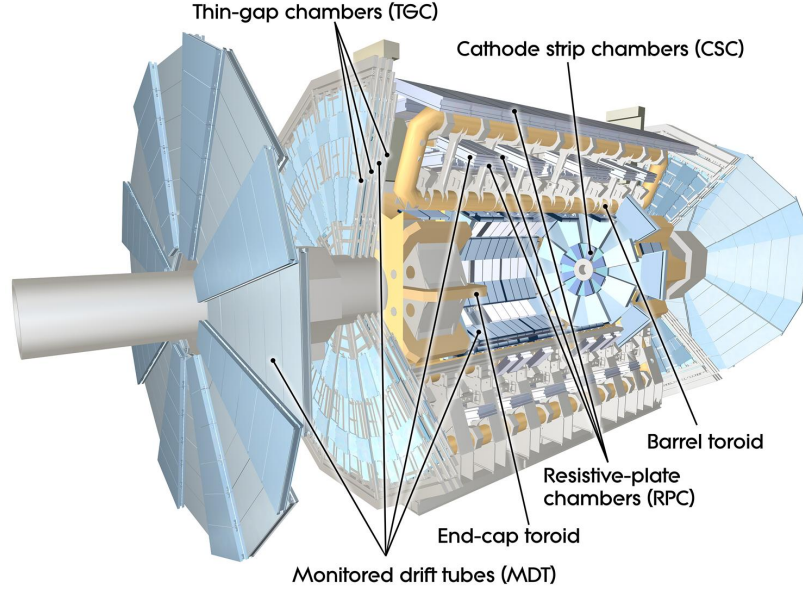


Figure 2.10: Cut-away view of the ATLAS Muon Spectrometer.

field generated by the toroids. To measure muon tracks with high precision and provide fast signal for triggering, the muon detector is instrumented with four types of detectors. The Monitored Drift Tubes (MDT) and the Cathode Strip Chambers (CSC) provide precision measurement of muon tracks, the Resistive Plate Chambers (RPC) and the Thin Gap Chambers (TGC) provide fast signal for the muon trigger system. Figure 2.10 shows the cut-away view of the MS.

The MS has a barrel region and two end-cap regions. In the barrel region, the MDT chambers are arranged in three cylindrical layers around the z -axis, at radial distance of approximately 5 m, 7.5 m and 10 m from the beam axis. The inner, middle and outer barrel layers are labeled as BIL/BIS, BML/BMS and BOL/BOS respectively. In the end-cap regions, there are four layers of MDT chambers, placed at distances of approximately 7 m, 11 m, 14 m and 21 m from the interaction point along the z -axis. The end-cap layers are labeled as inner (EIL/EIS), extra (EEL/EES), middle (EML/EMS) and outer (EOL/EOS). In the innermost end-cap region, CSC chambers are used because of their capability to cope with higher background rates. The RPC chambers are used for the muon trigger system in the barrel region, and the TGC chambers are used in the end-cap regions. Schematic drawings of the MS after LS1 (Run 2) are shown in Figure 2.11 and Figure 2.12.

The MDT chambers are deployed for most of the acceptance range, covering $|\eta| < 2.7$ except in the innermost end-cap, where the acceptance range is $|\eta| < 2.0$. CSC chambers are used to cover the larger pseudorapidity region $2.0 < |\eta| < 2.7$ in this layer. The MDT chambers are composed of two multi-layers, each formed by three or four layers of drift tubes. The gas mixture used is 93% Ar and 7% CO₂ with a small admixture of water vapor. The average space resolution obtained with a single tube is 80 μm , and with a multi-layer is about 50 μm . The CSC chambers are multi-wire proportional chambers with cathode strip read out. The gas mixture used in CSC is 80% Ar and 20% CO₂. Typical resolution obtained with this read out scheme is about 50 μm [85].

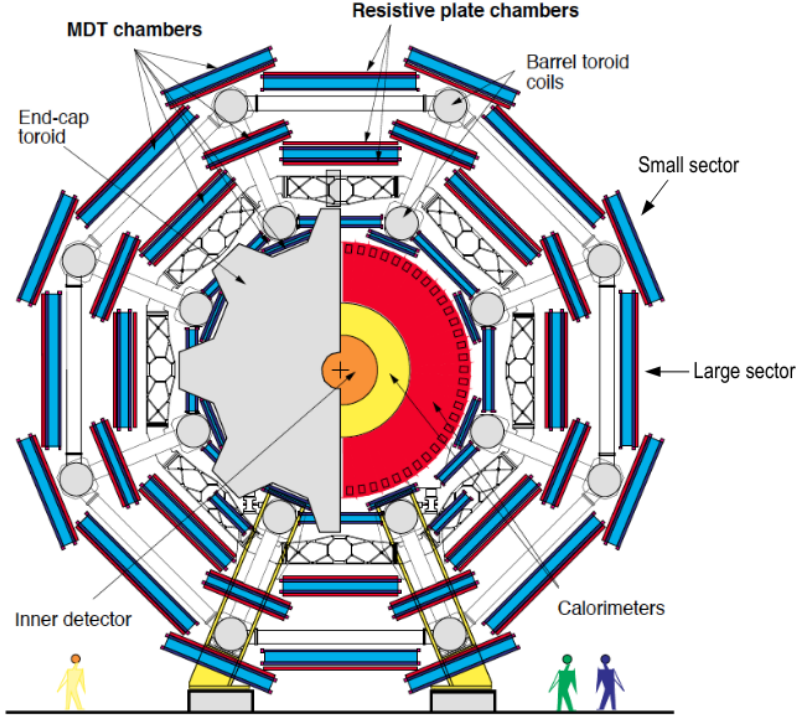


Figure 2.11: Schematic view of the ATLAS Muon Spectrometer in the x - y plane, transverse to the beam axis [84].

There are three layers of RPC chambers in the barrel region covering $|\eta| < 1.05$. Two layers of RPC chambers are installed on the sides of the BML MDT chambers to provide a low p_T trigger signal. The other layer is installed on the outer-side of the BOL MDT chamber, providing a high p_T trigger. The RPC chambers cover the barrel region with an acceptance range of $|\eta| < 1.05$. Each RPC chamber consists of two parallel Bakelite plates, the gap in-between is filled with gas mixture of 94.7% $C_2H_2F_4$, 5% C_4H_{10} and 0.3% SF_6 . The RPCs have a time resolution of about 1 ns. In the end-cap regions, two TGC doublets, one TGC triplet and an additional inner TGC layer are installed, giving a total of 3-4 layers of TGC chambers. The doublets are placed close to the outer-side of the EML MDT chambers, providing the low p_T trigger. The triplet is installed on the inner-side of the EML MDT chambers to provide the high p_T trigger. The inner TGC layer (TGC-FI) is placed on the inner-side of the EIL MDT chambers, close to the Tile calorimeter end-caps, to reject particles not originating from the interaction point. The TGC chambers cover the region $1.05 < |\eta| < 2.4$. The TGCs are multi-wire proportional chambers using 55% CO_2 and 45% n -pentane gas mixture, the time resolution is about 4 ns. The RPCs and TGCs also measure the coordinate of muons along the MDT wires.

The muon tracks can be reconstructed from the ID and the MS. The resolution of muon p_T measured in the MS is

$$\frac{\sigma_{p_T}}{p_T} = 0.02\% |p_T \text{ (GeV)}| \oplus 3.8\%. \quad (2.11)$$

When combined with the ID, the muon p_T resolution is better than 4% over a large range in momentum, and increases to around 10% at 1 TeV [77, 86]. This level of precision allows both SM measurements

2 The Large Hadron Collider and the ATLAS experiment

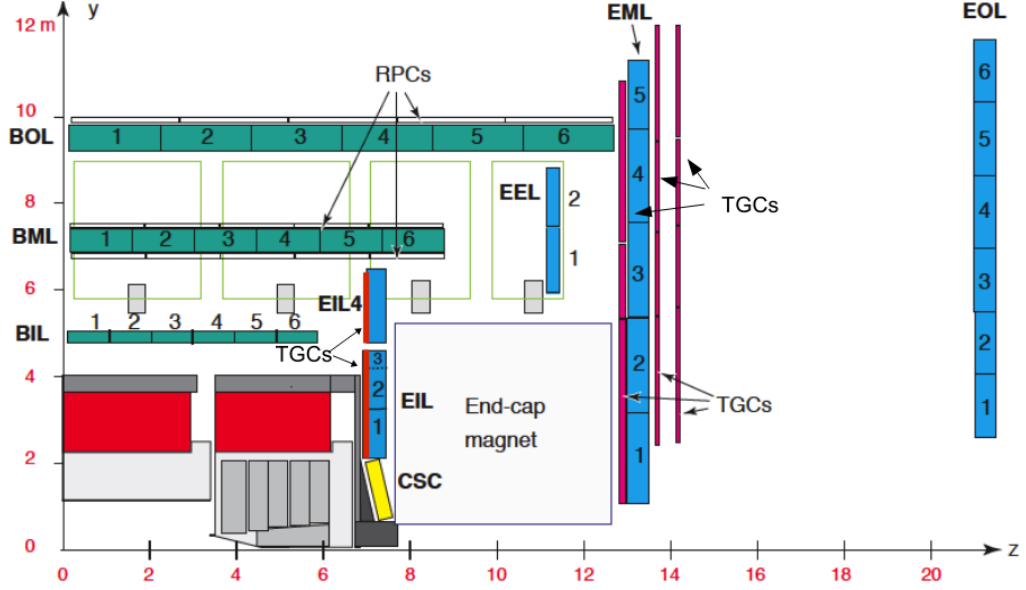


Figure 2.12: Schematic view of the Run 2 ATLAS Muon Spectrometer in r - z projections [84]. The cyan-green colored blocks are MDT chambers in the barrel region, the blue colored blocks are MDT chambers in the end-cap regions. The CSCs are shown in yellow. The TGCs are shown in red and pink. The RPCs are shown in white.

and new physics searches in processes containing high momentum muons in the final states. The muons are reconstructed from a combination of the ID and the MS. For muons with $p_T > 100$ GeV, the momentum resolution is mainly affected by the MS, and the drift tube hit resolution becomes the dominate factor above 250 GeV. For low p_T muons, the momentum resolution is dominated by the ID due to the energy loss in the calorimeter. With the combination, the muons can be reconstructed precisely in a wide p_T range from 3 GeV to 3 TeV.

2.2.5 Trigger and data acquisition system

The ATLAS Trigger and Data Acquisition (TDAQ) system [87] is designed to collect the electric signal of the detector. The proton-proton collisions take place at a very high rate, but only a small fraction of the collisions contain hard scattering processes that are useful to the physics analyses. It is impractical to record all the collision data due to limited computing resources. The TDAQ system performs a selection to decide whether an event could be interesting, and only the selected events will be recorded. During Run 2 of the LHC, the collisions take place every 25 ns, at a rate of 40 MHz, with a peak luminosity of up to $10^{34} \text{ cm}^{-2}\text{s}^{-1}$. The ATLAS trigger system has two levels, a hardware based Level-1 (L1) trigger which selects events at a rate of 100 kHz, and a software based High-Level trigger (HLT) which further selects events at a rate of 1 kHz. The structure of the ATLAS TDAQ system is shown in Figure 2.13.

The L1 trigger is based on detector modules that are able to provide fast signal for objects like electrons,

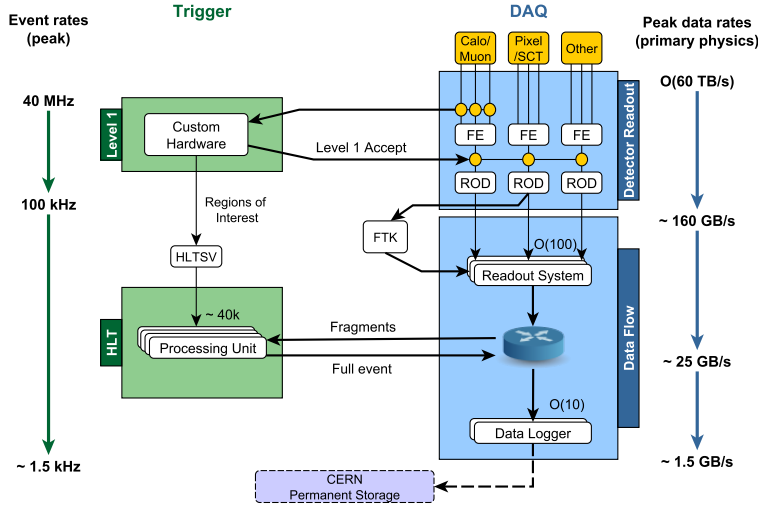


Figure 2.13: Functional diagram of ATLAS Trigger and Data Acquisition system in Run 2 [88].

photons, taus, jets and muons. Inputs from the reduced granularity sub-modules of the calorimeter and the muon spectrometer are used to make the decision. The L1 trigger system is formed by the L1 calorimeter (Calo) trigger and the L1 Muon trigger, as well as the L1 Central Trigger Processor (CTP), which takes inputs from the L1 Calo and the L1 Muon triggers.

The L1 Calo trigger processes signals from the calorimeter. The trigger algorithms identify Regions-of-Interest (RoIs) in the detector modules. For electron/photon and tau trigger algorithms, the RoIs are 2×2 trigger tower clusters in the electromagnetic calorimeter. The sum of the transverse energy from at least one of the four possible pairs of nearest neighbor towers (1×2 or 2×1) has to exceed a predefined threshold. The E_T threshold is set differently for different η regions with a granularity of $\Delta\eta \times \Delta\phi = 0.1 \times 0.1$. The jet trigger RoIs are 4×4 or 8×8 trigger tower clusters, the summed electromagnetic and hadronic transverse energy also needs to exceed predefined thresholds.

The L1 Muon trigger processes signals from the RPC and the TGC chambers of the MS. There are three RPC layers in the barrel region as described in Section 2.2.4. During the long shutdown after Run 1 (LS1), some outer layer RPC chambers in the projective region of the acceptance holes were commissioned and became operational in Run 2. The TGC chambers are arranged in four layers, the two doublets and the triplet were already operational in Run 1. The inner TGC layer was added during LS1, covering $1.3 < |\eta| < 1.9$.

The L1 trigger decision is made by the L1 CTP. The CTP receives signals from the L1 Calo and the L1 Muon triggers, and performs selection based on the geometric or kinematic association between the trigger objects. Calculation of global event quantities such as E_T^{miss} are performed at the L1 CTP. The completed L1 trigger system is able to make decision in less than 2.5 μs .

The HLT trigger system is a set of software based algorithms running offline. The trigger algorithms take inputs from the L1 CTP, and process events with full granularity information which are not

2 The Large Hadron Collider and the ATLAS experiment

available at L1. The inner detector tracking algorithm provides electron trigger and muon trigger at low- p_T . The calorimeter reconstruction algorithms convert electronic signals into cell energies and clusters, which are used as discriminants for particle identification and backgrounds rejection. The cell energies and clusters also serve as inputs to reconstruction of the physics objects including electrons, photons, tau leptons, jets and E_T^{miss} . The muon tracking algorithms make use of L1 trigger together with precision measurements from the MDT and CSC chambers to further refine muon candidates.

Finally, the data collected for physics analyses must be of good quality. The ATLAS DAQ system operates at the L1 trigger output rate. The events passing the L1 trigger are temporarily stored and read-out by the detector modules as the Read-Out Drivers (RODs). The DAQ system receives inputs from the RODs, then with further selection from the HLT system, the event rate is reduced to around 1 kHz. The selected events are sent to permanent storage for physics analyses.

3 Object reconstruction

The electric signals recorded by the ATLAS detector are processed by several event reconstruction algorithms to identify particles and determine their kinematics. The objects are used to select events with certain final states and kinematic signatures for physics analyses. This section describes the reconstruction of physics objects from detector signals. Tracks are reconstructed from signals in the ID. Energies of electrons and photons are measured by the EM calorimeter. Hadrons are absorbed by the hadronic calorimeter. Muons leave signals in the ID and the MS. Neutrinos do not leave observable signals in the ATLAS detector, but they can be identified through the imbalance of the total transverse momentum of the event, which is the missing transverse momentum. Figure 3.1 illustrates the passage of some particles in the ATLAS detector system.

3.1 Charged particle track reconstruction

Track is the trajectory of a charged particle passing through the ID and the MS. The reconstructed tracks are used in almost every element of reconstruction, including primary vertices, leptons, pile-up

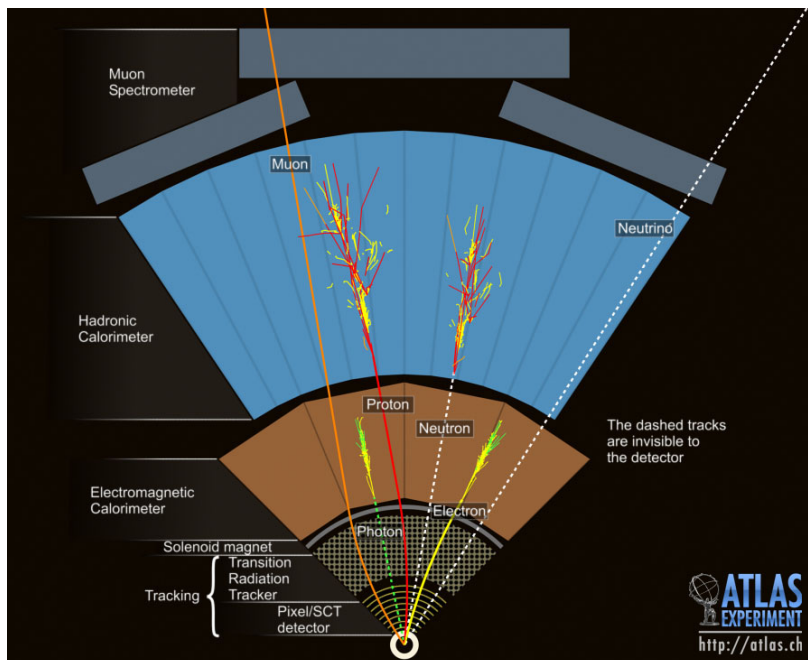


Figure 3.1: Overview of particles passing through the ATLAS detector.

3 Object reconstruction

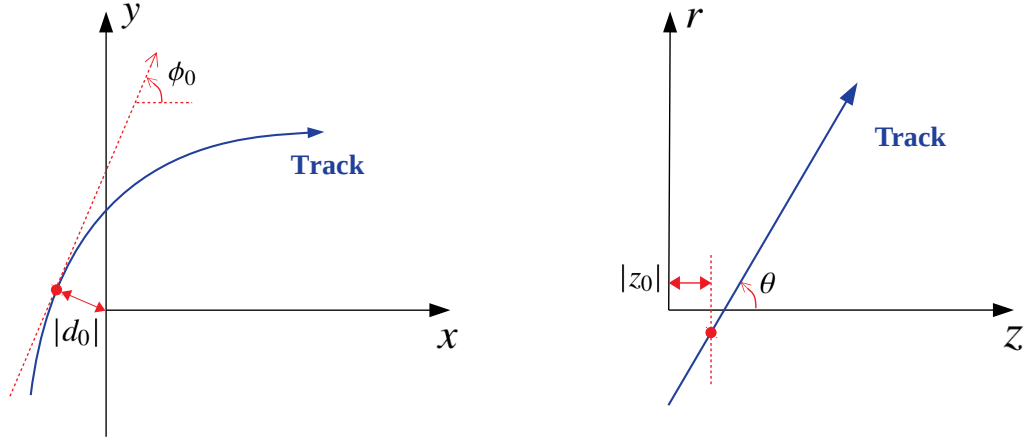


Figure 3.2: Illustration of the track parameters in the transverse plane (left) and the $r - z$ plane (right).

removal for jets and E_T^{miss} , jet flavor tagging. The high luminosity of the LHC puts stringent challenges to the track reconstruction algorithm. Each event could contain around 600 tracks in Run 2. The algorithm needs to reconstruct all the tracks with high efficiency and low fake rate. The current ATLAS track reconstruction algorithm follows the New Tracking (NEWT) approaches [89]. It is a software system that integrates a lot of well performing reconstruction algorithms. NEWT has been the default ATLAS ID reconstruction program and many components of NEWT are also used in MS track reconstruction.

A track of a charged particle in a magnetic field requires five track parameters ($\mathbf{q}$) to describe:

$$\mathbf{q} = (d_0, z_0, \phi_0, \theta, q/p), \quad (3.1)$$

where d_0 is the transverse impact parameter, the absolute value of d_0 is defined as the shortest distance between a track and the beam line in the transverse plane; z_0 is the longitudinal impact parameter, defined as the z coordinate of the point on the track to evaluate d_0 . The transverse impact parameter is signed to further discriminate the track in b -tagging. The sign is defined as positive (negative) if the track intersects the jet axis, to which it belongs, in front of (behind) the primary vertex along the flight direction. The azimuthal angle ϕ_0 describes the direction of the track at the point to evaluate d_0 in the transverse plane. The polar angle θ describes the direction of the track in the $r - z$ plane. q/p is the ratio of the charge and the momentum of the particle. An illustration of the track parameters is shown in Figure 3.2.

Reconstruction of the tracks

The reconstruction of tracks starts from grouping pixels and strips into clusters. The energy deposits of a charged particle are usually collected in multiple pixels. The average size of pixel clusters is about two pixels in the $r - \phi$ plane and from one to three pixels in the longitudinal direction. The

3.1 Charged particle track reconstruction

pixels with energy deposits above a certain threshold are taken into account. The reconstruction algorithm creates three-dimensional position measurements from the clusters, which is referred to as space-points. In the pixel detector, each cluster equates to one space-point. In the SCT, the two strip clusters on both sides of a layer are combined to form one space-point. Then, track seeds are formed with combinations of three space points. Different combinations of space-points in the pixel detector and the SCT are considered. A combinational Kalman filter [90] is applied to build track candidates from the track seeds. The remaining space-points are incorporated into the track candidate. The Kalman filter creates multiple track candidates if more than one compatible trajectories exist.

The best track candidate is selected with the ambiguity solver. In the ambiguity solver, the track candidates are evaluated individually and assigned with a score. The score increases when the cluster resolution and the fitting of track become better, decreases when a sensitive detector element along the track does not provide a matched cluster. Such sensitive elements without a matching cluster are the so-called "holes". With the scores, the ambiguity solver deals with clusters used by multiple track candidates. A cluster cannot be shared by more than two tracks, and one track cannot contain more than two shared clusters. The clusters that do not meet this requirement are removed, and the track score is re-calculated. A set of quality criteria is applied to the track candidate at this stage, any track candidate that do not meet the requirement are removed. Finally, a high resolution fit is performed on the remaining track candidates, with the extended TRT hits to estimate the final track parameters.

Performance of the track reconstruction

The performance of track reconstruction algorithm has been studied with Run 2 data and MC simulated samples. Two sets of quality criteria, referred to as Loose and Tight Primary, are studied. The Loose selection criteria allows more tracks to be kept but comes with higher fake rate. The Tight Primary criteria has lower tracking efficiency but improves fake rejection. The tracking efficiency is defined as the ratio of the number of reconstructed tracks and the number of truth charged particles. It is measured in bins of truth particle p_T and η . Figure 3.3 shows the tracking efficiency as a function of truth p_T and η . The tracking efficiency reaches to 90% and 85% in central η region for Loose and Tight Primary selections, respectively. Not all reconstructed tracks correspond to a charged particle traversing the detector. Such tracks are the so-called "fake" tracks. The fake rate is estimated by assuming the number of non-fake tracks is to first order proportional to the number of pile-up interactions. Any deviation from the linearity is considered to be caused by fake tracks. The estimated fake rate increases with pile-up. It is negligible at $\langle\mu\rangle = 0$, increases to around 5% at $\langle\mu\rangle = 25$ for Loose selection. The fake rate can be controlled at very low level with Tight Primary selection [91].

The impact parameter resolution is measured with respect to the primary vertex. The resolution of the primary vertex is unavoidably folded into the measurement, so the impact parameter resolution needs to be unfolded from its reference point. The results of the unfolded resolution measurements is shown in Figure 3.4. The resolution of d_0 varies from 20 μm to 120 μm . The resolution of z_0 differs vastly in different η regions. The resolution measured in simulation is slightly better than in data, due to the discrepancies in material description used in simulation and simplified energy deposit model.

3 Object reconstruction

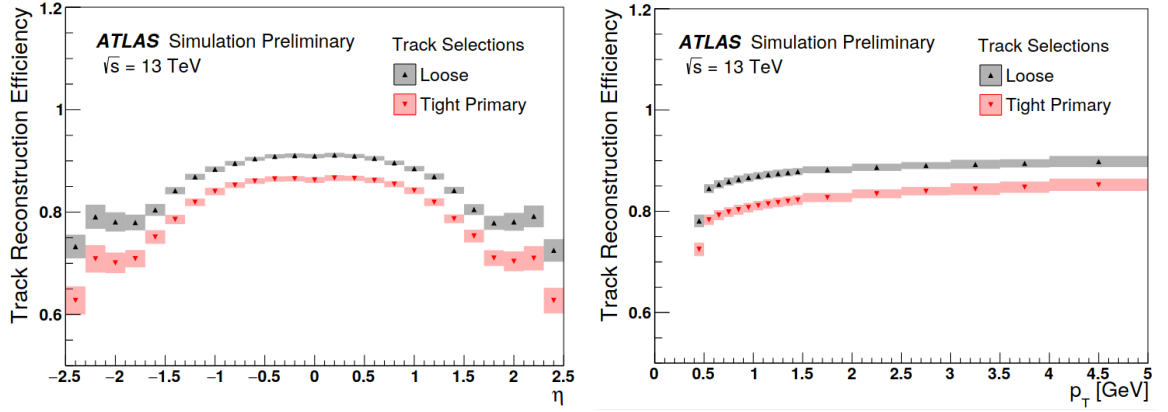


Figure 3.3: Tracking efficiency evaluated with early Run 2 data and minimum bias MC simulation, as a function of truth particle p_T and η [91].

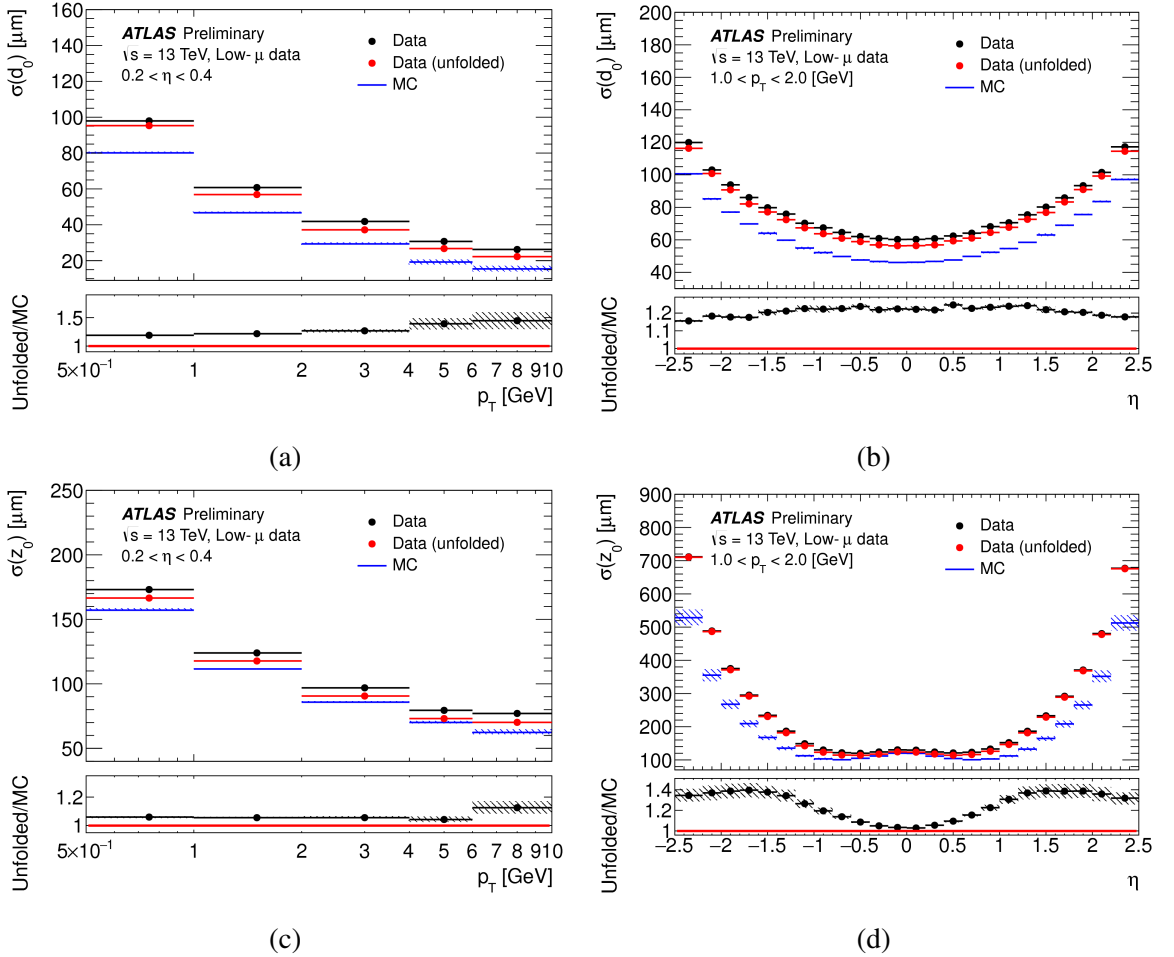


Figure 3.4: Resolution of transverse impact parameter d_0 (a, b) and longitudinal impact parameter z_0 (c, d) as a function of p_T (for $0.2 < \eta < 0.4$) and η (for $1 < p_T < 2$ GeV) [78].

3.2 Primary vertex

Primary vertices are the primary interaction points of the pp collisions. The reconstruction of primary vertex is based on the reconstruction of charged particle tracks. The process is divided into two stages: the primary vertex finding algorithm that associates tracks to vertex candidates, and the vertex fitting algorithm that reconstruct the vertex position and its covariance matrix. For vertex reconstruction in Run 2, a very tight selection is first applied to select quality tracks [92]:

- $p_T > 400$ MeV, $|\eta| < 2.5$,
- number of silicon hits larger than 9 for $|\eta| \leq 1.65$, larger than 11 for $|\eta| > 1.65$,
- at least 1 IBL hits or B-layer hits,
- No more than 1 shared pixel hit or 2 shared SCT hits,
- No pixel holes,
- No more than 1 SCT hole.

A vertex candidate requires at least two tracks passing these selection criteria associated to it. The primary vertex finding procedure is an iterative process starting from vertex seed position determined with the beam spot constraint. After the seed has been determined, the vertex finding algorithm performs an adaptive vertex fitting with the seed position and the track parameters. The selected tracks are assigned with weights to reflect its compatibility with the vertex seed. The vertex position is then re-calculated with the weighted tracks. After the last iteration, the tracks found incompatible with the vertex are returned as unused tracks for the next vertex finding iteration.

The vertex reconstruction efficiency has been measured from Run 2 data. It is evaluated from the ratio between events with a reconstructed vertex and events with at least two quality tracks. Figure 3.5 shows the vertex reconstruction efficiency measured from low pile-up (low- μ) data, compared to MC simulations. The efficiency is higher than 99% for events with more than four tracks. The vertex reconstruction performance was extended to the high pile-up regime using data collected in Run 1 [93], with values of $\langle\mu\rangle$ between 55 and 72. The simulation samples spanned values of $\langle\mu\rangle$ from 38 to 72. The efficiency decreases with increasing pile-up. Figure 3.6 shows the average number of reconstructed primary vertices compared to truth vertices as a function of $\langle\mu\rangle$, where matched vertex is vertex with tracks mainly from one generated interaction, merged vertex is vertex with tracks from two or more generated interactions. The position resolution of the reconstructed vertex measured from the same low pile-up data is shown in Figure 3.7. The resolution in the transverse direction is around 20 μm for events with more than 25 tracks. For events with a very small number of tracks, the resolution could exceed 100 μm . The resolution in the longitudinal direction is around 30 μm for events with more than 25 tracks.

3.3 Electron reconstruction

The electrons are reconstructed from energy deposits in the EM calorimeter with associated charged particle tracks in the ID [94]. A schematic illustration of the sub-detectors crossed by an electron is

3 Object reconstruction

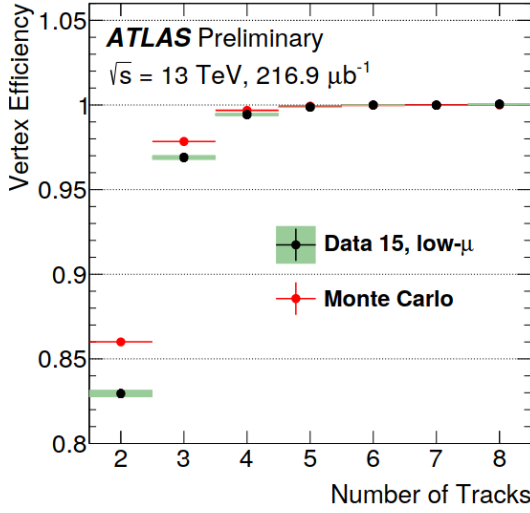


Figure 3.5: Vertex reconstruction efficiency measured from low- μ data collected in Run 2 and minimum-bias MC simulation, as a function of number of tracks in the events [92].

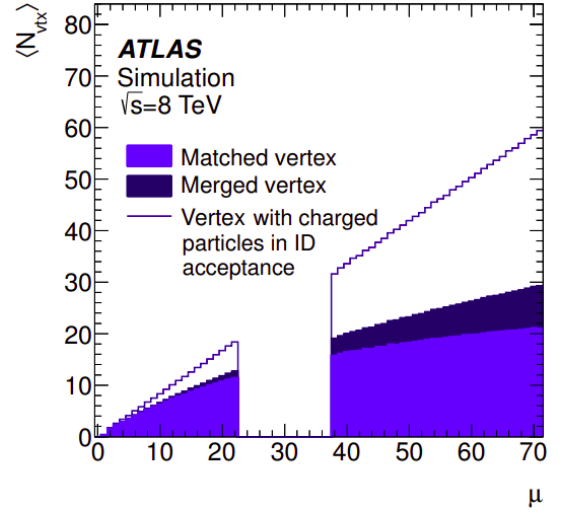


Figure 3.6: Average number of reconstructed primary vertices compared to the total number of truth vertices as a function of $\langle\mu\rangle$, measured from simulated samples in Run 1 [93].

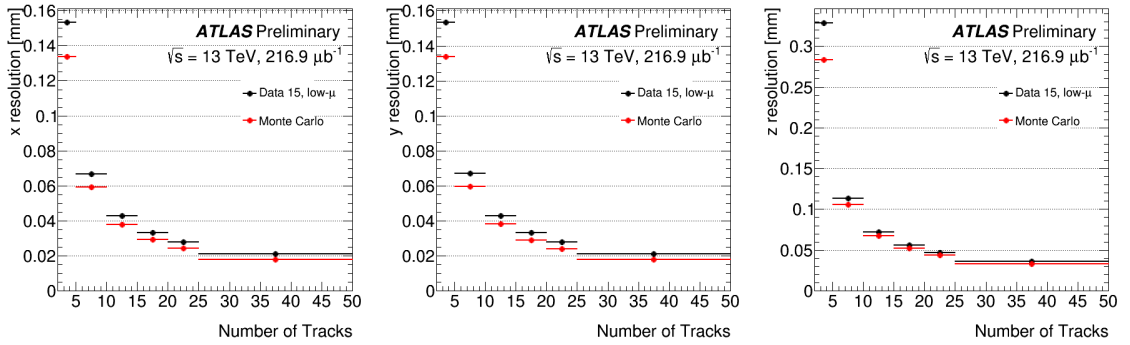


Figure 3.7: Position resolution of the reconstructed primary vertex in the x -axis direction (left), y -axis direction (middle), z -axis direction (right) [92].

shown in Figure 3.8. The complete reconstruction process includes the electron reconstruction, the electron identification and the performance studies.

The energy deposits in the calorimeters are reconstructed as topological clusters. For the reconstruction of electrons, seed clusters formed from EM calorimeter towers are identified. The η - ϕ space of the EM calorimeter is divided into a grid of $N_\eta \times N_\phi = 200 \times 256$ towers of size $\Delta\eta \times \Delta\phi = 0.025 \times 0.025$, corresponding to the granularity of the EM calorimeter middle layer. The energy of a tower is obtained from the sum of all the cell energies in the shower-depth layers. For a cell that spans several towers, the energy is assigned evenly among the towers without taking into account any geometrical weighting. The seed clusters are searched for by a sliding-window algorithm [96] with a window size of 3×5 towers in the η - ϕ space. The total transverse energy of the seed cluster has to be larger than 2.5 GeV.

3.3 Electron reconstruction

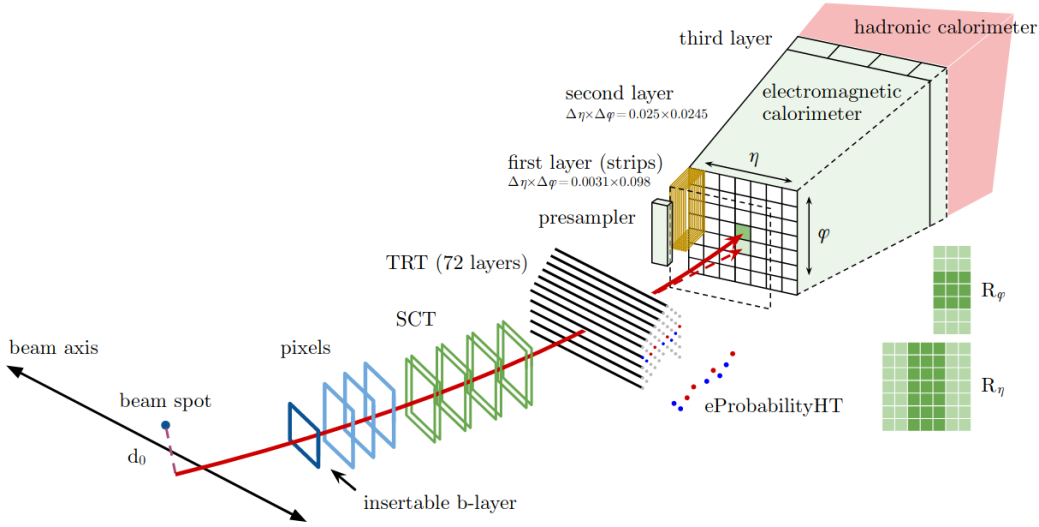


Figure 3.8: Schematic illustration of the sub-detectors crossed by an electron. The solid red curve represents the trajectory of the electron, the dashed red curve represents the photon produced by the interaction between the electron and the detector materials [95].

Electron track reconstruction follows the methods described in Section 3.1. In addition to the standard track reconstruction process, an electron-specific pattern recognition and fitting algorithm is developed to recover the energy losses from bremsstrahlung. The standard track reconstruction algorithm uses pion hypothesis for energy loss of the charged particles, so a track seed with a transverse momentum larger than 1 GeV cannot be extended to a full track with less than seven silicon hits. For such cases, a new pattern recognition using electron hypothesis is re-tried. This modification allows up to 30% energy loss in the material. It is integrated into the standard track reconstruction above 1 GeV, which improves the performance and has minimal interference with the main track reconstruction. The fitting of the tracks with electron hypothesis takes into account the increase in χ^2 due to bremsstrahlung. A track candidate is fitted with the standard fitting (pion hypothesis) first. If it fails due to large χ^2 , it is then re-fitted with the electron hypothesis.

The tracks are loosely matched to the seed clusters based on the distance in η and ϕ . The matching conditions allow disagreement between momentum measured in the ID and energy reconstructed in the calorimeter to account for bremsstrahlung energy loss. After the matching, the track parameters of these selected tracks are precisely estimated with an optimized fitter, the Gaussian Sum Filter (GSF) algorithm [97], which is a non-linear generalization of the Kalman filter that accounts for non-linear bremsstrahlung effects. The final electron candidate is reconstructed by matching the selected tracks to the seed cluster with tighter selection than the previous step. An electron has to have at least one track matched to the seed cluster. If more than one tracks match to a seed cluster, the best matched track is selected as the primary track. The selection is based on the track-cluster distance and the number of pixel hits, especially hits in the IBL in the track. The electron candidates without any pixel hits are removed and considered to be photons. The kinematics of the electrons is computed using the calibrated energy of the cluster and the primary track. The reconstruction efficiency of electrons is shown in Figure 3.9. The electron reconstruction efficiency becomes closer to the track reconstruction

3 Object reconstruction

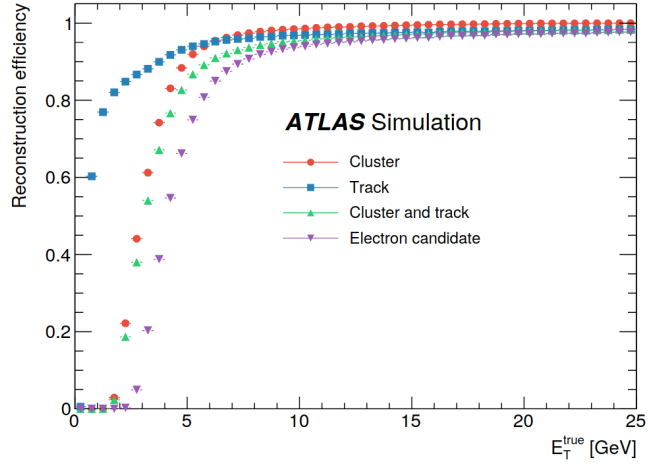


Figure 3.9: Reconstruction efficiency of the seed clusters, electron tracks, matched clusters and tracks and the electrons as a function of truth electron E_T [98].

efficiency as E_T increases.

3.3.1 Electron identification

Not all reconstructed electron candidates are prompt electrons originate from hard scattering physics processes. Fake electrons refer to objects such as hadronic jets, electrons from photon conversions that are reconstructed as electron candidates. It is necessary to apply electron identification algorithms to reject these backgrounds while keeping the prompt electrons. The electron identification algorithm for Run 2 is a likelihood-based method that uses multivariate analysis (MVA) to make selection decisions. The distribution of the likelihood discriminant peaks at 1 and 0 for real and fake electrons respectively.

Three levels of identification criteria are typically provided: Loose, Medium and Tight. The selections are defined with the same discriminant, but the cuts on this discriminant are different in order to achieve different efficiencies. The electrons selected by Medium criteria are a subset of those selected by Loose, and Tight electrons are a subset of those passing Medium criteria.

The efficiencies of the electron identification algorithms are measured in data and simulation. Processes with very clean signature is used: $Z \rightarrow ee$ for electrons with $E_T > 15$ GeV and $J/\psi \rightarrow ee$ for electrons with $E_T < 15$ GeV. The efficiencies are calculated as the ratio of the number of electrons passing the identification criteria and the number of electrons with a good quality associated track. The tag-and-probe method [94] is used to select unbiased electron candidates in the samples. One of the electron (tag) must satisfy a strict requirement on the discriminant, and the other electron (probe) is selected with loose requirement. The tag is used to select events, the probe is used to measure the efficiency. Several other event selection cuts are applied, including E_T , η thresholds and invariant mass window. The measured electron identification efficiency is shown in Figure 3.10. The efficiency varies from 58% to 88% for Tight operating point, and from 86% to 95% for Loose operating point.

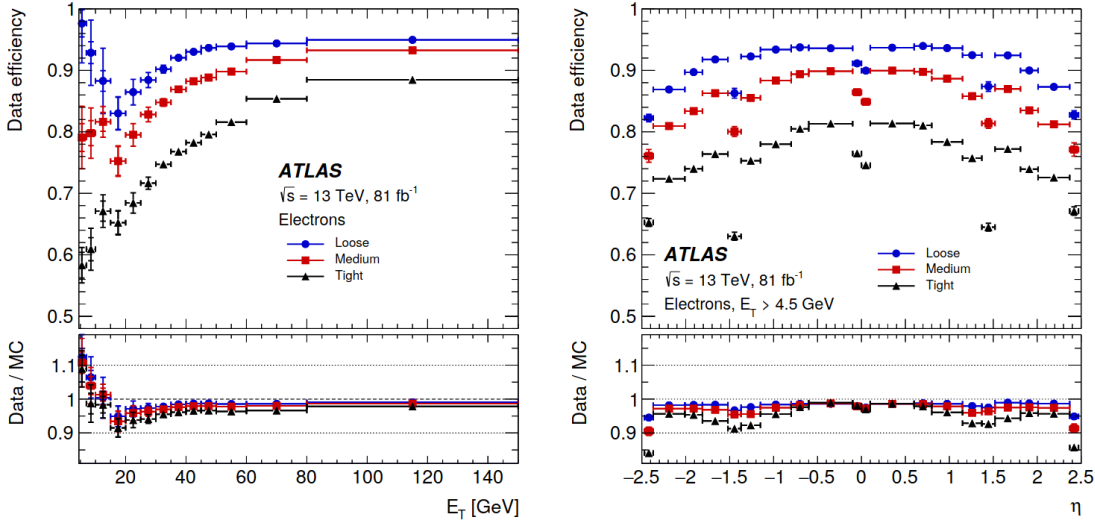


Figure 3.10: Electron identification efficiency measured from $Z \rightarrow ee$ events in data collected in Run 2, as a function of E_T (left) and η (right) for Loose, Medium, Tight operating points [98].

3.3.2 Electron isolation

Except the electron identification criteria, some isolation requirements can further discriminate between signal and background. The isolation discriminating variables are able to reject objects such as electrons from photon conversions, electrons from heavy hadron decay, and light hadrons mis-identified as electrons. Two variables that are typically designed for this purpose are:

- calorimeter isolation variable E_T^{coneX} . It is defined by the sum of E_T of all the positive energy topo-clusters within $\Delta R = X$ around the seed cluster at electromagnetic scale, the energy of the electron is subtracted.
- track isolation variable p_T^{varconeX} . It is defined by the sum of p_T of all the quality tracks within $\Delta R = \min(X, 10 \text{ GeV}/E_T)$ around the candidate track of the electron. Tracks matched to the electron or converted photon are excluded.

Typical value for X is 0.2. Working points of the isolation selection are defined in two ways: a fixed cut on variables or a fixed efficiency. The fixed efficiency working point is named as the "Gradient". It is designed to provide an efficiency of 90% at $E_T = 25$ GeV and an efficiency of 99% at $E_T = 60$ GeV. The other three working points are fixed cuts on the variables defined as [98]:

- HighPtCaloOnly, $E_T^{\text{cone0.2}} < \max(0.015 p_T, 3.5 \text{ GeV})$.
- Loose, $E_T^{\text{coneX}}/E_T < 0.2$, $p_T^{\text{varconeX}}/p_T < 0.15$.
- Tight, $E_T^{\text{coneX}}/E_T < 0.06$, $p_T^{\text{varconeX}}/p_T < 0.06$.

Efficiencies of these working points measured from $Z \rightarrow ee$ events are shown in Figure 3.11.

3 Object reconstruction

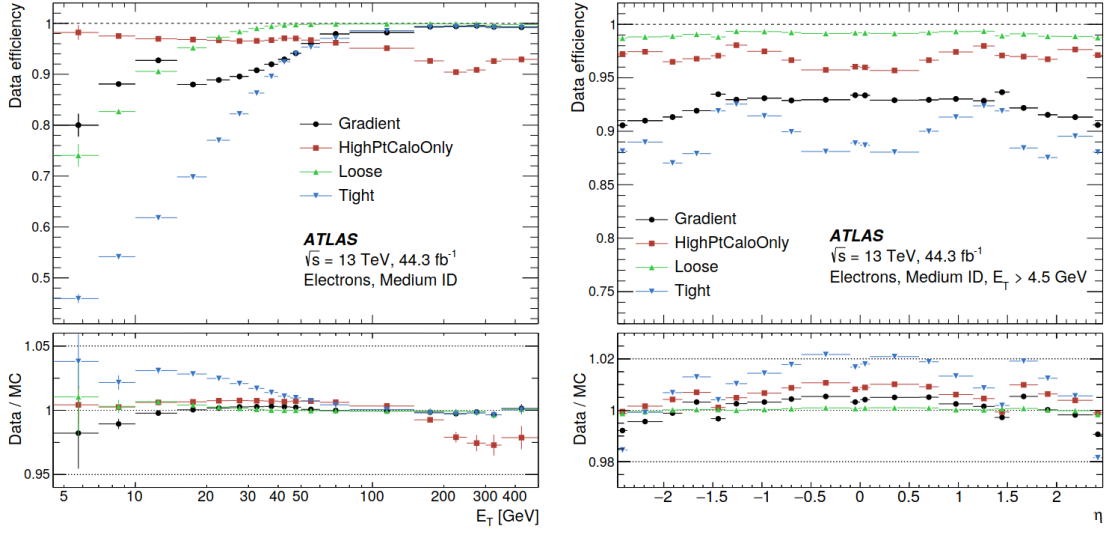


Figure 3.11: Electron isolation efficiencies measured from $Z \rightarrow ee$ events in data collected in Run 2 as a function of E_T (left) and η (right) for different working points [98].

3.4 Muon reconstruction

Muons can pass through all the sub-detectors. The muon trajectories are reconstructed from hits in the ID and the MS independently. Muons also interact with the calorimeter material. The deposited energy is limited but can be used in the muon identification. Muon track reconstruction in the ID follows the same methods like any other particles as described in Section 3.1. Muon reconstruction in the MS [99, 100] starts from identification of short straight-line local track segments. A feature extraction technique called as the Hough transform [101] is used to identify segments aligned with the hits in the bending plane of the MS. Segments in different detector modules are combined into preliminary track candidates. Loose constraints on the originating point and the bending of the track are applied. Then, a global χ^2 fitting is performed to reconstruct muon tracks from segments in different layers. A track needs to have at least two segments, except in the barrel end-cap transition region, where a single high quality segment can be used to build a track.

The muon candidate can be identified with information from different combinations of sub-detectors. There are five types of reconstructed muon:

- Combined (CB): combination of tracks reconstructed independently in the ID and the MS. A global fit re-calculates the parameters of the combined track.
- Inside-out combined (IO): track reconstructed in the ID with at least three loosely aligned MS hits. A combined fit reconstructs the track with the ID hits and the MS hits.
- Segment-tagged (ST): track reconstructed in the ID with at least one associated segment in the MDT or CSC chambers.
- Calorimeter-tagged (CT): track reconstructed in the ID with a matched energy deposit in the calorimeter, corresponding to a minimum ionizing particle.

- MS extrapolated (ME) muon: track reconstructed only in the MS, the track can be extrapolated to the IP with a loose requirement.

The overlap between different muon types is resolved by always giving priority to CB muons, then to IO muons, then to ST muons, and finally to CT muons. The ME muons are mainly used in the region not covered by the ID, $2.5 < |\eta| < 2.7$.

3.4.1 Muon identification

Similar to electron reconstruction, not all reconstructed muons are prompt muons. Fake muons can originate from pion and kaon semileptonic decays, punch-through jets not fully contained in the calorimeter, or from cosmic rays. Muon identification algorithms select high quality muon candidates using number of hits in the ID and the MS, track fit properties and variables that test the compatibility of the measurements in the two sub-detectors. For each type of muons, several sets of selection requirement with different working points (WP) are provided. Depending on the efficiency of muon candidates passing the cuts, typical working points are defined as Loose, Medium and Tight. Muons passing Medium selection constitute a subset of those passing Loose selection. Muons passing Tight selection are also a subset of muons passing Medium selection. The Medium WP is used by many analyses. The Loose WP is specially designed for Higgs to four-muon decays. The Tight selection provides the highest background rejection rate, mainly used in analyses that are limited by background from non-prompt muons. Two additional working points called as High- p_T WP and Low- p_T WP are designed for analyses that explore extreme phase space with muon p_T above 100 GeV or the lowest p_T muons.

The Medium identification criteria is the most widely used WP. It is defined only for CB and IO muons. The quantities used to define the WP are

- precision hit (hole) station: number of MS stations in which muon has at least three hits (holes) in the MDT and CSC chambers, where "hole" means a sensitive detector element along the track that does not provide a matched cluster;
- q/p compatibility: the absolute difference between the ratio of the charge and momentum of the muons measured in the ID and MS, divided by the sum in quadrature of the corresponding uncertainties,

$$q/p \text{ compatibility} = \frac{|q/p_{ID} - q/p_{MS}|}{\sqrt{\sigma^2(q/p_{ID}) + \sigma^2(q/p_{MS})}}; \quad (3.2)$$

- ρ' : the absolute difference between the ID and the MS p_T measurement divided by the p_T of the combined muon track,

$$\rho' = \frac{|p_{T,ID} - p_{T,MS}|}{p_{T,combined}}; \quad (3.3)$$

- normalized χ^2 of the combined track fit.

3 Object reconstruction

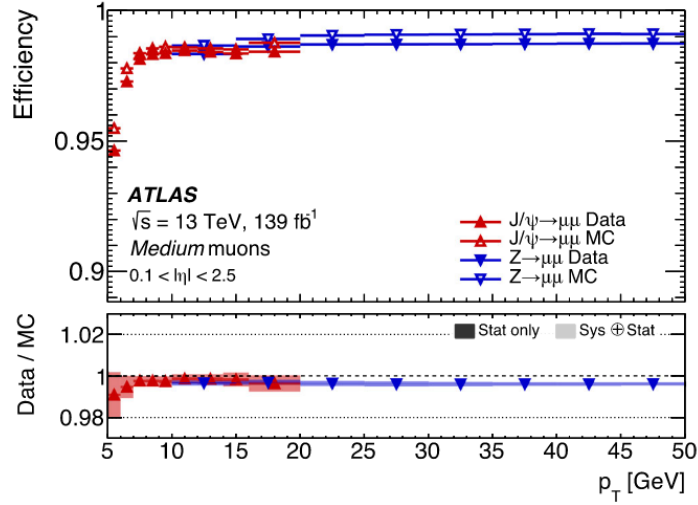


Figure 3.12: Muon reconstruction and identification efficiency measured in $Z \rightarrow \mu\mu$ and $J/\psi \rightarrow \mu\mu$ events in Run 2, for Medium selection criteria [100].

The muon candidates are required to have hits in at least two precision stations, except in $|\eta| < 0.1$, where at least one hit in precision station and no more than one hole in the precision station is enough. The q/p compatibility is required to be less than seven. More than 98% muons passing the medium identification selection are CB muons. Based on the Medium selection cuts, the Tight identification continues to require normalized χ^2 being less than eight. Further requirements on q/p compatibility and ρ' are placed in some regions. The Loose identification is defined to accept all types of muons. CT and ST muons are accepted in $|\eta| < 0.1$. IO muons with low p_T ($p_T < 7 \text{ TeV}$) and only one precision station are accepted in $|\eta| < 1.3$.

3.4.2 Muon reconstruction and identification efficiency

The efficiencies of muon reconstruction and identification are measured in $Z \rightarrow \mu\mu$ and $J/\psi \rightarrow \mu\mu$ events. The tag-and-probe method is used to select samples containing muon pairs. Additional cuts applied in event selection include p_T , η thresholds and invariant mass window. The selected signal events have a purity of 99.9%. $Z \rightarrow \mu\mu$ samples are used to measure efficiency for muon with p_T greater than 15 GeV. $J/\psi \rightarrow \mu\mu$ events are used to measure efficiencies in $3 < p_T < 20 \text{ GeV}$ with small uncertainties. The measured muon reconstruction and identification efficiencies for Medium criteria is shown in Figure 3.12. The measurement from $Z \rightarrow \mu\mu$ and $J/\psi \rightarrow \mu\mu$ in overlap region between 10 GeV and 20 GeV agree well with each other. The efficiency exceeds 98% for muons with $p_T > 10 \text{ GeV}$.

3.5 Jet reconstruction

Partons hadronize to produce a spray of particles that are called jets. The reconstruction of jets is essential to many physics analyses including $H \rightarrow b\bar{b}$, as the b -hadrons in the final state are embedded in b -jets. The complete reconstruction of jets has several steps. The actual reconstruction form jets from calorimeter clusters, then the reconstructed jets are processed by several algorithms to calibrate the kinematic variables. This section describes the complete reconstruction process. One step of the reconstruction, the Monte-Carlo based calibration, which is part of my own original contribution, is described in detail in Chapter 4.

3.5.1 Reconstruction of clusters

The first step of the jet reconstruction is the building of calorimeter topological clusters (topo-clusters) [102]. Topo-clusters are groups of calorimeter cells spanning in three-dimensional space of η , ϕ and r . The observable that controls the cluster formation is the cell signal significance, defined as

$$S_{\text{cell}}^{\text{EM}} = \frac{E_{\text{cell}}^{\text{EM}}}{\sigma_{\text{noise,cell}}^{\text{EM}}}, \quad (3.4)$$

where $E_{\text{cell}}^{\text{EM}}$ is the energy of the cell, $\sigma_{\text{noise,cell}}^{\text{EM}}$ is the average noise in a cell, both measured on the electromagnetic (EM) scale. This energy scale measures the energy deposits of electrons and photons correctly, but does not accounts for corrections of the energy of hadrons due to the non-compensating character of the ATLAS calorimeter. The formation of a topo-cluster follows the 4-2-0 approach. The algorithm starts from seed cells with $|S_{\text{cell}}^{\text{EM}}| > 4$. Then all the cells neighboring a seed cell with $|S_{\text{cell}}^{\text{EM}}| > 2$ are added. If a neighboring cell passes the seed cell requirement, then the two seed cells are merged. Finally, all the neighboring cells with $|S_{\text{cell}}^{\text{EM}}| > 0$ are added. "Neighboring" is generally defined as a calorimeter cell and the formed cells in the topo-cluster being directly adjacent in a given sampling layer, or having partial overlap in the η - ϕ plane if in adjacent layers. The cell collections can span modules within the same calorimeter as well as calorimeter sub-detector transition regions.

Clusters built following this approach can be too large to provide a good measurement of energy, so clusters with more than one signal peaks are split. The signal peak is defined as a cell with $E_{\text{cell}}^{\text{EM}} > 500$ MeV, and this cell has at least four neighbors with no larger cell energy. In the splitting, each cell could be split no more than one time. The energy of a shared cell is split between the two clusters based on geometrical signal weights, calculated as:

$$\omega_1 = \frac{E_{\text{clus},1}}{E_{\text{clus},1} + rE_{\text{clus},2}}, \quad \omega_2 = 1 - \omega_1, \quad (3.5)$$

where $r = \exp(d_1 - d_2)$, d_1 and d_2 are the distances from the cell to the gravity center of the two clusters, measured in units of a typical electromagnetic shower size (approximately 5 cm) scale in the ATLAS calorimeters. Every cell has a geometrical signal weight defined for convenience. For a cell that is included in two clusters, the geometrical signal weights are calculated as in Eq. 3.5. For a cell that is not split, its weight ω is defined as 1.

3 Object reconstruction

Energy of the topo-cluster is calculated as weighted sum of all the cell energies,

$$E_{\text{cluster}} = \sum_{i=1}^{N_{\text{cell}}} \omega_i \cdot E_{\text{cell}}. \quad (3.6)$$

The cell energy can be calculated at EM scale or local cell weight (LCW) scale. The LCW scale accounts for the energy loss of hadrons. It is mainly used in reconstruction of large- R jets. The variable ω_i is the geometrical signal weights introduced by cluster splitting. Cells with negative energy are results of fluctuations introduced by pile-up and electronics noise. The absolute value of $\zeta_{\text{cell}}^{\text{EM}}$ is used in the clustering process, so cells with negative energy are also collected. Topo-clusters with negative energy seed are likely to have negative energy as well. These clusters are typically generated by out-of-time pile-up. The negative energy topo-clusters are not included in the jet reconstruction, but they can be used as an average global cancellation of contributions from pile-up. The cluster direction is calculated as weighted average of cell positions,

$$\eta_{\text{clus}} = \frac{\sum_{i=1}^{N_{\text{cell}}} \omega_i \cdot E_{\text{cell}} \cdot \eta_{\text{cell},i}}{\sum_{i=1}^{N_{\text{cell}}} \omega_i \cdot E_{\text{cell}}}, \quad \phi_{\text{clus}} = \frac{\sum_{i=1}^{N_{\text{cell}}} \omega_i \cdot E_{\text{cell}} \cdot \phi_{\text{cell},i}}{\sum_{i=1}^{N_{\text{cell}}} \omega_i \cdot E_{\text{cell}}}. \quad (3.7)$$

The momentum of a topo-cluster is calculated using its reconstructed energy and position,

$$\vec{p}_{\text{clus}} = E_{\text{clus}} \cdot (\sin \theta_{\text{clus}} \cos \phi_{\text{clus}}, \sin \theta_{\text{clus}} \sin \phi_{\text{clus}}, \cos \theta_{\text{clus}}). \quad (3.8)$$

The topo-clusters are interpreted as massless pseudo-particles in object reconstruction. The jets also do not have a reconstructed mass. However, the mass can be defined with a valid physics hypothesis.

Track-Calo Cluster

The default topo-clusters are reconstructed using only information from the calorimeter. It is possible to improve the jet resolution by combining the tracker and the calorimeter information. Track-Calo Cluster (TCC) [103] is a combination of track spatial coordinates and topo-cluster energy components. The idea of TCC is to maximize the performance of the reconstructed objects by exploiting the strengths of each detector. The ATLAS ID has excellent position resolution. However, the ID can only measure charged particles and the momentum is calculated from the curvature. In contrast, the calorimeter has an excellent energy resolution, while the angular resolution is limited.

The TCC building algorithm attempts to match tracks and topo-clusters if they meet the following requirements:

$$\sigma_{\text{trk}} < \sigma_{\text{clus}}, \quad \Delta R < \sqrt{\sigma_{\text{clus}}^2 + \sigma_{\text{trk}}^2}, \quad (3.9)$$

where σ_{clus} is the width of the topo-cluster, σ_{trk} is the uncertainty on the track extrapolated to the topo-cluster, ΔR is the angular distance between the track and the topo-cluster. The successfully matched track and topo-cluster pairs are then used to build the TCC 4-vector. Figure 3.13 shows the different types of TCC objects. The simplest type is a single match between a track and a topo-cluster, as the case of the first TCC object. The 4-vector is built from the track coordinates and the cluster energy.

$$\text{TCC}_1 : (p_T^{c1}, \eta^{t1}, \phi^{t1}, m = 0). \quad (3.10)$$

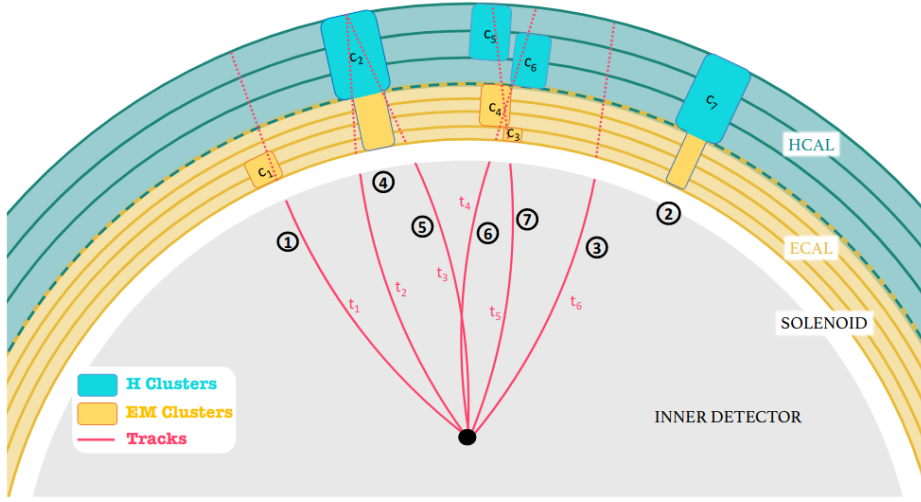


Figure 3.13: Schematic demonstration of the seven types of TCC objects. The red curves labeled as "t" represent tracks. The orange blocks represent EM calorimeter clusters. The blue blocks represent hadronic calorimeter clusters [103].

Topo-clusters without a matched track are used to build TCC directly, as the case of type 2.

$$\text{TCC}_2 : (p_T^{c_7}, \eta^{c_7}, \phi^{c_7}, m = 0). \quad (3.11)$$

Tracks without a matched topo-cluster are also used to build TCC directly, such as the type 3.

$$\text{TCC}_3 : (p_T^{t_6}, \eta^{t_6}, \phi^{t_6}, m = 0). \quad (3.12)$$

When multiple tracks matched to a single topo-cluster like the forth and the fifth cases, each track is used to build a TCC with part of the topo-cluster. The topo-cluster energy is shared between the two tracks based on their momentum fraction.

$$\text{TCC}_4 : (p_T^{c_2} \frac{p_T^{t_2}}{p_T[\mathbf{p}^{t_2} + \mathbf{p}^{t_3}]}, \eta^{t_2}, \phi^{t_2}, m = 0), \quad (3.13)$$

$$\text{TCC}_5 : (p_T^{c_2} \frac{p_T^{t_3}}{p_T[\mathbf{p}^{t_2} + \mathbf{p}^{t_3}]}, \eta^{t_3}, \phi^{t_3}, m = 0), \quad (3.14)$$

where $p_T[\mathbf{p}^{t_2} + \mathbf{p}^{t_3}]$ is the transverse momentum of the 4-vector sum of track t_2 and t_3 .

More complicated cases are multiple tracks matched to multiple topo-clusters, like the examples of type 6 and type 7. Each track is still used as a seed to build a TCC. The topo-cluster energy is shared between tracks based on their momentum fraction, and the track momentum used to calculate the fraction needs to be weighted with the cluster energy fraction because multiple clusters matched to a track. The weight is calculated as:

$$w_\tau^{c_i} = \frac{p_T^{c_i}}{p_T [\sum_{c_k} p^{c_k}]}, \quad (3.15)$$

3 Object reconstruction

where the sum runs over all the clusters (c_k) matched to a specific track τ , c_i represents the topo-cluster that is being shared. The p_T fraction of track τ used to calculate the energy contribution from topo-cluster c_i to a TCC object becomes:

$$f^{\tau, c_i} = \frac{p_T^\tau \cdot w_{\tau}^{c_i}}{p_T \left[\sum_{t_j} p_{t_j}^\tau \cdot w_{t_j}^{c_i} \right]}, \quad (3.16)$$

where the sum runs over all the tracks (t_j) matched to topo-cluster c_i .

Finally, the 4-vector of a TCC object built from track τ is:

$$\text{TCC}_\tau : \left(\sum_{c_i} p_T^{c_i} \cdot f^{\tau, c_i}, \eta^\tau, \phi^\tau, m = 0 \right), \quad (3.17)$$

where the sum runs over all the topo-clusters matched to the track τ .

TCC objects have better resolution compared to topo-clusters, especially in environments with several close-by or even merged jets. However, jets reconstructed from TCC objects cannot perform as well as jets reconstructed from topo-clusters in low p_T regions. To improve the performance over a wide p_T range, TCC is integrated into a new object called Unified Flow Objects introduced in the following sections.

Particle Flow Objects

Particle Flow (PFlow) [104] is another way to combine tracking and calorimeter information. It subtracts calorimeter energy deposits that can match an extrapolated track. The reconstructed PFlow object is a combination of tracks and remaining topo-clusters. The PFlow algorithm first attempts to match tracks to one topo-cluster. The topo-clusters used in the reconstruction have to capture at least 90% of the total energy of a particle. The tracks are selected with strict quality criteria. At least nine hits in the silicon detectors and no holes in the Pixel detector are required. The topo-clusters are ranked by a modified distance metric:

$$\Delta R' = \sqrt{\left(\frac{\Delta\phi}{\sigma_\phi} \right)^2 + \left(\frac{\Delta\eta}{\sigma_\eta} \right)^2}, \quad (3.18)$$

where $\Delta\phi$ and $\Delta\eta$ are the angular distance between the barycenter of the topo-cluster and the track, σ_ϕ and σ_η are the angular topo-cluster width. The topo-clusters are preliminary matched to the tracks if they satisfy $E^{\text{clus}}/p^{\text{trk}} > 0.1$. Then the selected topo-clusters with the smallest $\Delta R'$ is taken to be the matched one. Figure 3.14 illustrates four types of PFlow. These cases use π^+ and π^0 as examples. The particles deposit their energy in one or two topo-clusters. The four examples demonstrate the majority of the PFlow objects.

For the case of a single charged particle matched to a topo-cluster as in Figure 3.14 (a), the topo-cluster is subtracted and the final PFlow object becomes the track. Particles do not always deposit their energy in a single topo-cluster, two topo-cluster case is very common, as shown in Figure 3.14 (b). The algorithm checks if the shower is split across multiple clusters, and then takes additional clusters

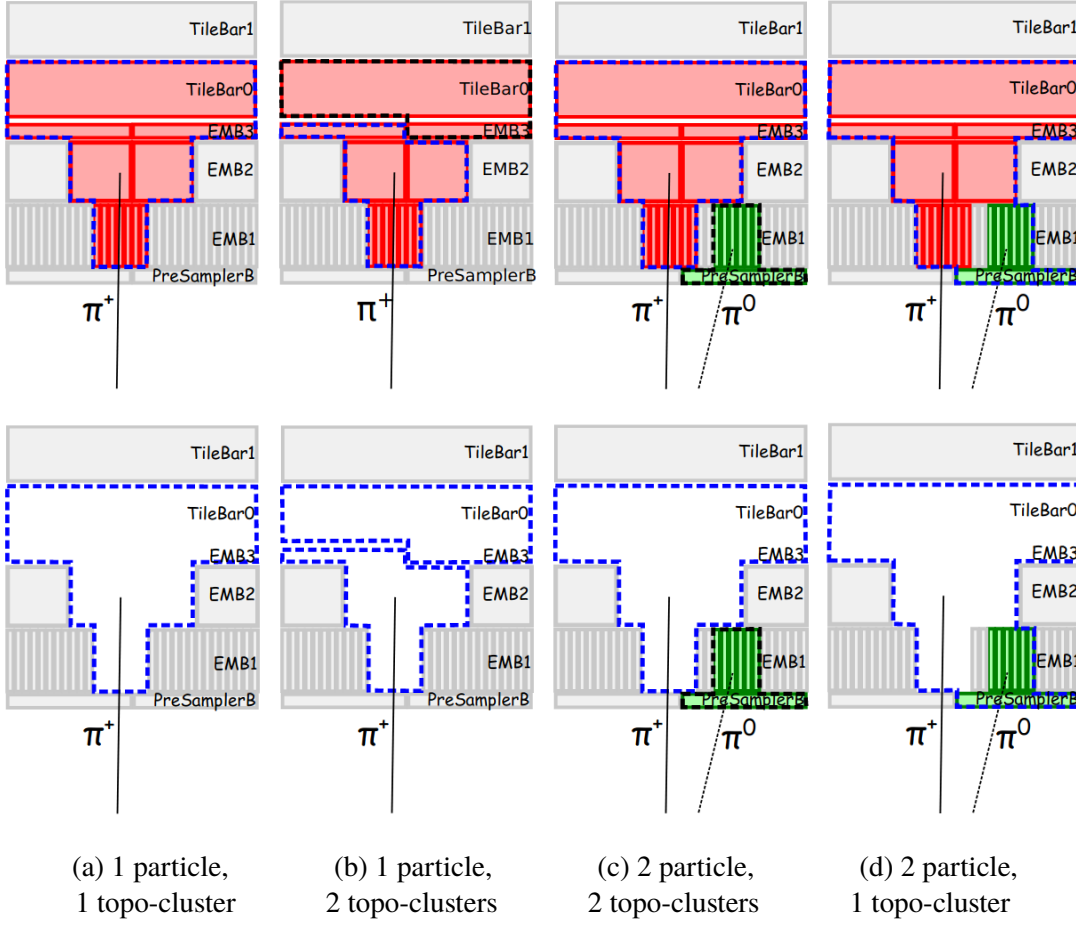


Figure 3.14: Illustration of PFlow matching (upper) and subtraction (lower) process [104]. (a) One track matched to one topo-cluster. (b) One track matched to two topo-clusters. (c) A charged particle and a neutral particle deposited as two topo-clusters. (d) A charged particle and a neutral particle deposited as one topo-cluster.

into consideration. The energy deposits in multiple topo-clusters matched to a charged particles are subtracted and replaced with the track, similar to the single topo-cluster case. If a topo-cluster does not match to any tracks as the case in Figure 3.14 (c), then it is assumed to be originated from a neutral particle. The topo-cluster remains unchanged and becomes the final PFlow object. For the case of a single topo-cluster deposited from multiple particles, as in Figure 3.14 (d), the energy deposits matched to charged particles are subtracted. The energy deposits of charged particles are estimated by

$$\langle E_{\text{dep}} \rangle = p^{\text{trk}} \langle E_{\text{ref}}^{\text{clus}} / p_{\text{ref}}^{\text{clus}} \rangle, \quad (3.19)$$

where $E_{\text{ref}}^{\text{clus}} / p_{\text{ref}}^{\text{clus}}$ is the ratio of energy deposits measured in calorimeter and momentum measured in the ID, with subscript ‘ref’ to indicate it is determined from single pion samples. The final PFlow object is a modified topo-cluster with the estimated energy replaced using track. PFlow is able to improve energy and angular resolution of jets compared to topo-clusters. It has been integrated into jet reconstruction for Run 2. The algorithm is robust against pile-up, and has good performance for jets with p_T up to 100 GeV.

3 Object reconstruction

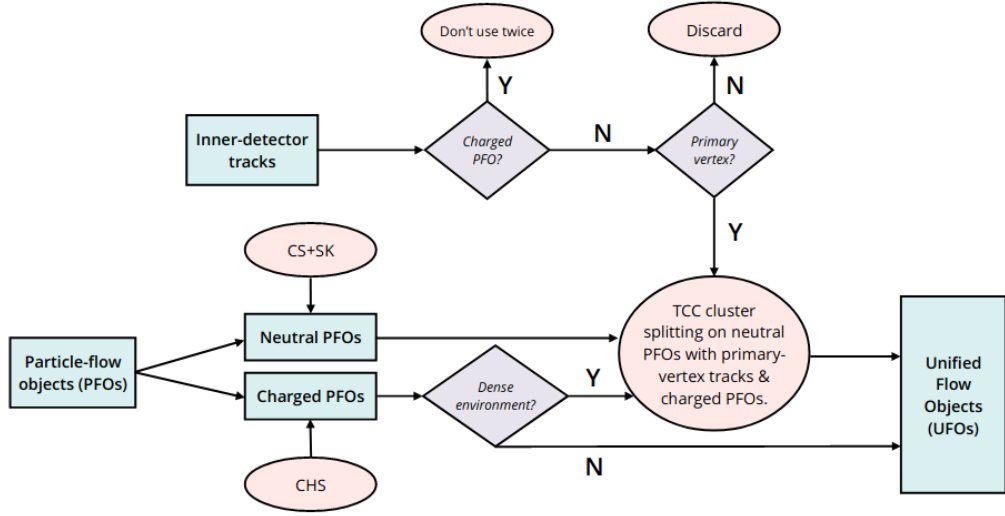


Figure 3.15: An illustration of the UFO reconstruction algorithm [105].

Unified Flow Objects

Unified Flow Object (UFO) is a union of TCC objects and PFlow objects [105]. The motivation of UFO is to obtain a better and more stable performance in a wide p_T range. TCC is able to improve the jet reconstruction performance significantly at high p_T , but their performance is worse than the baseline topo-clusters at low p_T . Jets reconstructed from PFlow objects are improved overall, but do not achieve the same level as TCC at high p_T .

The procedure of UFO reconstruction algorithm is illustrated in Figure 3.15. The algorithm starts from reconstructed PFlow objects (PFO). The PFOs are divided into three categories: neutral PFOs, which are the same as topo-clusters; charged PFOs with energy subtracted, which are PFOs reconstructed correctly; charged PFOs without energy subtraction, which are PFOs mis-matched in the so-called "dense" environment that has multiple close-by objects from high p_T particle decays. Pile-up mitigation algorithms such as Constituent-Subtraction (CS) and Soft-Killer (SK) [106] may be applied to the PFOs. For the charged PFOs, they come out directly as UFOs if they are not in a track dense environment. Otherwise, the charged PFO is processed by the TCC algorithm together with primary vertex tracks that have not been used in PFO reconstruction. The reconstructed TCC object becomes an UFO. The neutral PFOs are also processed by the TCC algorithm, and the reconstructed TCC objects are UFOs.

UFOs are very stable with respect to pile-up. It offers enhanced resolution and better jet reconstruction, calibration and tagging performance. Large- R jets reconstructed from UFOs have been recommended as the default in the ATLAS experiment for future analyses in Run 3.

3.5.2 Jet clustering

Jets are reconstructed from topo-clusters, PFOs, UFOs, tracks and truth particles using anti- k_t algorithm [107]. To apply the algorithm, distance quantities of d_{ij} and d_{iB} are introduced. The variables are defined as:

$$d_{ij} = \min\left(\frac{1}{p_{Ti}^2}, \frac{1}{p_{Tj}^2}\right) \frac{\Delta R_{ij}^2}{R^2}, \quad (3.20)$$

$$d_{iB} = \frac{1}{p_{Ti}^2}, \quad (3.21)$$

where p_{Ti} and p_{Tj} are the transverse momentum of topo-cluster i and topo-cluster j , ΔR_{ij} is the angular distance between i and j , R is the radius parameter that decides the radius of the reconstructed jets. The clustering proceeds by identifying the smallest among all the d_{ij} and d_{iB} for the topo-cluster i . If the smallest is d_{ij} , the topo-cluster i and the topo-cluster j are merged. If it is d_{iB} , then topo-cluster i is identified as a jet and removed from the list of constituents. The process is repeated with the remaining topo-clusters until no constituents are left.

The functionality of the anti- k_t algorithm can be understood by considering an event that contains a few well-separated hard scattered particles and many soft constituents. The quantity d_{ij} between a hard scattered particle and a soft constituent is determined by the particle p_T and the angular distance ΔR_{ij} . Soft constituents with similar ΔR have much larger d_{ij} because their p_T is much smaller than the hard scattered particle. As a result, soft constituents are much more likely to be clustered with nearby hard scattered particles. If a hard particle has no neighboring hard particles within an angular distance of $2R$, it will simply accumulate all the soft constituents within a circle of radius R , resulting a perfect conical jet.

Jets can be reconstructed with different radius to suit different needs. One of the most widely used radius is 0.4. Jets with $R = 0.4$ (small- R) can be used to reconstruct most of the hadron decay products. For analyses that target high transverse momentum regions like the boosted $VH(b\bar{b})$ analysis, jets reconstructed with $R = 1.0$ (large- R) are needed. In the boosted $VH(b\bar{b})$ analysis, the small- R jets are reconstructed from topo-clusters with their energy calibrated at EM scale, the large- R jets are reconstructed from topo-clusters with LCW scale energy. Several new inputs have been studied for both small- R jets and large- R jets to develop a precision recommendation for Run 3. The new small- R jets are reconstructed from PFlow objects. The large- R jets are reconstructed from UFOs. The precision recommendation for small- R jets is used in the combined $VH(b\bar{b})$ legacy analysis.

Another jet collection used in $VH(b\bar{b})$ analysis is track jet formed from charged particle tracks, with a variable radius (VR). Instead of clusters, the anti- k_t algorithm is applied to tracks. The radius parameter is proportional to the inverse of the jet p_T : $R = \rho/p_T$, where ρ is set to 30 GeV. The minimum and maximum of the radius parameter are set to 0.002 and 0.4, respectively. The other two clustering algorithms are the k_t algorithm [108] and the Cambridge-Aachen (C-A) algorithm [109, 110]. These algorithms are not used in jet reconstruction directly, but they are useful in the calibration processes to remove the soft constituents. The k_t algorithm has the same steps as the anti- k_t algorithm, but it uses p_T^2 instead of $1/p_T^2$ in the calculation of the distance variables.

$$d_{ij} = \min\left(p_{Ti}^2, p_{Tj}^2\right) \frac{\Delta R_{ij}^2}{R^2}, \quad (3.22)$$

3 Object reconstruction

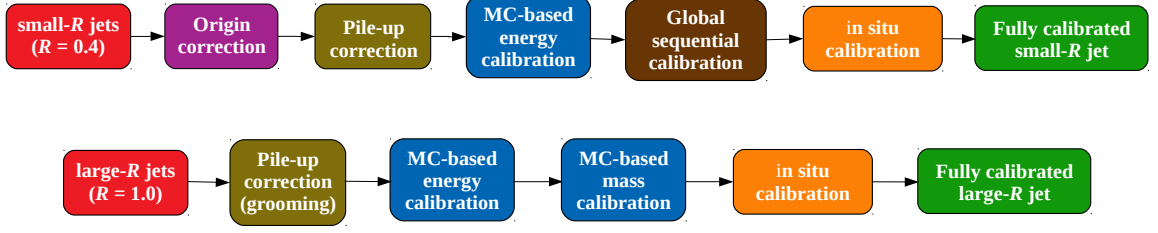


Figure 3.16: An overview of the calibration chain of the small- R jets (upper) and the large- R jets (lower).

$$d_{iB} = p_{Ti}^2. \quad (3.23)$$

The k_t clustering algorithm identifies the smallest distance and proceeds in the same way as the anti- k_t algorithm. The C-A algorithm simply merges the pair of clusters with the smallest ΔR_{ij} until all the distances between objects are above a threshold R_0 . These objects are identified as jets. Typical value for R_0 is 1.0.

3.5.3 Jet Calibration

After the reconstruction of the jets, several steps are needed to calibrate the jets [111]. The calibration chain of the reconstructed small- R jets and large- R jets is shown in Figure 3.16. The jet calibration can be separated into three major steps. First the pile-up contributions are suppressed at jet level, then the jets are calibrated to MC truth scale, and the last step accounts for the data MC difference and derives the systematics. Small- R jets have an origin correction to correct the jet axis to point to the hard scatter vertex of interest. It does not affect the jet energy. This section introduces the major steps of the calibration for both small- R jets and large- R jets.

Pile-up removal

The small- R jet pile-up removal is composed of a correction with the energy density ρ times jet area $\mathcal{A}_{\text{jet}}$ and a residual correction with the pile-up parameters. The jet area is evaluated for each individual jet, providing a measure of the local size of the jet. The energy density is calculated for individual event within $|\eta| < 2.0$, providing a measure of global pile-up energy presented in the event. The final correction can be written as:

$$p_T^{\text{corr}} = p_T - \rho \mathcal{A}_{\text{jet}} - \alpha(N_{\text{PV}} - 1) - \beta\mu. \quad (3.24)$$

The correction factors for N_{PV} and μ are derived from specific MC simulated samples, and then used in pile-up removal for all the jets. The factor α is parameterized as a function of p_T and N_{PV} , which corrects the jet energy dependence on number of primary vertices. The factor β is parameterized as a function of p_T and μ , which corrects the jet energy dependence on the mean number of pp interactions per bunch crossing.

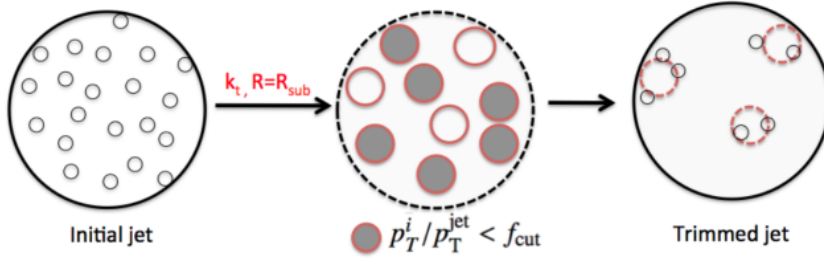
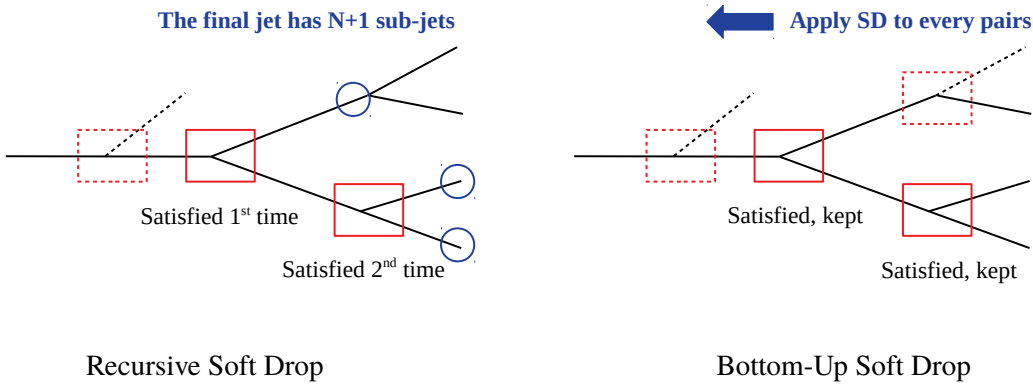


Figure 3.17: Process of the Trimming algorithm [112].



Recursive Soft Drop

Bottom-Up Soft Drop

Figure 3.18: An illustration of Recursive Soft Drop and Bottom-Up Soft Drop. The branch pairs satisfied the SD termination condition are marked with solid red box. The branches failed the termination condition are marked with dashed lines.

The large- R jet pile-up removal algorithm is grooming. Several kinds of grooming techniques have been developed. One of the grooming algorithm used for jets reconstructed from LCW scale topo-clusters is Trimming. It is designed to remove soft constituents in the jet. This algorithm first re-clusters the constituents of a large- R jet into several sub-jets of radius R_{sub} using k_t algorithm. Then any sub-jets with p_T fraction less than a cut f_{cut} are removed. This process is illustrated in Figure 3.17. Typical value of R_{sub} is 0.2, and of f_{cut} is 0.05.

Another grooming algorithm that is selected to be used with jets reconstructed from UFOs is Soft Drop (SD). It is designed to remove soft and wide radiation from a jet. This algorithm proceeds by splitting a jet into two branches using the last step of the C-A algorithm. If the two branches hold the condition:

$$\frac{\min(p_{T1}, p_{T2})}{p_{T1} + p_{T2}} < z_{\text{cut}} \left(\frac{\Delta R_{12}}{R_0} \right)^\beta, \quad (3.25)$$

the lower p_T branch is removed and the process is repeated with the updated jet. Otherwise, the process terminates and the remaining constituents form the jets. Typical values for z_{cut} and β are 0.1 and 1.0, respectively. Such configuration is to achieve a split that each one of the branches cannot have a p_T that is too small, and the angle between the two branches can not be too large. Two similar algorithms referred to as Recursive Soft Drop and Bottom-up Soft Drop are developed as alternatives of the SD.

3 Object reconstruction

Recursive SD applies SD to the reclustered jet until the termination condition is satisfied several times. For UFO jets, the algorithm proceeds the same as SD in the beginning. However, it does not stop when the termination condition is met for the first time. The Recursive SD continues applying SD to the branches in the remaining constituents. The algorithm proceeds until the termination condition can be met two times continuously, then the remaining constituents form the jets. Bottom-up SD applies the SD condition to every pair of branches in a completely reclustered jet. Any soft branches will be removed. An illustration of Recursive SD and Bottom-Up SD is shown in Figure 3.18.

Monte-Carlo calibration

Monte-Carlo (MC) calibration is to calibrate the jet energy and mass to particle scale at truth level. It is important because a large fraction (up to 60%) of the energy in a hadronic shower is invisible to the detector. For example, some of the energy can be deposited in materials that are not active sensors, the energy absorbed by nuclei is also not observed. Such lost energies have been taken into account partially at cluster level. A MC calibration at jet level is needed to completely recover the lost energies. The MC calibration compares reconstructed jets to jets built from simulated particles, referred to as truth jets. The response is defined as the ratio of the reconstructed quantity and the truth quantity:

$$R = \frac{X^{\text{reco}}}{X^{\text{truth}}}. \quad (3.26)$$

The jet energy scale (JES) calibration is derived in fine bins of energy and η . Corrections are applied to reconstructed jet energy using numerical inversion of the response. The JES calibration has the same procedure for both small- R and large- R jets, except the binning is different. The jet energy can be calibrated to truth scale with uncertainties smaller than 1%.

The large- R jets have a well-defined mass, so they are used to distinguish the hadronic decay of a massive particle. The calorimeter based mass is defined as the "invariant mass" of all the constituent in the jet,

$$m^{\text{calo}} = \sqrt{\left(\sum_{i \in J} E_i\right)^2 - \left(\sum_{i \in J} \vec{p}_i\right)^2}, \quad (3.27)$$

where i stands for constituents in a large- R jet. The other definition used to improve the resolution of the mass is the track-assisted mass, defined as

$$m^{\text{TA}} = \frac{p_{\text{T}}^{\text{calo}}}{p_{\text{T}}^{\text{track}}} \times m^{\text{track}}, \quad (3.28)$$

where $p_{\text{T}}^{\text{calo}}$ and $p_{\text{T}}^{\text{track}}$ are the jet momentum measured by the calorimeter and the tracker. For jets reconstructed from topo-clusters, the final mass is a combination of the calorimeter based mass and the track assisted mass. For UFO jets, just the calorimeter-based mass is enough as the track information has been already applied to improve the resolution of the constituents. The jet mass calibration (JMS) is derived in bins of energy, $\ln(m/E)$ and η . The large- R jet mass is calibrated to truth scale with uncertainties smaller than 5% in phase spaces that are important to physics analyses in Run 2. More details about the MC calibration and calibration performance are presented in Chapter 4.

Global sequential calibration and in-situ calibration

After the MC calibration, the small- R jets are processed by a global sequential calibration (GSC). The GSC [113] makes use of several variables that carry information about calorimeter response to jets. The correction is a function of jet properties binned in η . The dependence of the response on these jet properties are removed after applying this correction, without changing the mean of jet energy in each bin.

The final step in the calibration chain is the in-situ calibration [114, 115]. The in-situ calibration is to correct the data-MC difference and provide scale uncertainties. The calibration that had been done in Run 1 and Run 2 dataset uses the R -track method. The method uses the double ratio of measurements in data and MC to derive the correction factors,

$$R_{\text{trk}} = \frac{(X^{\text{jet}}/X^{\text{ref}})_{\text{data}}}{(X^{\text{jet}}/X^{\text{ref}})_{\text{MC}}}. \quad (3.29)$$

Here X can be jet mass, p_T or substructure variables. The superscript "jet" refers to variables of the jet. The superscript "ref" represents well measured reference objects like tracks, electrons or photons that are produced together with the jet in the same pp collision. The in-situ p_T calibration uses Z -jets and γ -jets events, where Z bosons decay to lepton pairs. The well measured Z or γ works as the reference object because their uncertainties are much smaller than the jet uncertainties. This double ratio should be equal to unity if the simulation of jets agrees perfectly with data. Therefore any deviation from unity can be treated as a scale uncertainty or a correction to the measurement in data. Figure 3.19 shows the derived double ratio for large- R jet p_T . The in-situ mass calibration use

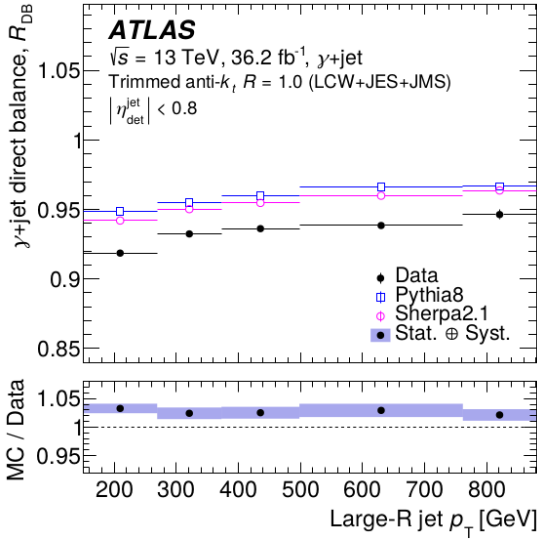


Figure 3.19: Double ratio of p_T measurement for large- R trimmed jets reconstructed from LCW topo-clusters, with JES and JMS calibration applied [115].

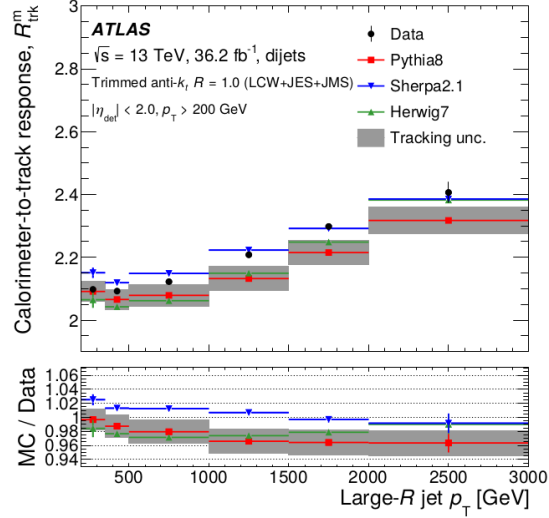


Figure 3.20: Double ratio of mass measurement for large- R trimmed jets reconstructed from LCW topo-clusters, with JES and JMS calibration applied [115].

calorimeter and tracker measurements to derive the double ratio. Figure 3.20 shows the double ratio of mass measurements derived from di-jet samples.

3 Object reconstruction

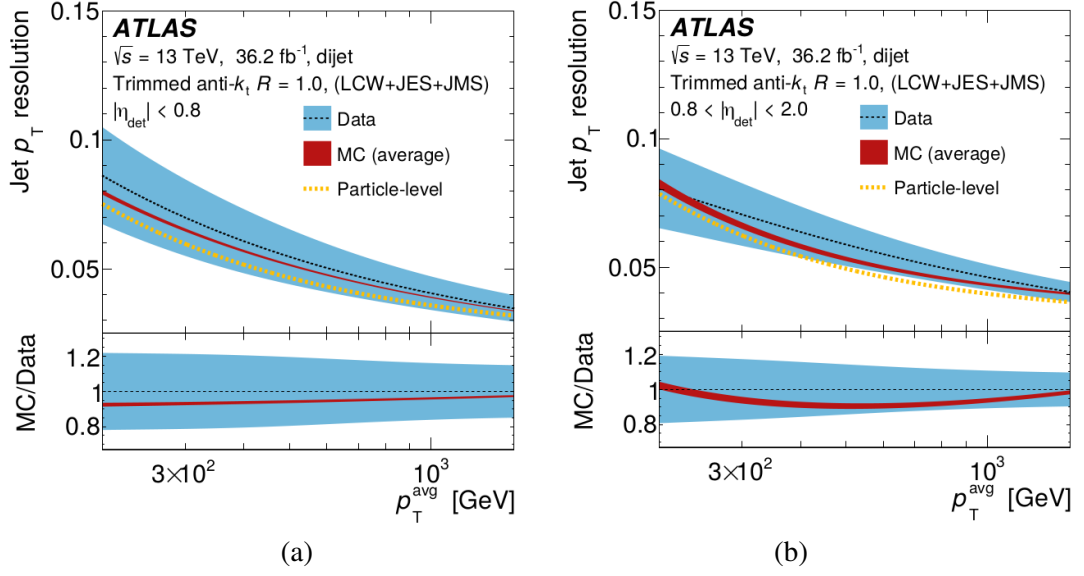


Figure 3.21: Measured jet p_T resolution for large- R jets in Run 2, as a function of p_T in two η regions: (a) $|\eta| < 0.8$, (b) $0.8 < |\eta| < 2.0$ [115].

R -track method can be used to study several sources of uncertainties. The baseline uncertainty accounts for the difference between data and the default MC generator. The modeling uncertainty is derived with alternative MC generators. The tracking uncertainty accounts for the effect of various tracking uncertainty sources. And the statistical uncertainty comes from the limited statistics in data and simulated samples. For $VH(b\bar{b})$ boosted analysis, the small- R jets had the in-situ calibration available, the large- R jets did not have the in-situ calibration because it was not ready when the analysis was done. For the $VH(b\bar{b})$ legacy analysis, both small- R and large- R jets have the in-situ calibration included.

The jet resolution σ is defined by the standard derivation of the Gaussian fit to the jet response distributions. The jet energy resolution (JER) is measured through the relative p_T resolution σ_{p_T}/p_T after the calibrations, as the relative energy resolution is approximately equal to the relative p_T resolution. The p_T resolution is estimated with the dijet balance method [115]. The measured relative p_T resolution is shown in Figure 3.21.

3.6 Missing transverse energy reconstruction

The neutrinos are invisible to the ATLAS detector. However, they can be indirectly identified through the imbalance of the transverse momentum of an event. The missing transverse momentum, E_T^{miss} , is defined as the imbalance of the total transverse momentum.

$$\vec{E}_T^{\text{miss}} = - \left(\sum_{\text{selected}} \vec{p}_T + \sum_{\text{soft}} \vec{p}_T \right), \quad (3.30)$$

3.6 Missing transverse energy reconstruction

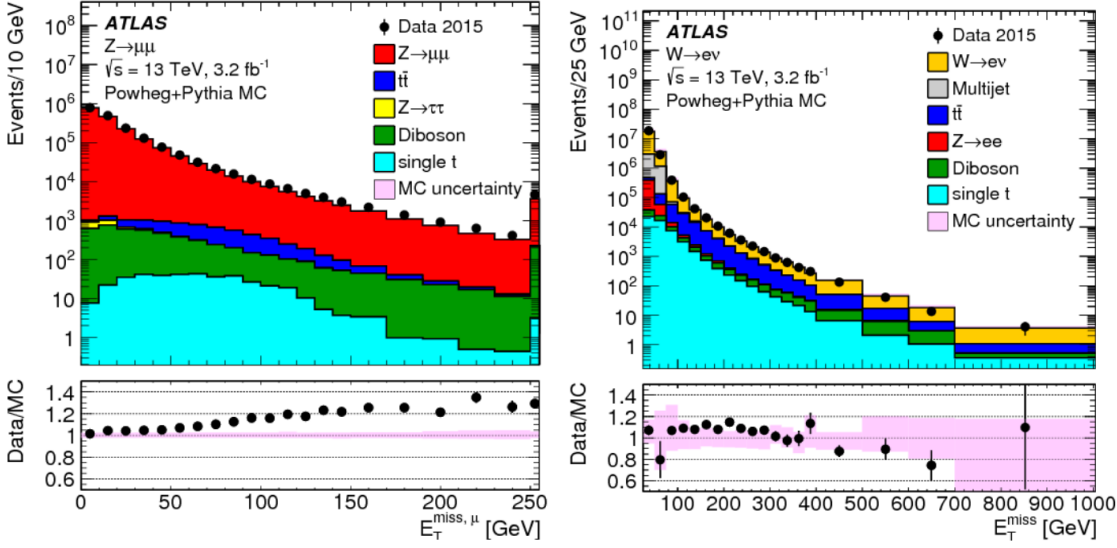


Figure 3.22: Distributions of E_T^{miss} for the samples of inclusive $Z \rightarrow \mu\mu$ (left) and $W \rightarrow e\nu$ (right) events in data, compared to MC simulations [116].

where the "selected" includes all the identified objects such as leptons, photons and jets, the "soft" includes all the remaining low- p_T objects that are associated with the hard-scatter vertex. The reconstruction of E_T^{miss} involves measurements from all the sub-detectors. Each kinds of object is reconstructed with a dedicated algorithm. The reconstructed objects are combined following the order: muons and electrons first, then photons, then hadronically decaying τ -leptons, jets and finally the soft objects.

Unlike fully reconstructed and calibrated particles or jets, E_T^{miss} reconstruction has a non-linear response. The resolution is also complex due to the composition of different objects. Therefore the performance of E_T^{miss} reconstruction is assessed by comparing variables related to E_T^{miss} in data and MC. The signal samples are $Z \rightarrow \mu\mu$ for a final state without genuine E_T^{miss} , and $W \rightarrow e\nu$ for a final states with genuine E_T^{miss} . Figure 3.22 shows the comparison of reconstructed E_T^{miss} in data and MC for the two final states. The data and MC simulations agree within 20% for the bulk of the E_T^{miss} distributions. The only larger differences exist in high E_T^{miss} regions. The E_T^{miss} resolution is determined by the width of the distribution of the differences between the measured E_T^{miss} and the truth $E_T^{\text{miss,true}}$.

$$\text{Resolution}_{x(y)} = \text{RMS}(E_{x(y)}^{\text{miss}} - E_{x(y)}^{\text{miss,true}}). \quad (3.31)$$

The resolution is measured with inclusive $Z \rightarrow \mu\mu$ samples and $W \rightarrow e\nu$, $W \rightarrow \mu\nu$ samples. Figure 3.23 shows the measured E_T^{miss} resolution. For the inclusive $Z \rightarrow \mu\mu$ sample, the resolution slowly increases with the total transverse momentum. The resolution measured with $W \rightarrow e\nu$ and $W \rightarrow \mu\nu$ samples are very similar. The MC simulations agree well with data.

3 Object reconstruction

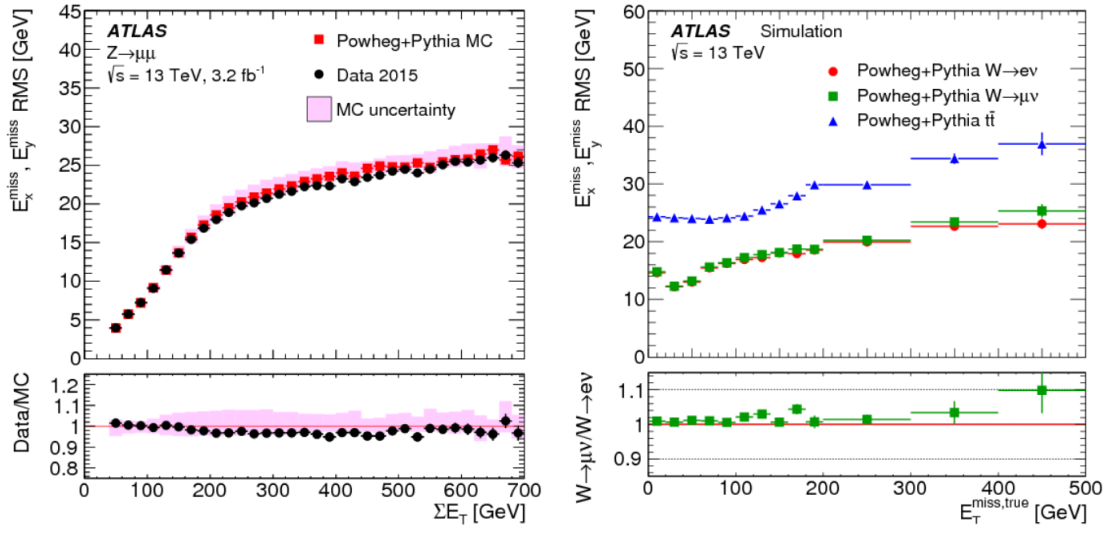


Figure 3.23: E_T^{miss} resolution measured in inclusive $Z \rightarrow \mu\mu$ samples (left) and $W \rightarrow e(\mu)\nu$ (right) samples [116].

4 Calibration of reconstructed jet energy and mass

The reconstructed jets have to be processed with several steps of calibration to calibrate the jet energy and mass. One of the steps is the Monte-Carlo (MC) based calibration. The MC based jet energy calibration (JES) is applied to both small- R jets and large- R jets. The MC based jet mass calibration (JMS) is applied to large- R jets, and has been studied with small- R jets. This chapter describes the procedure of MC calibration, and presents the calibration results of several jet collections.

4.1 Jet energy calibration

The MC JES calibration is derived in bins of pseudorapidity η and jet energy. The derivation follows the same procedure in every η bins. MC calibration compares reconstructed jet energy to truth jet energy, therefore the reconstructed jets need to be matched to the truth jets to start the calibration. The selection of the matching procedure requires the angular distance between the reconstructed jet and the truth jet being small,

$$\Delta R(\text{truth jet, reco jet}) < 0.3 \text{ (0.1 for tracks)}, \quad (4.1)$$

and both the reconstructed jet and the truth jet have to be isolated.

$$\begin{aligned} \Delta R(\text{reco jet, reco jet}) &> 1.5 \times \text{jet radius}, \\ \Delta R(\text{truth jet, truth jet}) &> 2.5 \times \text{jet radius}. \end{aligned} \quad (4.2)$$

The calibration can be divided in three major steps. In the first step, the algorithm derives the jet energy response as functions of truth jet energy and pseudorapidity.

$$\langle R \rangle = \langle E^{\text{reco}} / E^{\text{true}} \rangle \xrightarrow{\text{in bins of } E^{\text{true}}, \eta} R(E^{\text{true}}, \eta) \quad (4.3)$$

In the second step, the truth jet energy is replaced with the reconstructed jet energy estimated from the truth jet energy times the energy response. The jet energy response as a function of reconstructed jet energy and η can be used as the JES calibration factor.

$$E^{\text{calib}} = \frac{1}{R(E^{\text{true}} \cdot R(E^{\text{true}}))} \cdot E^{\text{reco}}. \quad (4.4)$$

Finally, a closure test is applied using the derived JES factor to evaluate the calibration performance. The following part of this section gives a detailed description of the first two calibration steps.

4 Calibration of reconstructed jet energy and mass

The goal of the first step is to get the jet energy response $R = E^{\text{reco}}/E^{\text{true}}$ as a function of E^{true} and η , where E^{reco} is the reconstructed jet energy, E^{true} is the truth jet energy. The calibration is derived in every η bins individually. The procedure of the derivation in one η bin can be summarized as:

1. Calculating the average of energy response R in every E^{true} bin. The average is evaluated using Gaussian fit to the distribution of the energy response. The fit can be invoked three times to optimize the fitting result. The fitting range is set to the mean value of the histogram plus and minus 1.5 times of the root mean square (RMS) initially, then modified to the mean value of the fitted function plus and minus 1.6 times its σ if the first fit is successful. If the modified fitting range exceeds the original, the original fitting range is kept.
2. Calculating the average of E^{true} with the jets used in the fit of the response. Together with the average of R , the algorithm gets a $\langle R \rangle$ vs $\langle E^{\text{true}} \rangle$ graph in every η bin.
3. Performing a fit to derive the response as a function of truth energy using the $\langle R \rangle$ vs $\langle E^{\text{true}} \rangle$ graphs. The methods of fitting are Polynomial regression [117] or a combination of grooming and Polynomial regression. The grooming works as an extrapolation of the fitting range. The model of the fit is a 8th-degree polynomial parameterized as:

$$R = p_0 + p_1 \ln(x) + p_2 [\ln(x)]^2 + \dots + p_8 [\ln(x)]^8. \quad (4.5)$$

Here x is E^{true} , and p_i are the coefficients derived from the fit.

This response function $R(E^{\text{true}}, \eta)$ cannot be used as the JES factor directly because the scale factor needs to be applied to reconstructed jet energy, while this response is a function of truth jet energy. A numerical inversion is applied to transform the response into a function of reconstructed energy in the second step.

The goal of the second step is to transform the response from a function of E^{true} to a function of E^{reco} . A numerical inversion technique is used, the procedure can be summarized as:

1. Calculating an estimated reconstructed energy for every reco-truth jet pairs,

$$E^{\text{reco(NI)}} = E^{\text{true}} \times R(E^{\text{true}}, \eta). \quad (4.6)$$

This estimated energy is not the same as the E^{reco} of the jet because it measures an average deviation of the reconstructed energy from the truth energy, while the deviation of the original E^{reco} is not averaged.

2. Calculating the average of $E^{\text{reco(NI)}}$ using the same jets as the ones used in the fit of the response in the first step. In other words, the fitted response in every E^{true} bin is not changed, but the average of energy is replaced with $\langle E^{\text{reco(NI)}} \rangle$. The algorithm gets a $\langle R \rangle$ vs $\langle E^{\text{reco(NI)}} \rangle$ graph in every η bin.
3. Performing a fit to the $\langle R \rangle$ vs $\langle E^{\text{reco(NI)}} \rangle$ graphs to get energy response as a function of $E^{\text{reco(NI)}}$, using the same method as the first step.

The inverse of this function $R(E^{\text{reco(NI)}}, \eta)$ is the JES factor. For a jet with reconstructed energy E^{reco} and pseudorapidity η , the energy is calibrated as:

$$E^{\text{calib}} = \frac{1}{R(E^{\text{reco}}, \eta)} \cdot E^{\text{reco}}, \quad (4.7)$$

where $R(E^{\text{reco}}, \eta)$ is the same function as $R(E^{\text{reco(NI)}}, \eta)$. Finally, a closure test is performed with the calibrated jet energy to evaluate the calibration performance.

4.2 Jet mass calibration

The JMS calibration differs vastly from the JES calibration in binning and the methods to derive the response, but the ideas are similar. The reconstructed large- R jet mass is also compared to truth jet mass using the same selected reco-truth jet pairs. The mass response $R_m = m^{\text{reco}}/m^{\text{true}}$ is derived in bins of E , $\ln(m/E)$ and η . The JMS calibration starts from JES calibrated jets, the JES factor is applied to both the reconstructed jet energy and the reconstructed jet mass.

Similar to the JES calibration, the JMS calibration can also be divided in three major steps. The algorithm derives the mass response as functions of E^{true} , $\ln(m^{\text{true}}/E^{\text{true}})$, η in the first step.

$$\langle R_m \rangle = \langle m^{\text{reco}}/m^{\text{true}} \rangle \xrightarrow[\ln(m^{\text{true}}/E^{\text{true}}), \eta]{\text{in bins of } E^{\text{true}}} R_m(E^{\text{true}}, \ln(m^{\text{true}}/E^{\text{true}}), \eta). \quad (4.8)$$

In the second step, E^{true} is replaced with E^{calib} , which is assumed to be very close to the truth jet energy. The truth jet mass is replaced with the estimated reconstructed jet mass. The inverse of jet mass response as a function of the estimated reconstructed jet mass, E^{calib} and η is the JMS calibration factor.

$$m^{\text{calib}} = \frac{1}{R_m(m^{\text{true}} \cdot R_m)} \cdot m^{\text{reco}}. \quad (4.9)$$

The third step is a closure test for evaluation of the JMS calibration performance. The following part of this section gives a detailed description of the first and the second steps of the calibration.

The JMS calibration is derived in every η bins individually. In each η bin, the calibration proceeds by calculating the average of R_m in every $(E^{\text{true}}, \ln(m^{\text{true}}/E^{\text{true}}))$ bins. The average is calculated using Gaussian fit to the distribution of R_m when the statistics suffices. The Gaussian fit uses the same method as the energy response fitting, but the fitting range is set to $[0, 2]$ initially, and then modified to the mean value of the fitted function plus and minus 1.3 times of its σ . If the R_m histogram does not have enough entries, the average is evaluated with the mean of the distribution. Unlike the JES calibration, the average of E^{true} and $\ln(m^{\text{true}}/E^{\text{true}})$ are represented using the bin centers in the JMS calibration. The results are 2D distributions of $\langle R_m \rangle$ vs E^{true} , $\ln(m^{\text{true}}/E^{\text{true}})$.

The mass response distributions need to be smoothed to get ride of the sharp edges between neighboring bins. The smoothing serves similar purpose compared to the fitting of $\langle R \rangle$ vs $\langle E \rangle$ graphs in JES calibration. It uses a 2D Gaussian kernel smoothing [118] procedure. The smoothing works by implementing this formula:

$$R(\mathbf{x}_0) = \frac{\sum_{i=1}^N K(\mathbf{x}_0, \mathbf{x}_i) R(\mathbf{x}_i)}{\sum_{i=1}^N K(\mathbf{x}_0, \mathbf{x}_i)}, \quad (4.10)$$

4 Calibration of reconstructed jet energy and mass

where $K(\mathbf{x}_0, \mathbf{x})$ is the Gaussian kernel, $R(\mathbf{x}_i)$ is the mass response at $\mathbf{x}_i$, $\mathbf{x}$ is the scaled coordinates for E and $\ln(m/E)$. The Gaussian kernel is a Gaussian function of $|\mathbf{x} - \mathbf{x}_0|$, with its peak at $\mathbf{x}_0$ and σ set to 0.3.

$$K(\mathbf{x}_0, \mathbf{x}_i) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{|\mathbf{x} - \mathbf{x}_0|^2}{2\sigma^2}\right). \quad (4.11)$$

The coordinates for E and $\ln(m/E)$ are scaled to be within $[0,1]$:

$$x_i \rightarrow \frac{x_i - x_0}{x_{\max} - x_0}, \quad (4.12)$$

where x can be E or $\ln(m/E)$, the subscript "0" represents the point that is being smoothed, the subscript "i" represents other points in the graph. The kernel smoothing is applied to every points in the graph. All the points within $|\mathbf{x} - \mathbf{x}_0| < 0.3$ are taken into account to calculate the kernel weighted average. The smoothed R_m distributions are used as mass response function $R_m(E^{\text{true}}, \ln(m^{\text{true}}/E^{\text{true}}), \eta)$.

The second step of the JMS calibration transforms the mass response from functions of truth jet parameters to functions of reconstructed jet parameters. The algorithm calculates an estimated reconstructed jet mass for the numerical inversion:

$$m^{\text{reco(NI)}} = m^{\text{true}} \times R_m(E^{\text{true}}, \ln(m^{\text{true}}/E^{\text{true}}), \eta). \quad (4.13)$$

Then the average of R_m is calculated in every $(E^{\text{calib}}, \ln(m^{\text{reco(NI)}}/E^{\text{calib}}))$ bins using the same method as in the first step. The results are distributions of R_m in bins of $E^{\text{calib}}, \ln(m^{\text{reco(NI)}}/E^{\text{calib}})$. These R_m distributions are also smoothed using the same method as in the first step. The final results are smoothed R_m distributions in bins of $E^{\text{calib}}, \ln(m^{\text{reco(NI)}}/E^{\text{calib}})$ for every η bins. Another option used in the calibration of UFO jets is the direct numerical inversion. Instead of re-fitting the mass response distributions, it modifies the y-axis of the R_m distributions from $\ln(m^{\text{true}}/E^{\text{true}})$ to $\ln(m^{\text{reco(NI)}}/E^{\text{true}})$ directly.

A flattening requirement can be enforced to the inverse of R_m to avoid sharp edges caused by purely computational reasons.

$$\frac{dC(m^{\text{reco(NI)}})}{dm^{\text{true}}} \cdot m^{\text{true}} \cdot \sigma(m^{\text{true}}) \leq 0.1, \quad (4.14)$$

where $C = 1/R_m$, and $m^{\text{true}} \cdot \sigma(m^{\text{true}})$ is an equivalent to the width of m^{reco} . In regions failed this requirement, the value of dC/dm^{true} is modified to $0.1/(m^{\text{true}} \cdot \sigma(m^{\text{true}}))$ to get rid of the sharp edges.

The JMS factor is evaluated from the smoothed R_m distributions:

$$\text{JES} : m^{\text{reco}} = \frac{1}{R(E^{\text{reco}}, \eta)} \cdot m^{\text{raw reco}}, \quad (4.15)$$

$$m^{\text{calib}} = \left[R_m(E^{\text{calib}}, \ln(m^{\text{reco}}/E^{\text{calib}}), \eta) \right]^{-1} \cdot m^{\text{reco}}. \quad (4.16)$$

Finally, a closure test is performed with the calibrated jet mass to evaluate the calibration performance.

4.3 Performance of the calibration

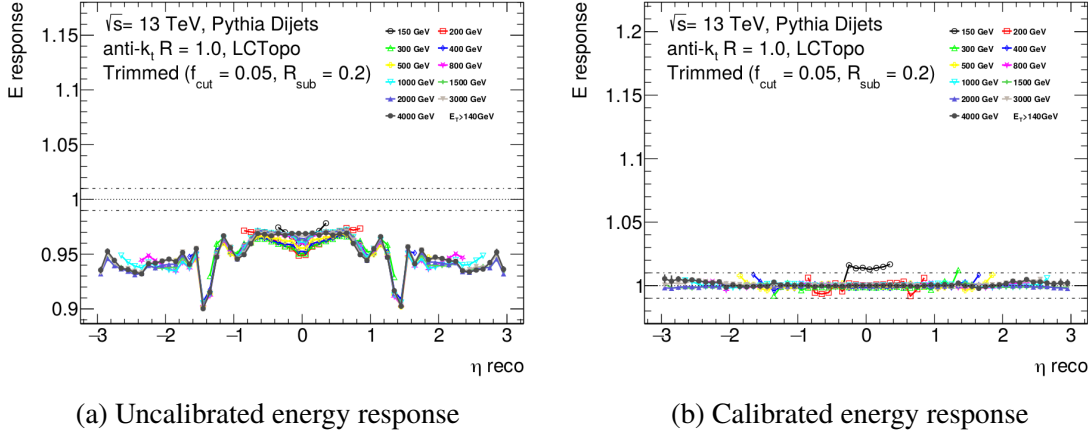


Figure 4.1: Distributions of energy response for LCTopo Trimmed jets, in bins of η at different E^{true} slices. The values of R are evaluated from R vs E^{true} graphs derived in the calibration and the closure test.

4.3 Performance of the calibration

Several large- R jet definitions have been studied during the optimization of the jet reconstruction. The MC JES and JMS calibration are derived for all the jet collections. This section presents the MC calibration performance for large- R jets reconstructed from LCW scale topo-clusters, TCC objects and UFO objects. A small- R jet collection for low pile-up environment is also studied, and can be found in Appendix A.

4.3.1 $R = 1.0$ jets reconstructed from topo-clusters

The large- R jet collection used in many analyses including $VH(b\bar{b})$ is built from local cell weight (LCW) energy scale topo-clusters, denoted as LCTopo in the following of this section. The LCW scale accounts for the energy loss in a hadronic shower. The MC calibration is able to calibrate the jet energy to truth scale completely. The grooming algorithm applied to this jet collection is the Trimming algorithm described in Section 3.5.3, with the parameters set to $f_{\text{cut}} = 0.05$, $R_{\text{sub}} = 0.2$.

The JES calibration is derived in η bins from -3.0 to 3.0 , with bin width of 0.1 , and E bins from 60 GeV to 5000 GeV, with various bin widths. The MC samples used to derive the calibration are QCD di-jet processes generated using Pythia 8. The calibration performance is evaluated through a closure test that derives the energy response using the calibrated jet energy. The closure test derive the calibrated R vs E^{true} graphs with the same procedure as the first step of the calibration. The distributions of energy response for uncalibrated jets and calibrated jets are shown in Figure 4.1. For a given energy slice and η bin, the value of energy response is evaluated from the R vs E^{true} graph at the energy slice and the matched η bin. Figure 4.1 (a) is the energy response for uncalibrated jets. Figure 4.1 (b) is the energy response for calibrated jets. The energy response becomes very close to unity after the calibration, which means the average of $E^{\text{calib}}/E^{\text{true}}$ is very close to unity in the range of the JES calibration. It proves the calibrated jet energy scale is very close to the truth jet energy scale and the uncertainty is within 1%.

4 Calibration of reconstructed jet energy and mass

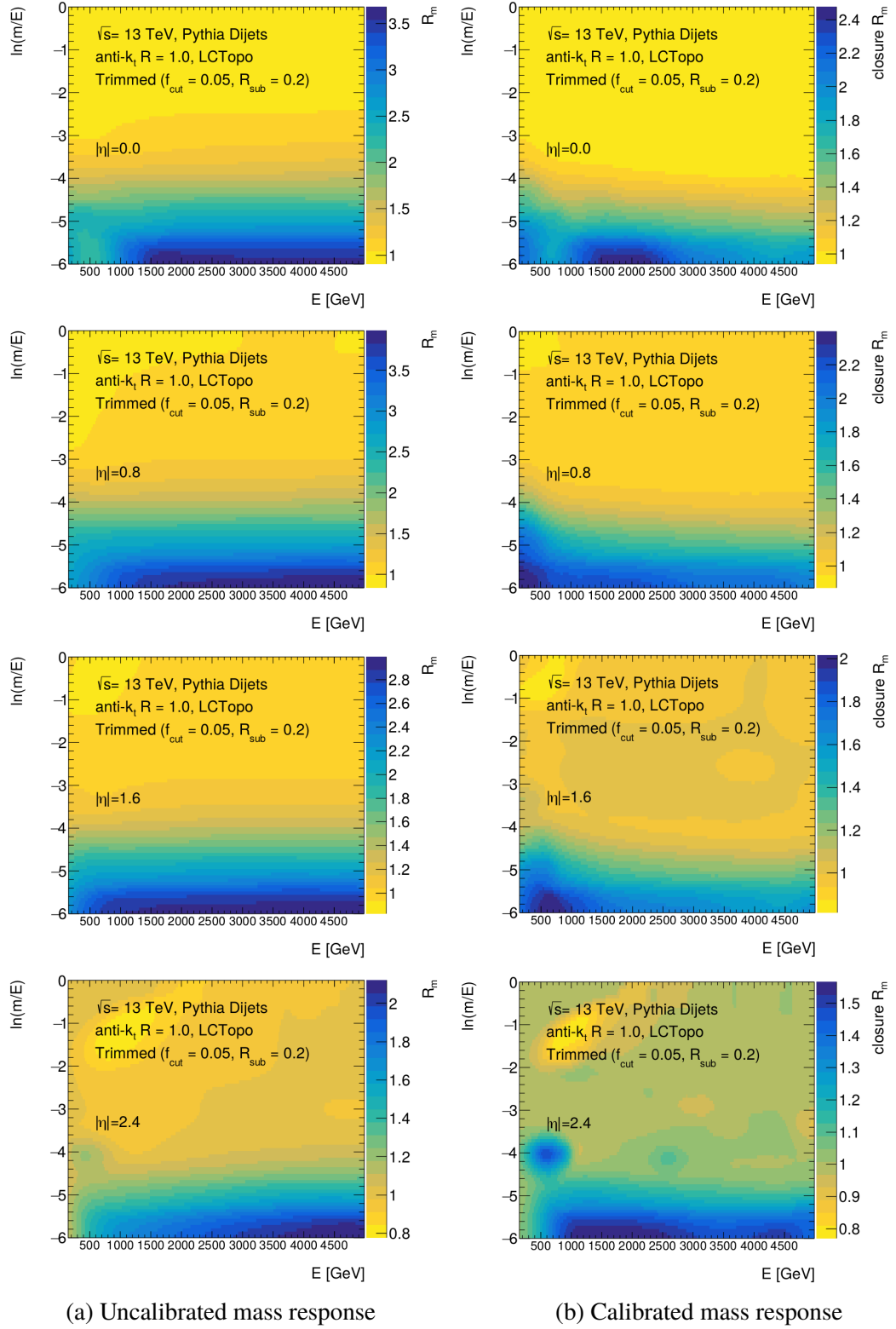


Figure 4.2: Distributions of calorimeter-based mass response for LCTopo Trimmed jets, plotted in bins of E^{true} , $\ln(m^{\text{true}}/E^{\text{true}})$ at several $|\eta|$ slices. The values are evaluated from R_m vs E^{true} , $\ln(m^{\text{true}}/E^{\text{true}})$, $|\eta|$ histogram generated in the JMS calibration and closure test at given positions.

4.3 Performance of the calibration

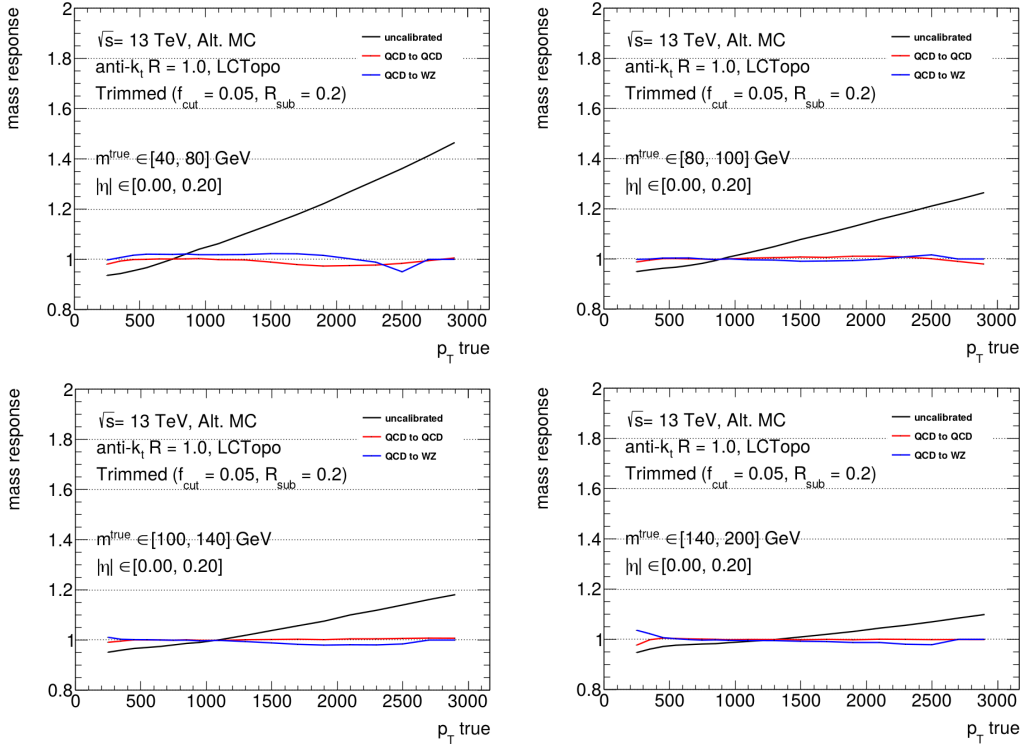


Figure 4.3: Distributions of R_m from closure test of JMS calibration derived with QCD di-jet samples applied to W/Z jets for LCTopo Trimmed jet collection. The examples are in central $|\eta|$ bin, with four representative mass bins covering W boson mass, Z boson mass, Higgs boson mass and top quark mass.

The JMS calibration is derived in $|\eta|$ bins from 0 to 3.0, with various bin widths of 0.1, 0.2, 0.25 and 0.3. The energy bins are from 200 GeV to 1200 GeV with bin width of 200 GeV, from 1200 GeV to 4000 GeV with bin width of 400 GeV, and a final bin from 4000 GeV to 6000 GeV. The $\ln(m/E)$ bins are from -6.0 to 0 with various bin widths. The MC samples used in the derivation are the same as the ones used in JES calibration. The JMS calibration performance is also evaluated through a closure test that derives mass response using the calibrated jets. The derived mass response of calibrated jets are stored in closure R_m vs E^{true} , $\ln(m^{\text{true}}/E^{\text{true}})$ histograms, so they can be compared to the R_m histograms generated in the first step of the calibration. The mass definition for LCTopo jets is a combination of calorimeter-based mass m^{calo} and track-assisted mass m^{TA} , as described in Section 3.5.3. Figure 4.2 shows the distributions of R_m derived from m^{calo} for uncalibrated jets and calibrated jets. Calibration of m^{TA} has similar level of performance. The mass response is close to unity in most of the studied regions after the JMS calibration. The only non-closures in the R_m maps occur in regions where $\ln(m/E)$ and E are small. These regions correspond to jets with $m < 40$ GeV, which is not used in the physics analyses in Run 2. These non-closures are caused by several factors. The most important reason is the reconstructed mass itself becomes more and more inaccurate with smaller mass, and the calibration performance depends on the precision level of the reconstructed mass.

The MC calibration derived using QCD di-jet samples is tested with alternative MC samples to make

4 Calibration of reconstructed jet energy and mass

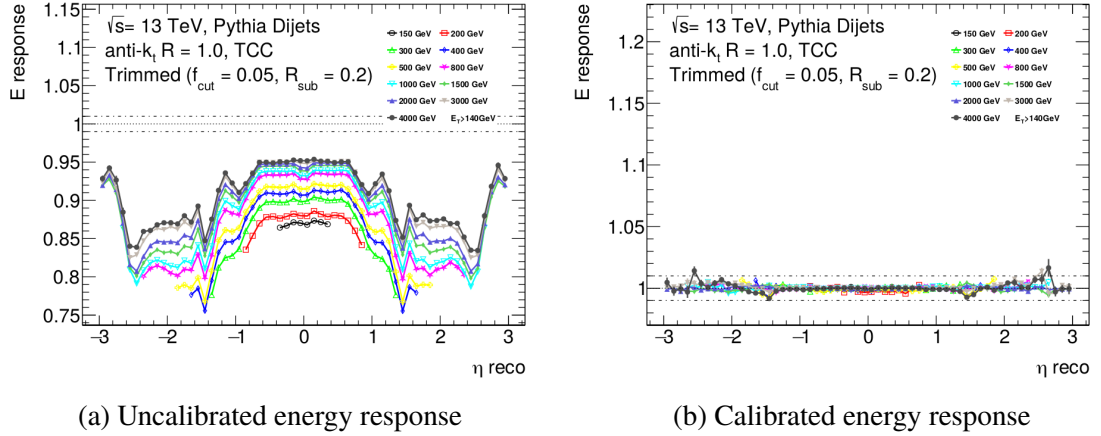


Figure 4.4: Distributions of energy response for TCC Trimmed jets, in bins of η at different E^{true} slices.

sure the calibration can be applied to jets originate from different processes. The alternative samples are large- R jets originate from hadronically decayed W or Z bosons. The JMS calibration derived with QCD di-jet samples is applied to W/Z jets in closure test. Examples of the calibrated mass response distributions are shown in Figure 4.3. Calibration derived with di-jet samples has similar level of performance when applied to alternative MC samples. The uncertainties can be limited within 5% in the regions that are important to Run 2 physics analyses. Some very high p_T regions with $p_T > 1500$ GeV and $\eta > 1.4$ have larger non-closure due to limited statistics. The non-closures in those regions do not affect the JMS calibration performance. LCTopo jet was selected as the standard large- R jet definition in the ATLAS experiment and has been successfully used in physics analyses including the boosted $VH(b\bar{b})$ analysis.

4.3.2 $R = 1.0$ jets reconstructed from Track Calo Clusters

The large- R jets reconstructed from TCC objects were studied together with LCTopo jets. The MC samples used in the derivation of the calibration are the same as the ones used in the calibration of LCTopo jets, but the jets are reconstructed using a different definition, which is the TCC algorithm. The grooming algorithm of this jet collection is Trimming as well.

The JES calibration and the JMS calibration for TCC jets have the same binning as the configuration in LCTopo jet calibrations. The JES calibration performance is shown in Figure 4.4. The calibrated jet energy scale is very close to the truth energy scale, with an uncertainty smaller than 1%. It is at similar precision level as the LCTopo jets. The performance of JMS calibration for TCC jets is also similar to LCTopo jets. Figure 4.5 shows an example of R_m distributions for uncalibrated and calibrated TCC jets. The calibrated mass response is close to unity in the regions that are important to physics studies. However, the reconstructed mass of TCC jet is topology dependent. The QCD derived calibration has around 10% non-closure everywhere when applied to W/Z jets, which can cause problems in W/Z jet tagging. It means the reconstructed mass for QCD jets and W/Z jets with the same level of truth mass are different, such difference cannot be calibrated at the same time. The difference in

4.3 Performance of the calibration

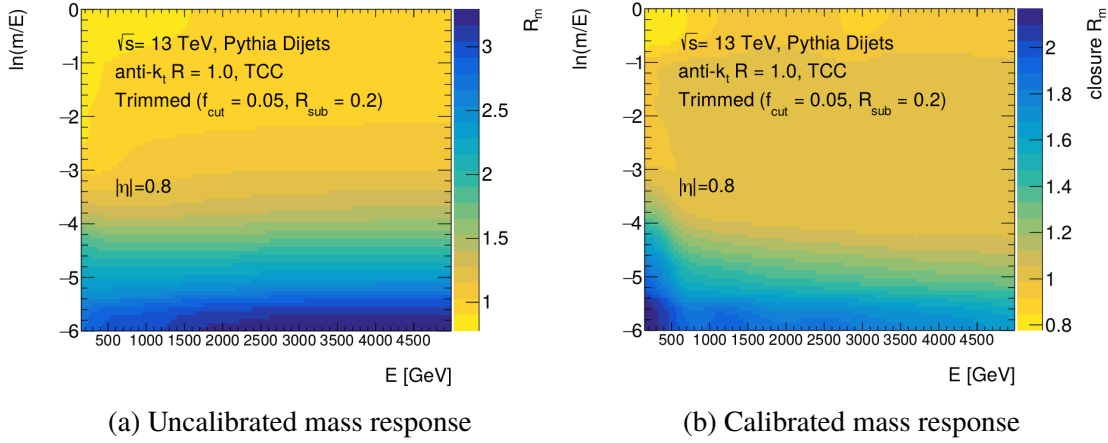


Figure 4.5: Example distributions of mass response for TCC Trimmed jets, plotted in bins of E^{true} , $\ln(m^{\text{true}}/E^{\text{true}})$ at $|\eta| = 0.8$.

reconstructed mass for jets with the same m^{true} comes from the the TCC algorithm, which is topology dependent so the different jet substructure can result in different reconstructed mass.

TCC jets did not become the default large- R jet collection in the ATLAS experiment. It was integrated into UFO reconstruction. The topology dependence is reduced significantly with the UFO jets, which is presented in the next section.

4.3.3 $R = 1.0$ jets reconstructed from Unified Flow Objects

Several kinds of jets built from UFOs are studied during the optimization process for a precision recommendation that can be used in Run 3. The jet collections are reconstructed from UFOs with pile-up mitigation algorithms CS and SK applied at cluster level, as described in Section 3.5.1. The grooming algorithms are Soft Drop, Recursive Soft Drop and Bottom-Up Soft Drop. Trimming is also studied as a backup. This section presents the calibration performance of the UFO CSSK Soft Drop jet collection, which is the one recommended as the default in the ATLAS experiment. The calibration performance of UFO jets with the other two Soft Drop algorithms are very close. The decision of recommendation was made with the completed performance studies including cluster reconstruction, jet calibration and tagging.

Three kinds of MC samples are used in the derivation and the evaluation of the calibration. QCD di-jet processes generated by PYTHIA 8 generator is used in the derivation of JES calibration and the majority of the JMS calibration. The other two processes: $W' \rightarrow WZ$ and $Z' \rightarrow t\bar{t}$ generated by PYTHIA 8, are used to provide large- R jets originate from W/Z boson decays and top quark decays. The events are processed by different jet definitions to form the UFO jets with different grooming algorithms. The JES calibration is derived in the same bins as the bins used in the calibration of LCTopo and TCC jets. The distributions of energy response for uncalibrated and calibrated UFO Soft Drop jets are shown in Figure 4.6. The energy response is very close to unity after the calibration, with an uncertainty smaller than 0.05% in all the calibrated regions. The energy of UFO jets is calibrated better than

4 Calibration of reconstructed jet energy and mass

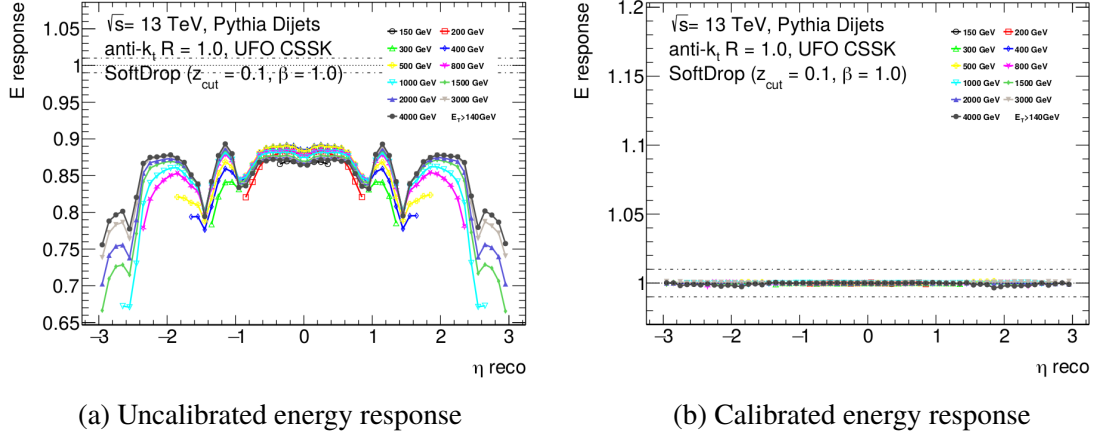


Figure 4.6: Distributions of energy response for UFO Soft Drop jets, in bins of η at different E^{true} slices.

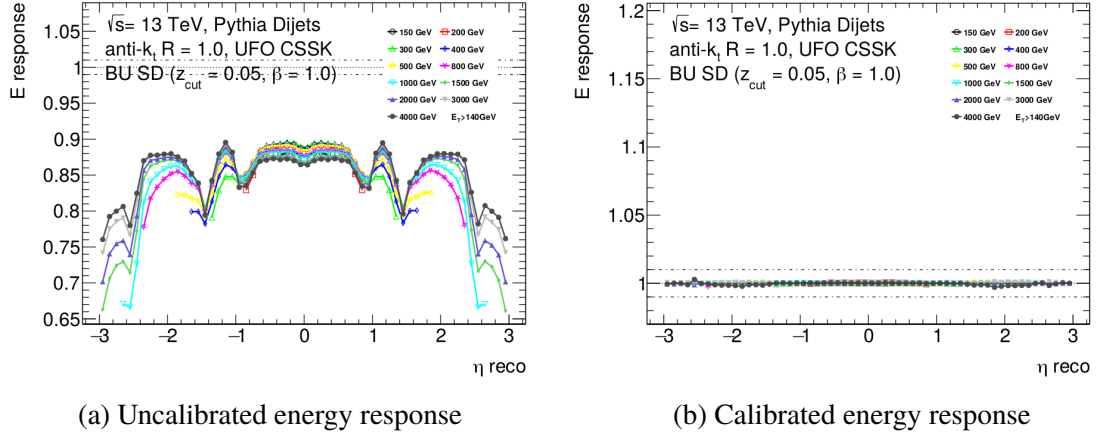


Figure 4.7: Distributions of energy response for UFO Bottom-Up Soft Drop jets, in bins of η at different E^{true} slices.

LCTopo jets, as a result of better reconstructed UFO jet energy. The JES calibration performance for UFO Bottom-Up Soft Drop jets is shown in Figure 4.7. The closure test result of Recursive Soft Drop jets is almost identical as the result of Bottom-Up Soft Drop jets. The energy of these two jet collections is calibrated to the same precision level as the standard Soft Drop jet collection.

The JMS calibration of UFO Soft Drop jets was originally derived using QCD di-jet samples. The binning is the same as the ones used in JMS calibration of LCTopo and TCC jets. Figure 4.8 shows the distributions of uncalibrated and calibrated mass response for UFO Soft Drop jets, in bins of E , $\ln(m/E)$ at different η slices. The calibrated mass response is close to unity in most of the studied regions. Some regions with $\ln(m/E) < 4$ still has large non-closure after the calibration, due to the inaccuracy of the reconstructed mass. These regions correspond to jet mass $m < 40$ GeV, so the non-closure does not have effect on the performance of the calibration.

The topology dependence observed in the calibration of TCC jets still exists in regions with

4.3 Performance of the calibration

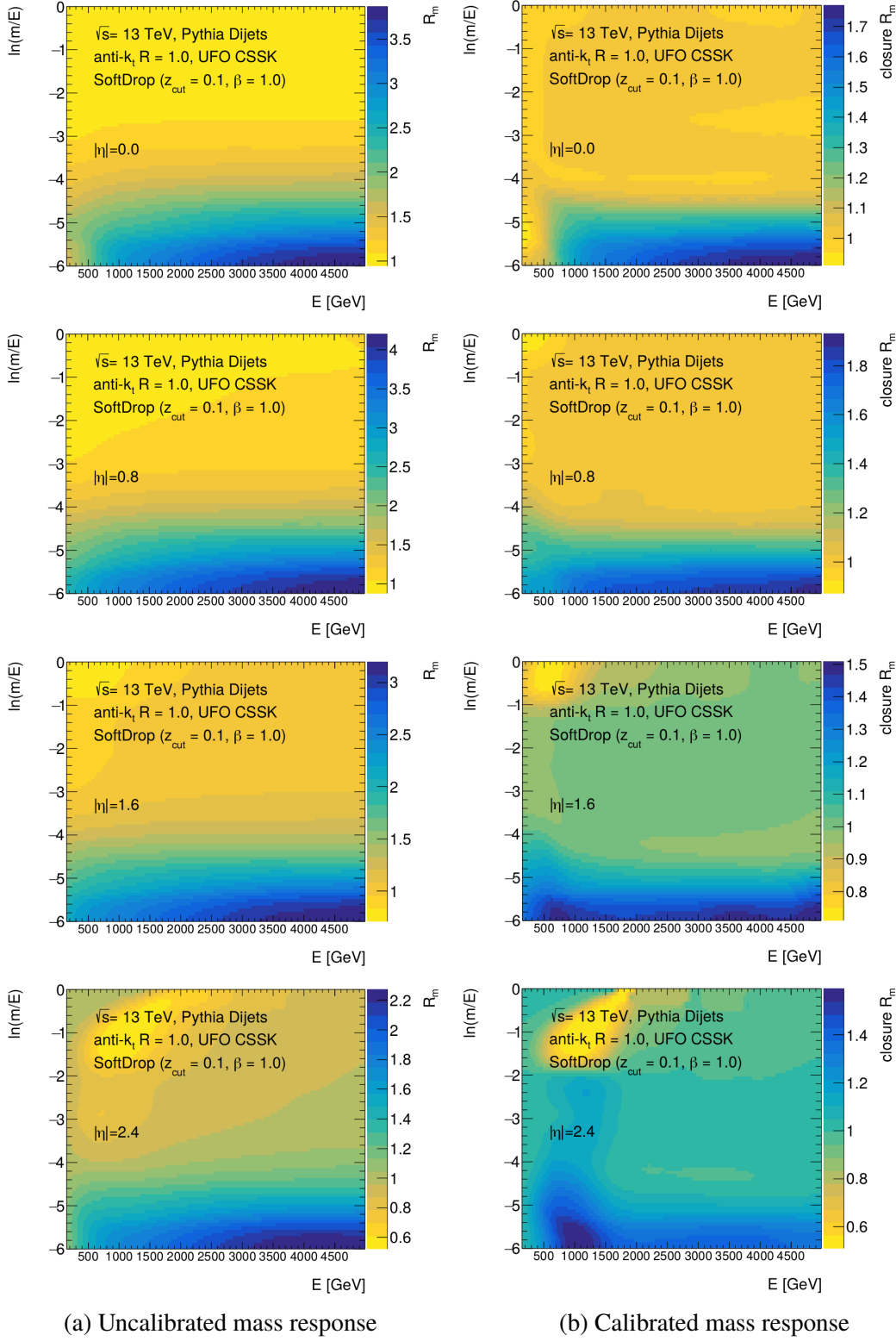


Figure 4.8: Distributions of mass response for UFO Soft Drop jets, plotted in bins of E^{true} , $\ln(m^{\text{true}}/E^{\text{true}})$ at several $|\eta|$ slices. The values are evaluated from R_m vs E^{true} , $\ln(m^{\text{true}}/E^{\text{true}})$, $|\eta|$ histogram generated in the JMS calibration and closure test at given positions.

4 Calibration of reconstructed jet energy and mass

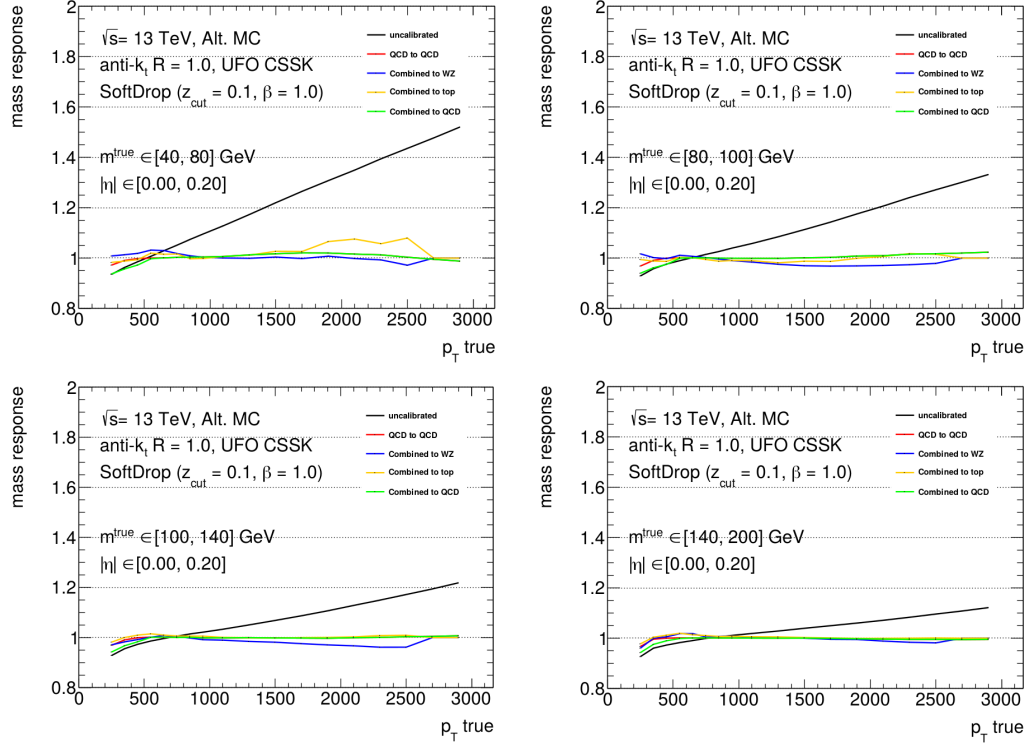


Figure 4.9: Closure test results of the combined JMS calibration being applied to jets originated from W/Z boson decays, top quark decays and QCD processes for UFO Soft Drop jet collection. The examples are in central $|\eta|$ bin, with four representative mass bins covering W boson mass, Z boson mass, Higgs boson mass and top quark mass.

$p_T < 500$ GeV in the calibration of UFO jets. The maximum non-closure can reach 10% at low p_T when apply JMS factors derived using QCD di-jet samples to W/Z jets. To reduce the non-closure in the calibration of W/Z jet mass, the final JMS factors in $p_T < 500$ GeV are replaced with JMS factors derived using W/Z jet samples. The rest of the JMS factors in $p_T > 500$ GeV are still derived using QCD di-jet samples to get sufficient statistics in a wide mass range. The combined JMS calibration has large non-closure when applied to QCD jets in $p_T < 500$ GeV, but the non-closure can be limited within 5% when applied to W/Z jets. The non-closure in the calibrated QCD jet mass can be tolerated because it does not have a physics hypothesis. Figure 4.9 shows the closure test results of the combined JMS calibration being applied to jets originated from W/Z boson decays, jets originate from top quark decays and QCD jets themselves. The combined calibration is able to calibrated W/Z jets and top jets mass with uncertainties limited within 5% in regions that are important to Run 2 physics analyses.

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

The boosted $VH(b\bar{b})$ analysis is developed to explore very high transverse momentum phase spaces, where the decay products of the Higgs boson can be reconstructed as a large- R jet. The vector boson decays leptonically to provide efficient rejection of the multi-jet background. The measurement is divided to three channels depending on the number of charged leptons in the final state: 0-lepton ($Z \rightarrow \nu\nu$), 1-lepton ($W \rightarrow l\nu$), 2-lepton ($Z \rightarrow ll$). The dominant backgrounds in the analysis are vector boson (W, Z) + jets production and top pair production. The other backgrounds are single top production, diboson production and multi-jet processes. Except the multi-jet background, all the signal and background processes are estimated using simulation. This chapter gives a detailed description of the boosted $VH(b\bar{b})$ analysis, including the data and simulated samples, the event selection and categorization, the statistical treatment and the final results.

5.1 Data samples and simulated event samples

5.1.1 Dataset and triggers

The data used in the analyses described in this thesis were collected by the ATLAS detector during full Run 2 (2015-2018) period of the LHC. The full Run 2 dataset corresponds to an integrated luminosity of 139 fb^{-1} at $\sqrt{s} = 13 \text{ TeV}$. The data events have to pass the Good Run List (GRL), which is a list of luminosity blocks when all systems of the ATLAS detector were functioning well.

The triggers used to select events in the boosted $VH(b\bar{b})$ analysis are E_T^{miss} trigger and single electron triggers. As described in Section 1.2.4, the analysis has three channels depending on the number of charged leptons in the final state: 0-lepton channel $ZH \rightarrow \nu\nu b\bar{b}$, 1-lepton channel $WH \rightarrow l\nu b\bar{b}$ and 2-lepton channel $ZH \rightarrow ll b\bar{b}$. The E_T^{miss} trigger is used in 0L channel and all the muon sub-channels. The single electron trigger is used in 1L and 2L electron sub-channels. The muon sub-channels can use E_T^{miss} trigger because muons are not included in the definition of E_T^{miss} at trigger level. The choice of E_T^{miss} trigger over single muon trigger in muon sub-channels is based on the higher efficiency of E_T^{miss} trigger in regions not covered by the muon trigger and regions with gaps between muon chambers.

The transverse momentum of Z boson in 0L channel is reconstructed as E_T^{miss} , since the Z boson decays to two neutrinos. The E_T^{miss} trigger used in data taking has an online threshold between 70 GeV and 110 GeV depending on the data taking period. The muon sub-channels use the same E_T^{miss} trigger. In electron sub-channels, the events have to pass the lowest p_T un-prescaled single electron trigger.

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

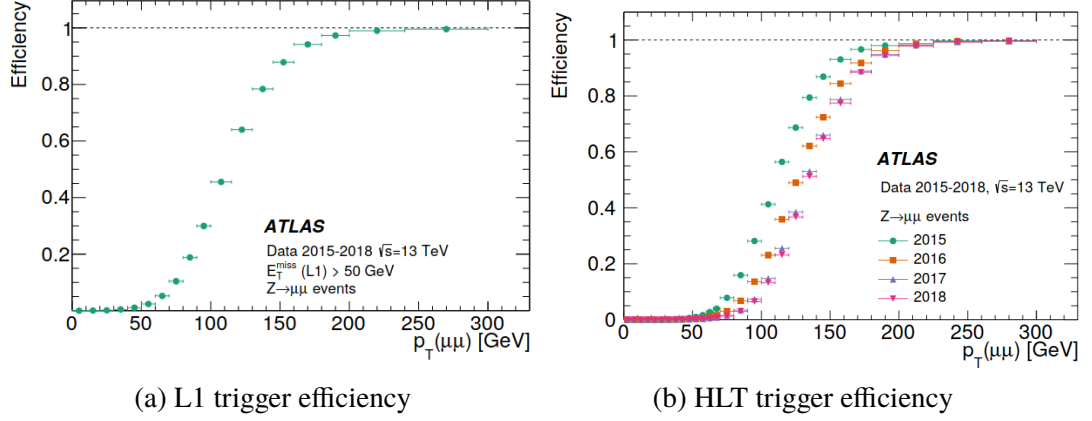


Figure 5.1: Measured L1 and HLT E_T^{miss} trigger efficiency for each year in Run 2, as a function of p_T of the di-muon system [119].

E_T^{miss} trigger

The E_T^{miss} trigger [119–122] is based on the momentum imbalance only in the calorimeter. Muons deposit very little energy in the calorimeter, so they are treated like neutrinos in the calculation. The trigger algorithm calculates the sum of E_T in all trigger towers at L1 level. An online threshold is applied to the towers: any towers with energy below this threshold is set to zero in the calculation. The L1 trigger can select events with $E_T^{\text{miss}} > 150$ GeV effectively.

The events selected by the L1 trigger are passed to the HLT. The HLT algorithm calculates E_T^{miss} from calibrated $R = 0.4$ jets. The pile-up removal and the calibration of jet energy have similar procedures as the calibration in off-line jet reconstruction. The value of E_T^{miss} is calculated as the negative transverse momentum vector sum of all jets with p_T above 7 GeV.

The trigger efficiency is defined as the fraction of events passing the trigger in the selected events measured in bins of E_T^{miss} . The E_T^{miss} trigger efficiency is measured in $Z \rightarrow \mu\mu$ samples. The off-line event selection requires the two muons with $p_T > 25$ GeV, and the invariant mass of the di-muon system between 66.6 GeV and 116.6 GeV. The trigger efficiency is measured as a function of p_T of the di-muon system. Figure 5.1 shows the efficiency of L1 E_T^{miss} trigger and the efficiency of HLT E_T^{miss} trigger. The efficiency reaches over 90% for $p_T > 150$ GeV.

Single electron trigger

The lowest p_T un-prescaled single electron trigger [123] is used to select events in all the electron sub-channels. The L1 electron trigger uses calorimeter information in the central region ($|\eta| < 2.5$) to build RoI. The RoIs are consisted of 2×2 trigger towers with granularity of 0.1×0.1 in η and ϕ . The energy of the trigger towers is calibrated at electromagnetic scale. A transverse energy threshold is applied to the nearest neighbor towers of the trigger tower. The E_T threshold can be η -dependent, it was around 20 GeV in Run 2.

5.1 Data samples and simulated event samples

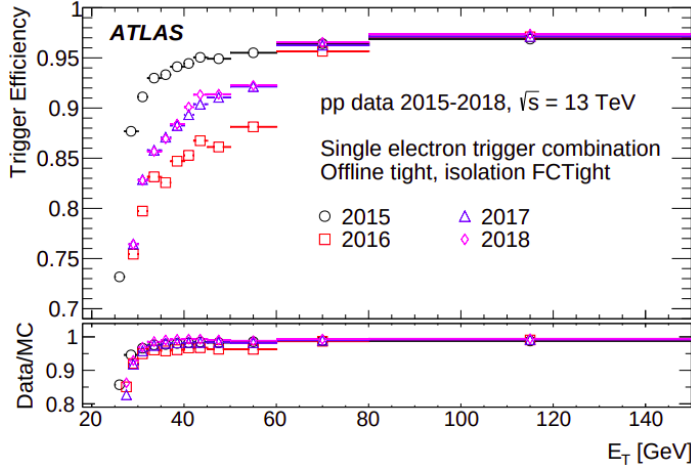


Figure 5.2: Evolution of the single electron trigger efficiency as a function of the offline electron E_T [123].

The HLT electron trigger reconstructs clusters from the calorimeter cells in regions slightly larger than the RoIs identified at L1. The electron candidates are identified through tracks matched to the corresponding clusters. The electron identification in trigger is designed to be as close as possible to the off-line electron identification used in analyses, with a few necessary differences. The identification algorithm relies on a multivariate analysis using likelihood method. Four efficiency working points are provided: ‘lhvloose’, ‘lhloose’, ‘lhmedium’, and ‘lhtight’. An optional isolation requirement that is similar to the electron isolation in the off-line reconstruction can also be applied in the trigger. The isolation requires the sum of p_T of all the quality tracks within a cone around the electron candidate smaller than 10% of the electron p_T . The nominal E_T threshold in the HLT triggers is 24 GeV in 2015 and 26 GeV in 2016, 2017 and 2018. Other thresholds at 60 GeV and 140 GeV (120 GeV in 2015) are also implemented.

The trigger efficiency is measured in an unbiased $Z \rightarrow ee$ sample selected using the tag-and-probe method. Figure 5.2 shows the single electron trigger efficiency in 2015-2018. The efficiency reaches over 95% for electrons with E_T larger than 60 GeV.

5.1.2 Simulated events

Simulated samples of VH signal and backgrounds are used to determine the expected signal and background distributions. The simulation is based on theoretical calculations using the SM predictions. Systematic uncertainties are also taken into consideration in the simulation. The signal processes in boosted $VH(b\bar{b})$ analysis include $qq \rightarrow ZH$, $qq \rightarrow WH$ and $gg \rightarrow ZH$. The major background processes are V +jets, top quark pairs, single top, diboson and QCD multi-jet. Except QCD multi-jet, all the other backgrounds are estimated using simulation. The QCD multi-jet process is estimated from data because the cross section of this process is very large and the rejection rate of this process is very high. The effects from pile-up are taken into account by overlaying the simulated samples with minimum bias events, which are simulated inelastic scattering events with as little selection bias as possible.

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

The simulation starts from generating physics process using MC event generators. At the generation level, the event contains information of the particles in the generated physics process. Then the actual simulation begins, it simulates the energy depositions of the particles in the ATLAS detector. Finally, the simulated events are digitized and reconstructed in the same way as the reconstruction of data events. The simulated samples for 2015 and 2016 data taking period are referred to as MC16a samples, 2017 data taking period are referred to as MC16d samples and 2018 data taking period are referred to as MC16e samples. The three data taking periods have the same generator configuration, but the pile-up configuration is different.

Several kinds of MC event generators are used in the simulation. Some are multi-purpose generators that are able to model a variety of processes. Others are specialized generators that are better at modeling specific processes. Every process have their best suited generator based on the strength of the generator. The generators used in the simulation for $VH(b\bar{b})$ analysis are

- POWHEG [124, 125]: An event generator that can generate the hard radiations using the exact next-to-leading order (NLO) matrix elements. It is often interfaced to other generators that simulate parton showers.
- PYTHIA [126]: A multi-purpose generator that can model all the components of the total cross section in hadronic collisions, including elastic, diffractive and non-diffractive topologies. It is often used to model parton showers, and the pile-up events.
- SHERPA [127, 128]: A general purpose event generator that is capable of simulating complete final states in high energy particle collisions. The physics processes that can be simulated with SHERPA cover all reactions in the SM.
- HERWIG [129, 130]: A general purpose event generator for simulation of high energy particle collisions with special emphasis on the accurate simulation of QCD radiation. It is not used as a stand alone generator in the simulation for $VH(b\bar{b})$ analysis, but many samples generated with PYTHIA are also generated with HERWIG to evaluate the modeling systematic uncertainties.
- MADGRAPH [131]: An event generator for automatically generating matrix elements for $2 \rightarrow n$ scatterings. It is not used as a stand alone generator in $VH(b\bar{b})$ analysis, but also for evaluating systematic uncertainties.

The simulated samples are generated with different combinations of the generators. The events are weighted to match the integrated luminosity of the data taking periods in Run 2.

Signal processes simulation

The VH signal processes of $VH(b\bar{b})$ analysis include $qq \rightarrow VH$ and $gg \rightarrow VH$. The leading order (LO) Feynman diagram of $qq \rightarrow VH$ is shown in Figure 5.3. Three kinds of processes are taken into account: $qq \rightarrow ZH \rightarrow \nu\bar{\nu}b\bar{b}$, $qq \rightarrow WH \rightarrow l\nu b\bar{b}$, $qq \rightarrow ZH \rightarrow llb\bar{b}$, where l can be e, μ, τ , corresponding to the main VH signal processes in 0L, 1L and 2L channel. The ZH signal can also be produced through $gg \rightarrow ZH$ processes. The LO Feynman diagrams are shown in Figure 5.4. The production is mediated by a top quark loop.

5.1 Data samples and simulated event samples

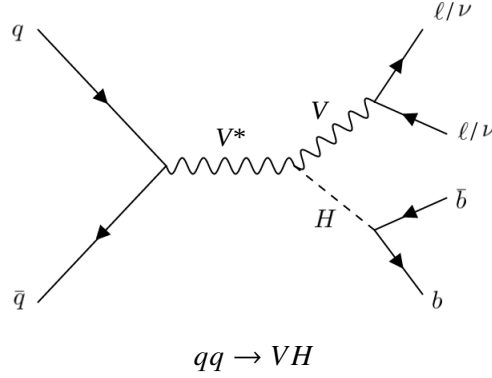


Figure 5.3: Leading order Feynman diagram of $qq \rightarrow VH$ processes.

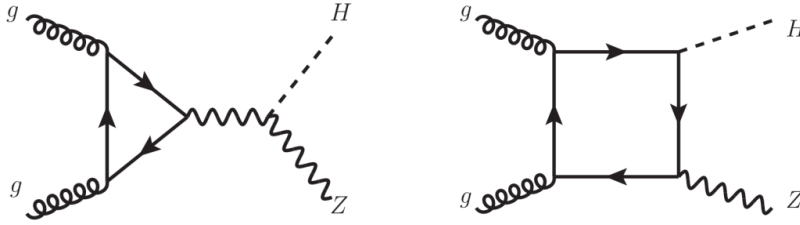


Figure 5.4: Leading order Feynman diagrams of $gg \rightarrow ZH$ processes.

The quark-induced VH signal samples are generated using POWHEG event generator with Multi-Scale Improved NLO (MiNLO) procedure [132], applying the AZNLO tune [133] with NNPDF3.0 [34] parton distribution function (PDF). The events are interfaced to PYTHIA 8 generator for modeling of the parton showers. The gluon-induced VH signal samples are generated using POWHEG event generator, also applying the AZNLO tune with NNPDF3.0 for the matrix element. PYTHIA 8 generator is used as interface to the event for modeling of parton showers. All the simulated samples are normalized to the best available theoretical prediction for the cross section where the Higgs boson mass is fixed to 125 GeV. The cross section of WH production is calculated at next-to-next-to-leading order (NNLO) for QCD processes and NLO for electroweak processes. The inclusive ZH production cross section is also calculated at NNLO for QCD and NLO for electroweak. The cross section of gluon-induced processes for ZH production is calculated at NLO, then the cross section of quark-induced processes is obtained from the difference between the inclusive ZH production cross section and the gluon-induced ZH production cross section to avoid double counting. The qg -initiated process has small contribution to the signal events. It is taken into account in qq -initiated production at NLO. The MC generators and cross sections of these signal processes are summarized in Table 5.1 [36].

Background processes simulation

Simulated background samples of the $VH(b\bar{b})$ analysis include vector boson (W , Z) + jets production, top quark pair production, single top production and diboson production. The production of V +jets is

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

Process	Generator	$\sigma \times \text{BR}$ [fb]
$qq \rightarrow ZH \rightarrow \nu\nu b\bar{b}$	POWHEG MiNLO + PYTHIA 8 (NNPDF3.0)	89.08
$qq \rightarrow WH \rightarrow l^+\nu b\bar{b}$	POWHEG MiNLO + PYTHIA 8 (NNPDF3.0)	164.6
$qq \rightarrow WH \rightarrow l^-\nu b\bar{b}$	POWHEG MiNLO + PYTHIA 8 (NNPDF3.0)	104.5
$qq \rightarrow ZH \rightarrow ll b\bar{b}$	POWHEG MiNLO + PYTHIA 8 (NNPDF3.0)	44.84
$gg \rightarrow ZH \rightarrow \nu\nu b\bar{b}$	POWHEG + PYTHIA 8 (NNPDF3.0)	14.30
$gg \rightarrow ZH \rightarrow l^+l^- b\bar{b}$	POWHEG + PYTHIA 8 (NNPDF3.0)	7.23

Table 5.1: MC event generators and cross sections of simulated VH signal processes used in Run 2.

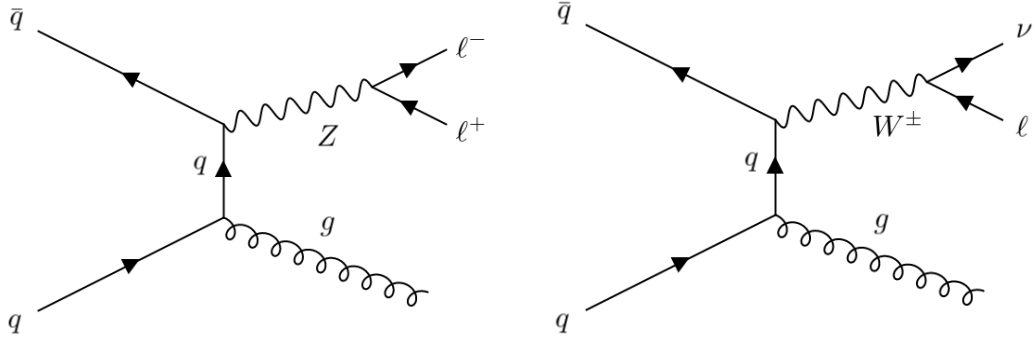


Figure 5.5: The lowest order Feynman diagrams for Z+jets and W+jets production at the LHC.

Process	Generator	$\sigma \times \text{BR}$ [pb]
V+jets ($Z \rightarrow \nu\nu$)	SHERPA 2.2.1	56280×0.200
V+jets ($W \rightarrow l\nu$)	SHERPA 2.2.1	183600×0.325
V+jets ($Z/\gamma^* \rightarrow ll$)	SHERPA 2.2.1	61940×0.101 ($m_{ll} > 40$ GeV)

Table 5.2: Event generators and cross section times branching ratio for simulated V+jets samples in Run 2.

the dominant background in all the sub-channels. It can be produced through a variety of processes. Figure 5.5 shows the lowest order Feynman diagrams for Z+jets and W+jets.

The production of V+jets is simulated with SHERPA 2.2.1 using NNPDF3.0 parton distribution functions for both the matrix element and the parton shower tuning. The samples are normalized using cross sections calculated at NNLO. The event generator and the SM predicted cross sections are summarized in Table 5.2.

The analysis aims to explore high p_T phase space. In order to generate sufficient statistics in high p_T regions, the simulation of V+jets samples are split in several $\max(H_T, p_T^V)$ regions, where H_T

5.1 Data samples and simulated event samples

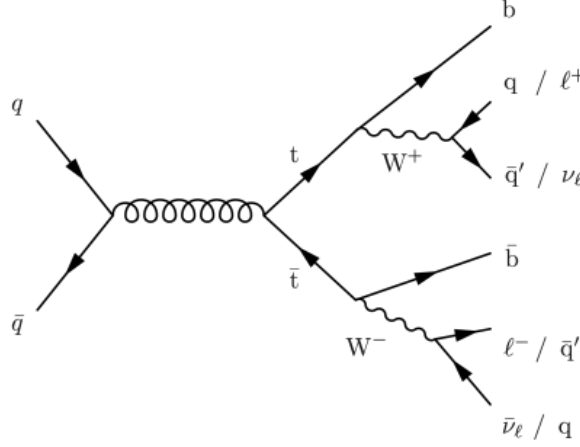


Figure 5.6: The lowest order Feynman diagram of top quark pair production at the LHC.

Process	Generator	$\sigma \times \text{BR}$ [pb]
$t\bar{t}$	POWHEG + PYTHIA 8	831.76
Single top		
s -channel	POWHEG + PYTHIA 8	10.32×0.325
t -channel	POWHEG + PYTHIA 8	216.96×0.325
Wt -channel	POWHEG + PYTHIA 8	71.7

Table 5.3: Event generator and cross section times branching ratio used in simulation of $t\bar{t}$ and single top production in Run 2.

is defined as the scalar p_T sum of all parton level jets with $p_T > 20$ GeV. The V +jets samples are produced in slices of $\max(H_T, p_T^V)$:

$$[0 - 70, 70 - 140, 140 - 280, 280 - 500, 500 - 1000, > 1000] \text{ GeV}. \quad (5.1)$$

Additionally, the analysis requires two b -tagged jets in the final states. Several flavor filters are applied to provide sufficient heavy flavor events. The filters require at least one b -hadron or c -hadron. The dominant V +jets background is a vector boson associated with heavy flavor (HF) jets, including bb , bc , bl , cc .

The other dominant background in 0L channel and 1L channel is the top quark pair production. The lowest order Feynman diagram of top quark pair production is shown in Figure 5.13. The $t\bar{t}$ samples are generated with POWHEG using NNPDF 3.0 for matrix element. The events are interfaced to PYTHIA 8 using NNPDF 2.3 parton distribution functions for parton showering. The samples are normalized using cross sections calculated at NNLO accuracy, including resummation of next-to-next-to-leading logarithmic (NNLL) soft gluon terms [134]. The generator and cross section of $t\bar{t}$ simulation is shown in Table 5.3. A Filter that requires at least one of the W boson decays leptonically is applied in the simulation to remove events that will be rejected in event selection of the analysis. E_T^{miss} filter

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

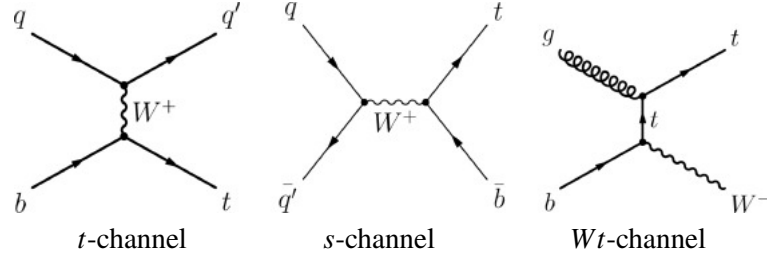


Figure 5.7: The lowest order Feynman diagrams of single top production at the LHC.

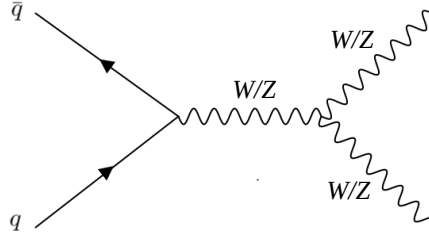


Figure 5.8: Example Feynman diagram of the lowest order $qq \rightarrow VV$ processes at the LHC.

is also used in the generation of di-lepton $t\bar{t}$ samples to remove events with less than two W bosons decay leptonically and provide sufficient statistics for the $VH(b\bar{b})$ analysis.

Single top quark production has three channels: s -channel, t -channel and Wt -channel. Figure 5.7 shows the lowest order Feynman diagrams of these single top quark processes. The single top background events enter the signal regions of $VH(b\bar{b})$ analysis due to the semileptonic decay of the top. The samples are generated using POWHEG with NNPDF 3.0 as the matrix element, and interfaced to PYTHIA 8 using NNPDF 2.3 parton distribution functions for parton showering. The single top samples are normalized using cross section calculated at NLO. A lepton filter is applied in generation of s -channel and t -channel single top samples to require the W boson from top decays leptonically. The samples of Wt -channel has no filter or a di-lepton filter applied to require both of the W boson decays leptonically. A dedicated removal is set to avoid the double counting between the two kinds of samples in Wt -channel.

Semi-leptonic diboson production has a variety of processes, including $qq \rightarrow WW$, $qq \rightarrow WZ$, $qq \rightarrow ZZ$, $gg \rightarrow WW$ and $gg \rightarrow ZZ$. One of the vector boson decays leptonically, the other one decays hadronically. Figure 5.8 is an example of the lowest order Feynman diagram of $qq \rightarrow VV$ process. The diboson samples are generated using SHERPA 2.2.1 with NNPDF 3.0 parton distribution functions for matrix element and parton showering. Samples for quark induced diboson with less than two jets are normalized using cross section calculated at NLO accuracy. Samples for diboson with two jets or three jets are produced to LO accuracy. Additionally, loop induced $gg \rightarrow VV$ samples are produced using LO accurate matrix element for final states with less than two jets. The event generator and cross section used in simulation of diboson samples are shown in Table 5.4.

5.1 Data samples and simulated event samples

Process		Generator	$\sigma \times \text{BR}$ [pb]
$qq \rightarrow WW$	$\rightarrow qq\ell\nu$	SHERPA 2.2.1	112.6×0.439
$qq \rightarrow WZ$	$\rightarrow \ell\nu qq$	SHERPA 2.2.1	50.3×0.227
	$\rightarrow qq\nu\nu$	SHERPA 2.2.1	50.3×0.135
	$\rightarrow qq\ell\ell$	SHERPA 2.2.1	50.3×0.0683
$qq \rightarrow ZZ$	$\rightarrow qq\ell\ell$	SHERPA 2.2.1	15.57×0.140
	$\rightarrow qq\nu\nu$	SHERPA 2.2.1	15.57×0.280
$gg \rightarrow WW$	$\rightarrow qq\ell\nu$	SHERPA 2.2.2	4.8×0.439
$gg \rightarrow ZZ$	$\rightarrow qq\ell\ell$ or $qq\nu\nu$	SHERPA 2.2.2	1.57×0.420

Table 5.4: Event generator and cross section times branching ratio used in simulation of diboson samples in Run 2.

5.2 Event Selection

The event selection of $VH(b\bar{b})$ analysis is to select regions with optimized signal background ratio, the so-called signal region. Regions dominated by a kind of background can also be selected to provide constraints on background events in the fitting, which is denoted as the control region. $VH(b\bar{b})$ boosted analysis has 10 signal regions and 4 top-enriched control regions defined. The development and optimization of the event selection were conducted without using the actual data to avoid bias in the selection, known as the "blinded analysis" policy. This section describes the baseline event selection, the tagging of Higgs candidates and the event categorization.

5.2.1 Higgs boson identification

The Higgs bosons decayed to b quark pairs are considered in the $VH(b\bar{b})$ analysis. To identify events with $H \rightarrow b\bar{b}$ decay, it is necessary to identify jets containing b -hadron decay products. The angular distance between the two b -jets originate from the Higgs boson decay depends on the Higgs boson p_T . Figure 5.9 is an illustration of the two b -jets from the decay of Higgs boson with low and high p_T . The ΔR between the two b -jets is approximately $2m_H/p_T$, where $m_H = 125$ GeV. The two b -jets can be separated clearly when p_T is not very large. When p_T increases to 250 GeV, the two b -jets start to merge into a large- R jet, which is the boosted topology. Therefore, the decay products of the Higgs boson is reconstructed as a large- R jet, the b -jets are recovered using variable radius (VR) track jets. The selection of the Higgs boson candidates in $VH(b\bar{b})$ boosted analysis requires at least two track jets matched to the leading large- R jet, and the two leading track jets have to be b -tagged. The selection is further optimized in $VH(b\bar{b})$ legacy analysis, introduced in Section 5.6.

B-tagging

The algorithm used to identify track jets containing b -hadron decay products in $VH(b\bar{b})$ boosted analysis is a multivariate method MV2c10 [135, 136]. It is upgraded to a combined b and c -tagger in the $VH(b\bar{b})$ legacy analysis. This section describes the MV2c10 b -tagging algorithm.

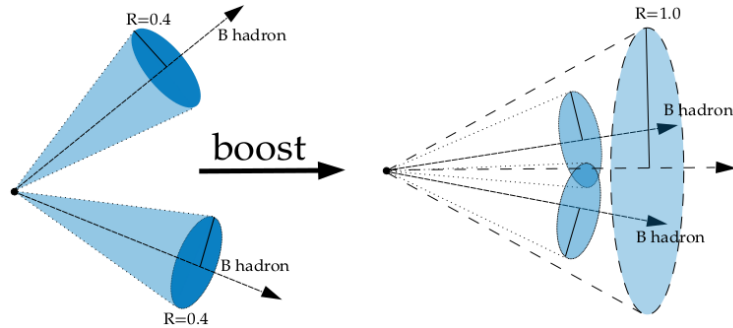


Figure 5.9: Illustration of b -jets from the decay of Higgs boson.

The b -hadrons have a lifetime of the order of 1.5 ps, corresponding to a significant mean flight length in the detector ($\langle l \rangle = \beta\gamma c\tau$) before decaying. This feature leads to at least one vertex displaced from the hard scattered interaction point. The b -tagging algorithms make use of the displaced vertex to identify jets originate from b -hadron. The algorithms can be divided into low-level and high-level. The low-level algorithms reconstruct the charged particle tracks associated with a hadronic jet, then the algorithms combine the tracks to explicitly reconstruct displaced vertices. The high-level algorithms combine results from the low-level algorithms to get the maximum performance.

The tracks are matched to jets using the angular distance ΔR between the track momenta and the jet axis. The ΔR requirement varies as a function of jet p_T , being wider for lower jet p_T , and narrower for higher jet p_T . The closest jet would be chosen if more than one jets can be matched to a track. These low-level algorithms are IP2D and IP3D [136], which are taggers based on a log-likelihood ratio discriminant separating tracks associated to jets. The IP2D tagger uses the signed transverse impact parameter significance d_0/σ_{d_0} as the discriminating variable. The IP3D tagger uses both the signed transverse impact parameter significance and the longitudinal impact parameter significance $z_0 \sin \theta / \sigma_{z_0 \sin \theta}$, in a two dimensional template to account for their correlation.

The secondary vertex finding algorithm SV1 [137] reconstructs a single displaced vertex in a jet. The SV1 algorithm identifies all the possible vertices with at least two tracks. All the tracks associated to the jet are taken into account, except tracks originate from long-lived particles decay, photon conversions or hadronic interactions with detector material. The track to vertex association is evaluated iteratively using a χ^2 test. In each iteration, the track with the largest χ^2 is removed and the fitting of vertex is repeated, until a vertex with acceptable χ^2 in the fit and an invariant mass less than 6 GeV is found.

To get maximum performance from the b -tagging algorithms, the results obtained from these low-level algorithms are combined using multivariate classifiers. The MV2 algorithm, used in the b -tagging of $VH(b\bar{b})$ boosted analysis, is based on a boosted decision tree (BDT). The BDT is trained using simulated $t\bar{t}$ and Z' samples containing b -jets. The BDT hyperparameters have been optimized to provide the best separation power between b -jets and c -jets or light jets. The fraction of b -jets, c -jets and light jets as a function of the MV2 discriminant is shown in Figure 5.10.

The b -tagging efficiency is defined as the fraction of b -jets passing the selection. For MV2 tagger, the selection is a simple fixed cut on the discriminant. The b -tagging discriminant distribution is divided into five bins, with the boundaries defined at b -tagging efficiency working points (WP) for 85%, 77%, 70%, 60%, bounded by the trivial 100% and 0% selections. The background jets rejection rates for different b -tagging WPs are shown in Table 5.5. The b -tagging algorithm can be applied to small- R jets and track jets. The tagging of track jets originate from b -hadrons is essential to the $VH(b\bar{b})$ boosted analysis.

Selection of Higgs boson candidate

As described in previous sections, the decay products of the Higgs boson are reconstructed as large- R jets. Therefore, the events in all the channels are required to have at least one large- R jets. The large- R jet with the highest p_T , denoted as the leading large- R jet, is used when more than one large- R jets exist. The events are required to have at least two VR track jets matched to the leading large- R jet.

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

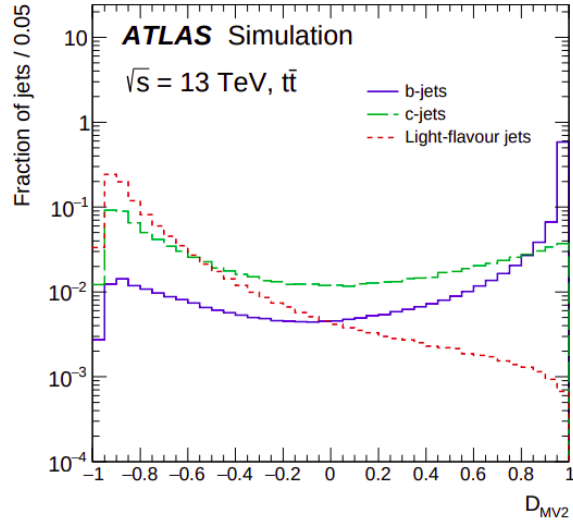


Figure 5.10: Distribution of the MV2 discriminant for b -jets, c -jets and light jets evaluated from simulated $t\bar{t}$ samples [135].

Efficiency	Cut value	Rejections		
		c -jet	τ -jet	light jet
60%	> 0.94	23	140	1200
70%	> 0.83	8.9	36	300
77%	> 0.64	4.9	15	110
85%	> 0.11	2.7	6.1	25

Table 5.5: Background jets rejection rates for different b -tagging efficiency working points, computed from simulated $t\bar{t}$ samples [135].

5.2 Event Selection

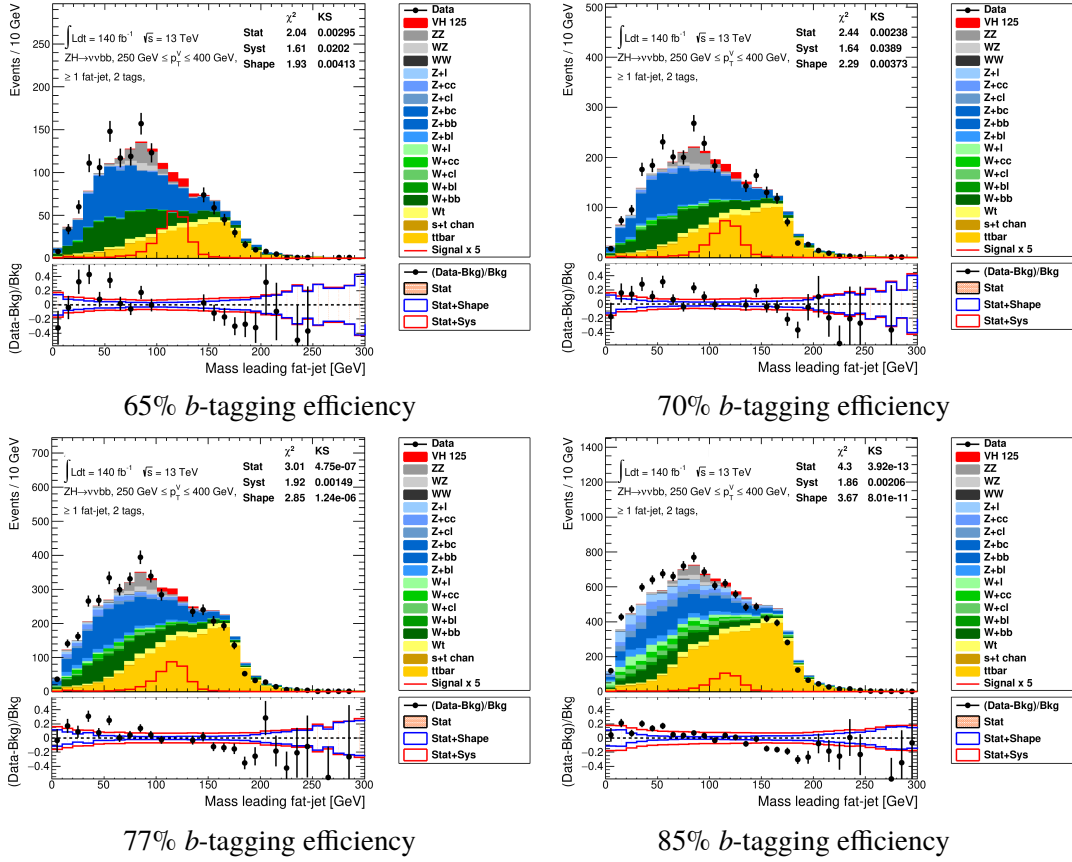


Figure 5.11: Distributions of m_J in the inclusive signal region in $250 \text{ GeV} < p_T^V < 400 \text{ GeV}$ for b -tagging efficiency WP: 60%, 70%, 77%, 85%. The distributions are plotted in 0L channel. Data points in bins with more than 5% signal events are not shown to follow the "blinded analysis" policy at this stage.

The matching process uses a "ghost-association" technique [138]. The matching is done with respect to the ungroomed large- R jets, mimicked by adding soft "ghost" constituents with infinitely small momentum to the jet. The ghosts do not affect the reconstruct jet because the transverse momentum is at least an order smaller than the jet constituents. The matching is integrated into the reconstruction of the jets, the track jets have to be contained in the large- R jet areas with the presence of the ghost constituents.

The b -tagging is applied to all the track jets. The two leading track jets matched to the leading large- R jet have to be b -tagged. The four b -tagging efficiency WPs have been studied and the one with the best expected significance is selected. Figure 5.11 shows the m_J distributions in the inclusive signal region (SR) in $250 \text{ GeV} < p_T^V < 400 \text{ GeV}$ for different b -tagging efficiency WPs, with the baseline event selections applied. The distributions for WPs at 60%, 70% and 77% are similar, while the 85% WP results in a significant increase of the V +jets background. The performance of these different efficiencies is evaluated through expected significance from fitting using m_J as the final discriminant, with flavor tagging systematics considered; and also through $s/\sqrt{b}$, where s and b denote the number of signal and background events in the inclusive SR in $p_T^V > 250 \text{ GeV}$ and $100 \text{ GeV} < m_J < 140 \text{ GeV}$.

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

b -tagging WP	60%	70%	77%	85%
$s/\sqrt{b}$	2.02	2.09	2.04	1.77
Expected significance	1.67	1.73	1.69	1.43

Table 5.6: Significance and $s/\sqrt{b}$ in 0L channel for b -tagging efficiency WP: 60%, 70%, 77%, 85%. The significance is calculated from fitting using m_J distributions, with flavor tagging systematics.

The expected significance is the significance calculated from fitting with a representative dataset created from the simulated signal and background samples, without using the actual dataset. Table 5.6 presents the significance and $s/\sqrt{b}$ for all the b -tagging efficiency WPs. The b -tagging efficiency at 70% has the largest expected significance and $s/\sqrt{b}$ among all the tested WPs. It is selected as the optimal b -tagging efficiency WP for $VH(b\bar{b})$ boosted analysis. Events with no b -tagged track jets, one b -tagged track jet and two b -tagged track jets compose the 0-tag category, 1-tag category and the 2-tag category. Only events in the 2-tag category are considered in the fitting as these regions have the largest sensitivity.

5.2.2 Baseline event selection

The baseline event selection is applied to data and simulated samples after the events passed the quality criteria, including the GoodRunList (data only), number of primary vertex and bad jet rejection. Then the events are selected based on the number of charged leptons in the final state. The analysis has three sub-channels: 0-lepton, 1-lepton and 2-lepton channel. The event selection is designed for each sub-channel based on their kinematic features. The trigger requirements are applied as described in Section 5.1.1. The vector boson transverse momentum p_T^V is required to be larger than 250 GeV in the stand alone $VH(b\bar{b})$ boosted analysis. The events are required to have at least one large- R jet with $p_T > 250$ GeV, $|\eta| < 2$ and $m_J > 50$ GeV. At least two track jets with $p_T > 10$ GeV, $|\eta| < 2.5$ need to match to the leading large- R jet, and both of the two leading track jets are b -tagged, as described in the Section 5.2.1. All the baseline event selection requirements are listed in Table 5.7.

0-lepton channel

The baseline event selection in 0-lepton (0L) channel requires zero VH -loose lepton, which is electron identified at Loose working point with Loose isolation requirement or muon identified at Loose working point with fixed cut Loose isolation requirement. The kinematic requirement on VH -loose electrons are $p_T > 7$ GeV, $|\eta| < 2.47$, $|d_0/\sigma(d_0)| < 5$ and $|z_0 \sin \theta| < 0.5$ mm. VH -loose muons are required to pass $p_T > 7$ GeV, $|\eta| < 2.7$, $|d_0/\sigma(d_0)| < 3$ and $|z_0 \sin \theta| < 0.5$ mm.

In addition to the Higgs tagging selection and p_T^V cut, 0L channel has several cuts to suppress multijet background. The multijet background enters the signal region due to jet energy mis-measurement. As a result, the E_T^{miss} is more likely to be aligned with the mis-measured jet. The cuts used to suppress multi-jet background are:

5.2 Event Selection

Selection	0L channel	1L channel		2L channel	
		e channel	μ channel	e channel	μ channel
Trigger	$E_{\text{T}}^{\text{miss}}$	single electron	$E_{\text{T}}^{\text{miss}}$	single electron	$E_{\text{T}}^{\text{miss}}$
Leptons	0 VH -loose lepton	1 WH -signal lepton no second VH -loose lepton $p_{\text{T}} > 27$ GeV $p_{\text{T}} > 25$ GeV		2 VH -loose leptons at least 1 ZH -signal lepton $p_{\text{T}} > 27$ GeV - opposite sign	
p_{T}^V	$p_{\text{T}}^V > 250$ GeV				
Large- R jet	at least 1 large- R jet, $p_{\text{T}} > 250$ GeV, $ \eta < 2.0$				
Track jet	at least two track-jets, $p_{\text{T}} > 10$ GeV, $ \eta < 2.5$, matched to the leading large- R jet				
b -tagging	leading two track jets matched to the leading large- R jet have to be b -tagged				
m_{J}	> 50 GeV				
$E_{\text{T}}^{\text{miss}}$	same as p_{T}^V	> 50 GeV	-	-	
$\min[\Delta\phi(E_{\text{T}}^{\text{miss}}, \text{small-}R \text{ jets})]$	$> 30^\circ$	-			
$\Delta\phi(E_{\text{T}}^{\text{miss}}, \text{large-}R \text{ jets})$	$> 120^\circ$	-			
$\Delta\phi(E_{\text{T}}^{\text{miss}}, \mathbf{p}_{\text{T}}^{\text{miss}})$	$< 90^\circ$	-			
$\Delta Y(V, \text{large-}R \text{ jet})$	-	$ \Delta Y(V, \text{large-}R \text{ jet}) < 1.4$			
m_{ll}	-			$66 \text{ GeV} < m_{ll} < 116 \text{ GeV}$	
Lepton p_{T} imbalance	-			$(p_{\text{T}}^{l_1} - p_{\text{T}}^{l_2})/p_{\text{T}}^Z < 0.8$	

Table 5.7: Baseline event selection requirements in all the sub-channels of boosted $VH(b\bar{b})$ analysis.

- $\min[\Delta\phi(E_T^{\text{miss}}, \text{small-}R \text{ jets})] > 30^\circ$, where small- R jets refer to $R = 0.4$ jets with $p_T > 70$ GeV, $|\eta| < 4.5$ not matched to the leading large- R jet.
- $\Delta\phi(E_T^{\text{miss}}, \text{large-}R \text{ jets}) > 120^\circ$.
- $\Delta\phi(E_T^{\text{miss}}, \mathbf{p}_T^{\text{miss}}) < 90^\circ$, where $\mathbf{p}_T^{\text{miss}}$ is the missing transverse momentum reconstructed from tracks.

These cuts are fine tuned to reduce the multijet background to a negligible level in 0L channel.

1-lepton channel

The event selection in 1-lepton (1L) channel requires exactly one WH -signal lepton. The WH -signal electron is identified at Tight working point with the calorimeter-based HighPtCaloOnly isolation requirement: $E_T^{\text{cone0.2}} < \max(0.015p_T, 3.5 \text{ GeV})$. The electron is required to have $p_T > 27$ GeV, $|\eta| < 2.47$, $|d_0/\sigma(d_0)| < 5$ and $|z_0 \sin \theta| < 0.5$ mm. The WH -signal muon is identified at Medium working point with the track-based HighPtTrackOnly isolation requirement: $p_T^{\text{cone0.2}} < 1.25$ GeV.

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

The kinematic requirements of WH -signal muons are $p_T > 25$ GeV, $|\eta| < 2.5$, $|d_0/\sigma(d_0)| < 3$ and $|z_0 \sin \theta| < 0.5$ mm. A veto on any additional VH -loose lepton is applied.

Except the common selections in all the sub-channels, a cut requires $E_T^{\text{miss}} > 50$ GeV is applied in electron sub-channel to suppress the multijet background. The remaining multijet background in 1L channel is estimated from data. The absolute difference in rapidity between the vector boson and the large- R jet $|\Delta Y(V, \text{large-}R \text{ jet})|$, is able to further discriminate between signal and backgrounds. The optimized value of this cut is $|\Delta Y(V, \text{large-}R \text{ jet})| < 1.4$. The z -component of the neutrino momentum used to calculate the rapidity of W boson is obtained from the W -mass constraint that minimizes the difference in longitudinal boosts of the reconstructed W boson and the large- R jet, $|\beta_W^z - \beta_H^z|$, where β^z is defined as $p_z/\sqrt{p_z^2 + m^2}$.

2-lepton channel

The event selection in 2-lepton (2L) channel requires exactly two same flavor VH -loose leptons, and at least one of the two have to be ZH -signal lepton. The VH -loose lepton requirement is the same as the requirement in 0L channel. The ZH -signal lepton is electron identified at Loose working point with Loose isolation requirement or muon identified at Loose working point with fixed cut Loose isolation requirement. The electrons are required to have $p_T > 27$ GeV, $|\eta| < 2.47$, $|d_0/\sigma(d_0)| < 5$ and $|z_0 \sin \theta| < 0.5$ mm. The muon selection requirements are $p_T > 27$ GeV, $|\eta| < 2.5$, $|d_0/\sigma(d_0)| < 3$ and $|z_0 \sin \theta| < 0.5$ mm. The selected muons are required to have opposite charge, while the selected electrons are not required due to higher rate of electron charge mis-identification.

Several additional cuts are applied in 2L channel other than the p_T^V cut and the Higgs tagging requirements. A Z mass window cut is applied to the invariant mass of the di-lepton system: $66 \text{ GeV} < m_{ll} < 116 \text{ GeV}$. The cut $\Delta Y < 1.4$ found useful in 1L channel can also further discriminate between signal and backgrounds in 2L channel. This cut can suppress Z +jets background, which is the dominate background in 2L channel. The other cut that is used to suppress Z +jets background is $(p_T^{l_1} - p_T^{l_2})/p_T^Z < 0.8$, where $p_T^{l_1}$ and $p_T^{l_2}$ are the lepton transverse momentum. This variable is found to be sensitive to Z boson polarization, which is different between ZH and Z +jets processes.

5.2.3 Event categorization

After applying the baseline event selection and the Higgs tagging selection, the events are categorized into several signal regions and control regions. The categorization is based on p_T^V , number of small- R jets not matched to the leading large- R jet and number of b -tagged track jets not matched to the leading large- R jet.

p_T^V splitting

The splitting based on vector boson transverse momentum is common for all three sub-channels. The p_T^V threshold for $VH(b\bar{b})$ boosted analysis is 250 GeV. The regions in $p_T^V > 250$ GeV are split to medium p_T^V ($250 \text{ GeV} < p_T^V < 400 \text{ GeV}$) and high p_T^V ($p_T^V > 400 \text{ GeV}$) regions. The p_T^V splitting is

5.2 Event Selection

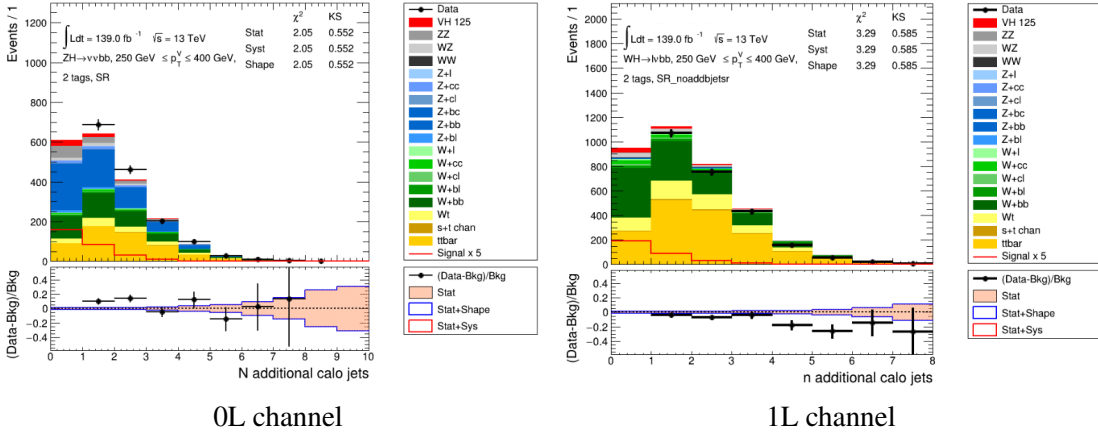


Figure 5.12: Number of small- R jets not matched to the leading large- R jet in 0L channel and 1L channel. Data points in bins with more than 5% signal events are not shown to follow the "blinded analysis" policy at this stage.

based on the fact that phase spaces with higher transverse momentum have higher signal-to-background ratio. The higher signal-to-background ratio improves the significance of the fitting. Additionally, the higher transverse momentum region can be more sensitive to potential BSM effects.

Signal region categorization

The signal regions (SR) in 0L channel and 1L channel are required to have no b -tagged track jets outside the leading large- R jet to enhance the top background rejection, while the signal regions in 2L channel do not have this requirement because 2L channel has very little top background.

The SRs in 0L channel and 1L channel are further categorized based on the number of small- R jet not matched to the leading large- R jet, denoted as additional calo jets. The small- R jets are matched to the large- R jet if $\Delta R < 1.0$. Figure 5.12 shows the distributions of number of additional calo jets in 0L channel and 1L channel. A high purity signal region (HP SR) and a low purity signal region (LP SR) are defined in each signal region using a cut on number of additional calo jets. The HP SR has zero additional calo jet, while the LP SR has at least one additional calo jets. This splitting is able to suppress $t\bar{t}$ background significantly, especially in 1L channel as it has the most $t\bar{t}$ background. In 2L channel, the splitting is not applied because the $t\bar{t}$ background is very minimal.

Control region categorization

The $t\bar{t}$ background is a dominant background in 1L channel and 0L channel. This process usually has an additional b -jet not matched to the leading large- R jet, as shown in Figure 5.13. The W boson that decays hadronically together with a b -jet could be identified as a large- R jet. The other W boson decays leptonically, so the event could be mis-identified as a Higgs candidate. The additional b -jet can be used to further discriminant between $t\bar{t}$ process and signal.

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

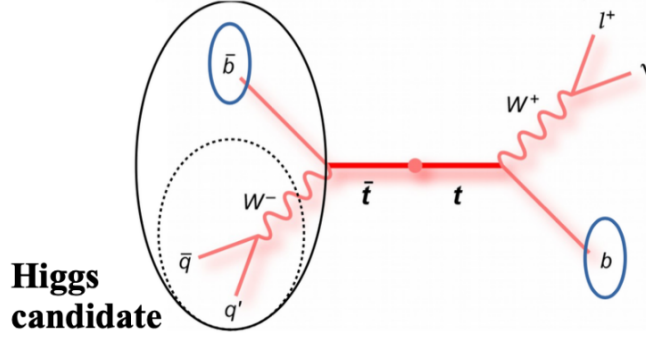


Figure 5.13: Illustration of the kinematic features of the $t\bar{t}$ process.

Channel	Categories					
	$250 \text{ GeV} < p_T^V < 400 \text{ GeV}$			$p_T^V > 400 \text{ GeV}$		
	0 add. b -tagged track-jet		≥ 1 add. b -tagged track-jets	0 add. b -tagged track-jet		≥ 1 add. b -tagged track-jets
	0 add. small- R jet	≥ 1 add. small- R jets		0 add. small- R jet	≥ 1 add. small- R jets	
0L	HP SR	LP SR	top CR	HP SR	LP SR	top CR
1L	HP SR	LP SR	top CR	HP SR	LP SR	top CR
2L	SR			SR		

Table 5.8: Summary of the signal regions and top control regions defined in 0L, 1L and 2L channel.

As described in previous sections, the signal regions in 0L and 1L channels require zero additional b -tagged track jet. While for control regions (CR) dominated by top backgrounds, the requirement can be inverted. The top control regions in 0L and 1L channels are defined with the requirement of at least one additional b -tagged track jet. The same b -tagging algorithm is applied to all the track jets including those not matched to the large- R jet through ghost association.

The control regions are able to constrain the $t\bar{t}$ background, which is one of the dominant backgrounds. There are ten SRs and the four top CRs defined in total. A summary of all the SRs and CRs is shown in Table 5.8.

5.2.4 Large- R jet mass

The large- R jet mass m_j is used as the final discriminant in the analysis. In order to improve the mass resolution and sensitivity, two corrections are applied to the large- R jet mass. The first one is the muon-in-jet correction. The semileptonic decays of the b -hadrons could result in muons in the final states, while the muons deposite little energy in the calorimeter. These muons can be reconstructed.

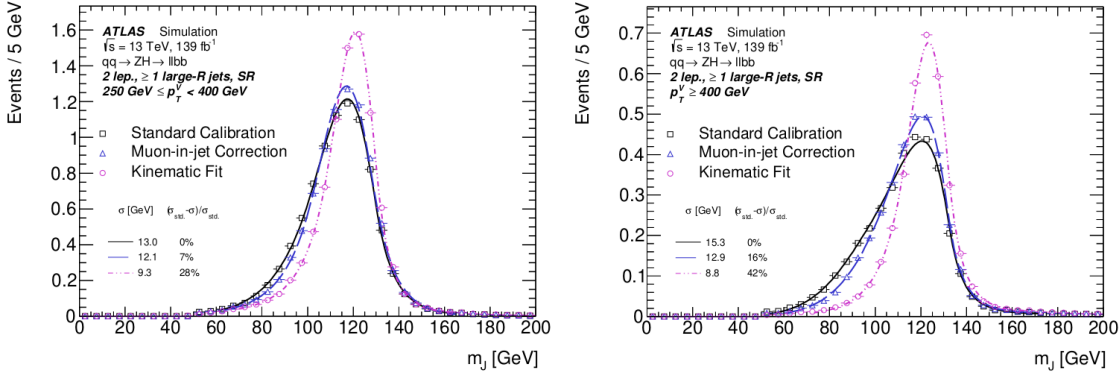


Figure 5.14: Distributions of large- R jet mass in 2L channel with muon-in-jet correction and kinematic fit applied [140].

The muon-in-jet correction [139] is applied to recover the energy carried by the reconstructed muons. The muons are identified at Medium working point and required to have $p_T > 10$ GeV. Muons that can be matched to the track jets and ghost associated to the large- R jet are used in the correction. The calorimeter-based large- R jet mass is corrected by adding the muon four-momentum, and subtracting the muon energy deposits in the calorimeter. The track-assisted mass is then corrected using the corrected calorimeter-based mass. Finally, the combined mass is corrected using both the updated calorimeter-based mass and the track-assisted mass.

The second correction is the kinematic fit applied in 2L channel. The decay products in 2L channel are completely reconstructed, so the transverse momentum of the di-lepton system can be used to improve the jet energy resolution. The jet energy resolution is typically 10% level, while the resolution of electrons and muons is typically 1% level. The kinematic fit constraints the large- R jet and the di-lepton system to be balanced in the transverse plane, then the large- R jet mass is scaled by the ratio of the fitted jet p_T and the original jet p_T . The kinematic fit also includes a constraint on the di-lepton system invariant mass to the Z boson mass.

The distributions of m_J for ZH signal events in 2L channel are shown in Figure 5.14. The muon-in-jet correction is able to improve the jet mass resolution by 5% - 10% depending on the channel. The kinematic fit brings an additional improvement by up to 40% in 2L channel [140].

5.2.5 Jet substructure variables

The large- R jet substructure information can be used to further discriminate between VH signal and background. The substructure variables commonly used in W/Z tagging are C_2 and D_2 [141]. These variables can also improve the event selection of $VH(b\bar{b})$ boosted analysis. The substructure variable C_2 and D_2 are defined from energy correlation functions (ECF). The N -point ECF is defined as:

$$\text{ECF}(N, \beta) = \sum_{i_1 < i_2 < \dots < i_N \in J} \left(\prod_{a=1}^N p_{Ti_a} \right) \left(\prod_{b=1}^{N-1} \prod_{c=b+1}^N R_{i_b i_c} \right)^\beta, \quad (5.2)$$

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

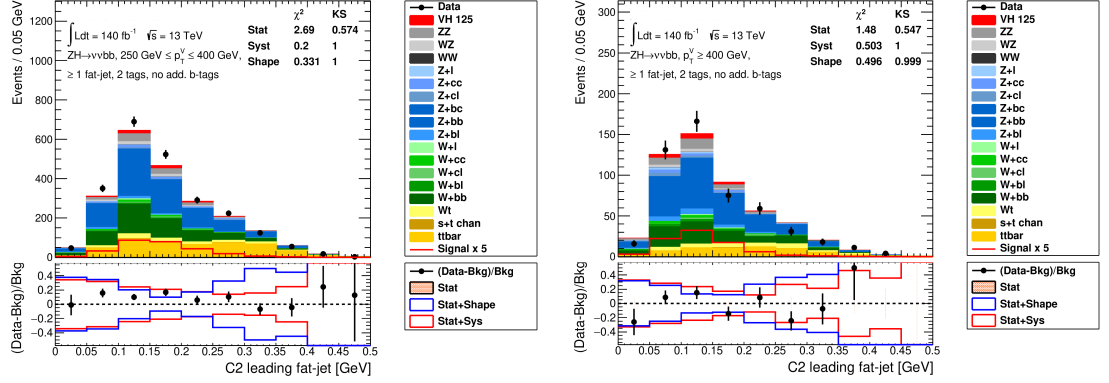


Figure 5.15: Distributions of C_2 in inclusive medium p_T^V SR and inclusive high p_T^V SR in 0L channel.

where the sum runs over all the constituents within the large- R jet, β is a fixed value parameter, $R_{i_b i_c}$ is the distance between i_b and i_c in the rapidity-azimuth angle plane. To define C_2 and D_2 , 3-point correlation functions are considered:

$$\begin{aligned} \text{ECF}(1, \beta) &= \sum_{i \in J} p_{Ti}, \\ \text{ECF}(2, \beta) &= \sum_{i < j \in J} p_{Ti} p_{Tj} (R_{ij})^\beta, \\ \text{ECF}(3, \beta) &= \sum_{i < j < k \in J} p_{Ti} p_{Tj} p_{Tk} (R_{ij} R_{ik} R_{jk})^\beta. \end{aligned}$$

By definition, a jet with fewer than N constituents has $\text{ECF}(N, \beta) = 0$. More generally, if a system has N sub-jets, then $\text{ECF}(N+1, \beta)$ should be significantly smaller than $\text{ECF}(N, \beta)$. The dimensionless large- R jet substructure variables are defined from the ECFs at $\beta = 1$ based on this feature.

$$\begin{aligned} C_2(\beta = 1) &= \frac{\text{ECF}(3) \times \text{ECF}(1)}{\text{ECF}(2)^2}, \\ D_2(\beta = 1) &= \frac{\text{ECF}(3) \times \text{ECF}(1)^3}{\text{ECF}(2)^3}. \end{aligned}$$

The values of C_2 and D_2 for large- R jets with two sub-jets are more likely to be smaller than large- R jets with more than two sub-jets. VH signal jets have two sub-jets, therefore a cut on C_2 and D_2 can improve the significance of the analysis.

Distributions of C_2 and D_2 in the inclusive SRs in 0L channel are shown in Figure 5.15 and Figure 5.16. As expected, signal events have smaller C_2 and D_2 values compared to background events in general. A scan is performed to decide the best cut point that maximize the expected significance for each substructure variable. The scan is first performed in wide ranges to decide the regions containing the optimal cut point, then a detailed scan in that region is performed to decide the cut point. The performance of the cuts is evaluated from the fit considering only statistical uncertainties first and then including systematic uncertainties. The result of the detailed scan for C_2 is summarized in Table 5.9 and Table 5.10. The same procedure is also applied to D_2 and other substructure variables.

5.2 Event Selection

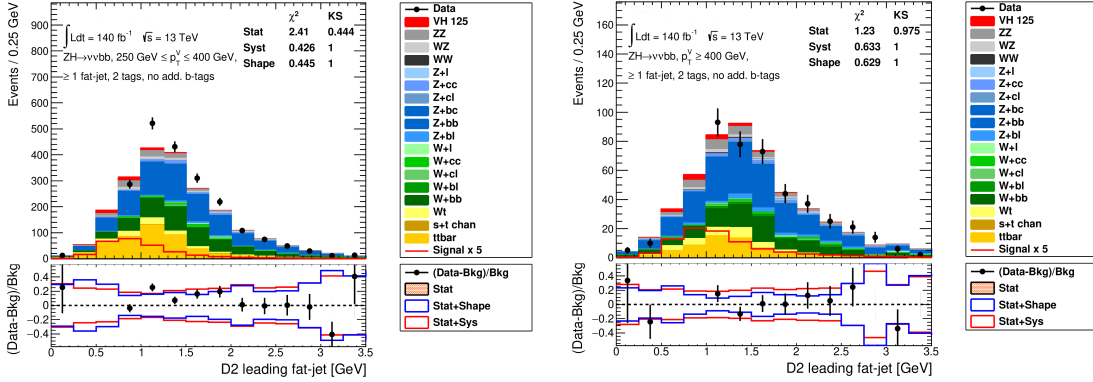


Figure 5.16: Distributions of D_2 in inclusive medium p_T^V SR and inclusive high p_T^V SR in 0L channel. Data points in bins with more than 5% signal events are not shown to follow the "blinded analysis" policy at this stage.

$C_2 <$	0.21	0.22	0.23	0.24	0.25	0.26	0.27	No cut
Significance 0L	1.980	2.020	2.035	2.043	2.038	2.028	2.022	1.910
Significance 2L	1.646	1.656	1.634	1.625	1.626	1.631	1.599	1.461

Table 5.9: C_2 cut scanning result in 0L channel and 2L channel. Significance is calculated from preliminary fitting with systematic uncertainties.

$C_2 < 0.24$ (0.22) is found to have the maximum increase on the expected significance in 0L (2L) channel. C_2 cut is also able to improve the significance in 1L channel when evaluated using $s/\sqrt{b}$ and fitting without systematic uncertainties induced by C_2 . However, it is not able to improve the fitting result when full systematic uncertainties are considered. This decrease of significance is caused by the poorly derived C_2 systematic uncertainties, especially the C_2 resolution. The C_2 resolution is estimated as 15% everywhere, which is implemented using Gaussian random number for the up and down variations. For some background MC samples like single top, diboson and W +jets, this approach suffers significant fluctuations in several bins with very limited statistics. The C_2 cut was not able to enter the baseline event selection in $VH(b\bar{b})$ boosted analysis due to this problem. It remains as a potential improvement for future analyses.

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

$C_2 <$	0.21	0.22	0.23	0.24	0.25	0.26	0.27	No cut
$s/\sqrt{b}$ 1L	1.333	1.346	1.356	1.358	1.359	1.354	1.347	1.246
Significance 1L								
with subR syst.	1.716	1.735	1.759	1.744	1.766	1.761	1.756	1.810
no subR syst.	1.716	1.743	1.774	1.780	1.805	1.802	1.809	1.810
no sub syst.	1.729	1.758	1.787	1.795	1.822	1.825	1.820	1.810

Table 5.10: C_2 cut scanning result in 1L channel. Significance is calculated from preliminary fitting with full systematic uncertainties (with subR syst.), with systematic uncertainties except the C_2 resolution (no subR syst.), without C_2 induced systematic uncertainties (no sub syst.).

5.3 Systematic uncertainties

The measurement is subject to a few systematic uncertainties as well as statistical uncertainties. The statistical uncertainties are results of stochastic fluctuations due to the limited statistics in the samples. The systematic uncertainties come from a variety of sources mainly in two categories: uncertainties related to experimental setup and uncertainties related to modeling of the MC samples. Each systematic uncertainty is implemented as a deviation in the respective model and it is treated as a nuisance parameter (NP). This section describes the experimental uncertainties and the modeling uncertainties in $VH(b\bar{b})$ boosted analysis.

5.3.1 Experimental systematic uncertainties

The experimental uncertainties are systematic uncertainties from the measurements in the detector and the experimental resolutions. The evaluation of these uncertainties follow the official ATLAS recommendations and they are evaluated in a similar manner for all the analyses. The experimental systematic uncertainties comes in the following categories.

- Event level: uncertainties related to the integrated luminosity and the pile-up modeling.
- Electron: uncertainties related to the electron trigger, the reconstruction and identification efficiency of the electrons, and the energy calibration.
- Muon: uncertainties for muon reconstruction, identification and isolation efficiency, muon trigger, momentum and track-to-vertex association.
- Small- R jets: jet energy scale and resolution uncertainties, pile-up correction and flavor tagging related uncertainties.
- Large- R jets: jet energy (mass) scale and resolution uncertainties.
- b -tagging: uncertainties related to b -tagging efficiency.
- E_T^{miss} and $E_{T,\text{trk}}^{\text{miss}}$: trigger efficiency, soft term and E_T^{miss} energy scale related uncertainties.

The dominant uncertainties in $VH(b\bar{b})$ boosted analysis are the large- R jet uncertainties and the b -tagging uncertainties. The following part of this section introduce each category of the systematic uncertainty briefly.

The integrated luminosity of Run 2 pp collision data recorded by the ATLAS detector is mainly measured by the Luminosity Cherenkov Integrating Detector (LUCID) [142], which is located at 17 m from the ATLAS interaction point. The measurement is complemented by other sub-detectors including the ATLAS beam conditions monitor (BCM) and the calorimeter. The systematic uncertainties arise from many sources. The contribution from each source is estimated through comparing measurements by the LUCID and other detectors. The total systematic uncertainty estimated for the integrated luminosity in Run 2 is 1.7% [143].

The pile-up uncertainties arise from the modeling of pile-up events. The simulated events are corrected to match the pile-up condition in data through a process called pile-up reweighting. This process

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

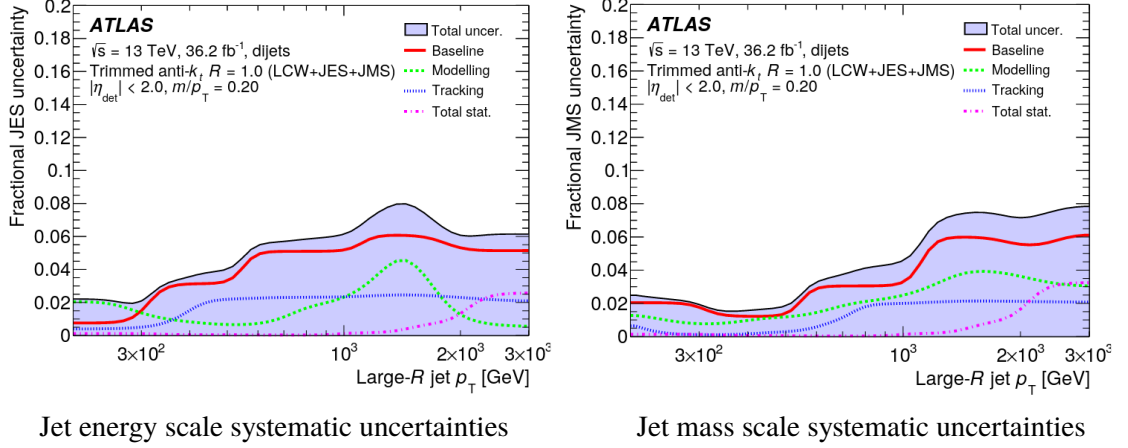


Figure 5.17: The total jet energy and mass scale uncertainties derived using R -track method for $R = 1.0$ LCTopo jet [115].

corrects the distribution of N_{PV} and $\langle\mu\rangle$ in simulated samples. The pile-up systematic uncertainty is estimated to be at the same level as the correction.

The electron related systematic uncertainties arise from the reconstruction, identification and isolation efficiencies, as well as the energy scale and resolution. The total electron energy scale uncertainties are estimated to be 4×10^{-4} in the barrel and 2×10^{-3} in the end caps. The energy resolution is of the order of 1% for electron with transverse momentum larger than 30 GeV. The systematic uncertainty in the identification efficiency depends on the electron p_T , being 7% at $p_T = 4.5$ GeV and reaching less than 1% for $30 \text{ GeV} < p_T < 250 \text{ GeV}$ [98].

The muon related uncertainties also arise from the reconstruction, identification and isolation efficiencies. Additionally, momentum scale and resolution uncertainties and track-to-vertex association efficiency uncertainty are included. The total uncertainty of muon reconstruction and identification efficiency is below 0.1% over a large muon momentum range. The total uncertainty from track-to-vertex association efficiency is around 0.01% for muons with $p_T > 3 \text{ GeV}$ [100]. The momentum scale uncertainty varies from a minimum of 0.05% for $|\eta| < 1$ to a maximum of 0.3% for $|\eta|$ at around 2.5. The muon momentum resolution is around 4% in a wide p_T range [99].

The systematic uncertainties that affect small- R jets are mainly the jet energy scale uncertainties and the jet energy resolution. These uncertainties are derived in the in-situ calibration as described in Section 3.5.3. The uncertainties related to large- R jets affect the analysis the most. These uncertainties are also derived in the in-situ calibration. The large- R jet uncertainties include the jet energy and mass scale uncertainties and the resolution. The scale uncertainties are derived using the R -track method. It covers the difference between data and simulated samples. Figure 5.17 shows the jet energy scale uncertainty and the jet mass scale uncertainties derived for the large- R jet collection used in $VH(b\bar{b})$ analysis. It has four components:

- baseline uncertainty covering the difference between data and the default MC generator;
- modeling uncertainty covering the difference between the default and alternative MC generators;

5.3 Systematic uncertainties

- tracking uncertainty arises from the effect of various tracking uncertainty sources;
- statistical uncertainty arises from the limited statistics in data and simulated samples.

The total scale uncertainties varies from 2% to 8% for different large- R jet p_T . The jet energy and mass resolution is estimated conservatively, with values up to 20% as shown in Section 3.5.3.

The b -tagging uncertainties are the other major component of the systematic uncertainties. These uncertainties account for the difference between simulation of heavy flavor production and data. The b -tagging uncertainties are estimated for the data-to-simulation scale factor $\kappa_{\epsilon_b}^{\text{data/MC}} = \epsilon_b^{\text{data}} / \epsilon_b^{\text{sim}}$, where ϵ_b denotes the b -tagging efficiency. The scale factor and its corresponding uncertainty are derived for b -jets, c -jets, light jets and τ -lepton jets separately. The total uncertainty derived for b -jets is about 5%, around 10% for c -jets, and up to 30% for light jets [144].

The uncertainties associated with the measurement of E_T^{miss} are provided for the E_T^{miss} scale and resolution. The extraction of the uncertainties is based on data-to-simulation comparisons of several observables related to the soft term E_T^{miss} and the overall E_T^{miss} [116].

5.3.2 Modeling uncertainties

The signal and background processes used in the $VH(b\bar{b})$ analysis are modeled by several MC event generators as described in Section 5.1.2. Each process is modeled by a nominal generator, which is the default generator used in the simulation, as well as a list of alternative generators that are used to extract the modeling uncertainties. The modeling uncertainties can be divided into three categories: cross section normalization, acceptance in various analysis regions and jet mass shape.

The cross sections as well as their uncertainties used in the simulation are taken from the most accurate theoretical predictions available up to date. The uncertainties of the theoretical predication are treated as the cross section normalization systematic uncertainties. For the dominant backgrounds (V + heavy flavor, $t\bar{t}$), the normalizations are free parameters floated in the fitting.

The acceptance systematic uncertainties are derived from comparisons between different MC generators normalized to the same cross section. It covers the effects from imperfect modeling of different phase spaces in the analysis. The relative acceptance uncertainty of a region B with respect to another region A is computed according to the double ratio formula:

$$\frac{N_A(\text{nominal})}{N_B(\text{nominal})} \cdot \frac{N_B(\text{alt.})}{N_A(\text{alt.})} - 1, \quad (5.3)$$

where $N_A(\text{nominal})$ and $N_A(\text{alt.})$ are the number of events generated by the nominal MC generator and the alternative MC generator in region A ; $N_B(\text{nominal})$ and $N_B(\text{alt.})$ are the number of events generated by the nominal MC generator and the alternative MC generator in region B . The acceptance uncertainties are derived using m_J distributions. The acceptance uncertainties considered in the $VH(b\bar{b})$ boosted analysis are:

- the migration between p_T^V regions (med-to-high p_T^V);
- The extrapolation from CR to SR (CR-to-SR);

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

- ratio of event yields in high purity and low purity SRs (LP-to-HP);
- flavor composition (specific for V +jets sample);
- distributions of events in different channels due to lepton and jet counting (channel extrapolation).

The large- R jet mass shape uncertainties are considered in each analysis region separately. The nominal and the alternative MC generators are normalized to have the same cross section. These uncertainties account for higher order QCD and electroweak corrections, variations of the parton shower and underlying event models.

Signal modeling systematic uncertainties

The VH signal processes include $qq \rightarrow VH$ and $gg \rightarrow VH$. The nominal qq -induced samples are generated using POWHEG MiNLO + PYTHIA 8, the nominal gg -induced samples are generated using POWHEG + PYTHIA 8 as described in Section 5.1.2. The systematic uncertainties considered in the modeling of signal processes are the following.

- **Branching Ratio:** a theoretical uncertainty of 1.6% has been used in the branching ratio of $H \rightarrow b\bar{b}$. It is a normalization uncertainty.
- **Electroweak (EW) Corrections:** uncertainty applied to account for missing higher order EW corrections. The uncertainty is a function of p_T^V . It is estimated as the maximum of the generic size of neglected NNLO EW effects, which is 1.5% in the case of WH process and 1.0% in the case of ZH process. The uncertainty has effects on m_J shape.
- **Parton Shower (PS):** uncertainties arising from the hadronization model and parton shower. These uncertainties are estimated by comparing the nominal MC samples to the alternative MC samples, where HERWIG 7 instead of PYTHIA 8 is used to model the parton showers. The signal acceptance uncertainty for analysis region i is derived as:

$$\frac{N_i^{\text{Herwig 7}}}{N_i^{\text{Pythia 8}}} - 1, \quad (5.4)$$

where N_i is the signal yields in the analysis region i , generated by HERWIG 7 or PYTHIA 8. Such acceptance uncertainties have been derived only in the SRs due to the low signal contamination in the CRs. The values of the signal acceptance uncertainties are summarized in Table 5.11. The change from PYTHIA 8 to HERWIG 7 also causes difference in m_J shape. The m_J variations are derived in every bins in the signal regions, the templates are then used in the fit.

- **QCD Scales:** uncertainties arising from higher orders in QCD variations of the renormalization scale μ_R and the factorization scale μ_F on the POWHEG samples. These uncertainties are obtained by varying μ_R and μ_F by a factor of 2 up and down. The QCD scale uncertainties have been arranged into seven nuisance parameters:
 1. Δ_Y : the overall cross section uncertainty;
 2. $\Delta_{75}, \Delta_{150}, \Delta_{250}, \Delta_{400}$: acceptance uncertainties across different p_T^V boundaries (75 GeV, 150 GeV, 250 GeV, 400 GeV).

5.3 Systematic uncertainties

	0L		1L		2L
	HP SR	LP SR	HP SR	LP SR	SR
med. p_T^V	2.4%	4.4%	0.9%	5.2%	3.7%
high p_T^V	0.0%	7.4%	2.8%	6.5%	4.7%

Table 5.11: Signal acceptance uncertainties arising from parton showers.

p_T^V	Δ_Y^{qq}	Δ_{75}^{qq}	Δ_{150}^{qq}	Δ_{250}^{qq}	Δ_{400}^{qq}
[250, 400] GeV	0.7%	3.3%	1.3%	1.4%	-0.38%
[400, +∞] GeV	0.7%	3.3%	1.3%	1.4%	1.8%
	Δ_Y^{gg}	Δ_{75}^{gg}	Δ_{150}^{gg}	Δ_{250}^{gg}	Δ_{400}^{gg}
[250, 400] GeV	25%	26%	13%	14%	-1.3%
[400, +∞] GeV	25%	26%	13%	14%	15%

Table 5.12: Values of QCD scale uncertainties for nuisance parameters in STXS bins split in p_T^V [145].

3. Δ_0, Δ_1 : acceptance uncertainties across HP SR with 0 additional calo jet and LP SR with at least 1 additional calo jet.

The QCD scale uncertainties are calculated separately for qq -induced and gg -induced signal samples. The impact of these uncertainties is derived in the maximum splitting STXS bins. It is possible to merge them to adapt the STXS bins used in the analysis. The impact of these uncertainties for the seven nuisance parameters in each bin are summarized in Table 5.12 and Table 5.13.

- **Parton Distribution Function (PDF) + α_S :** these uncertainties arise from variations due to limited statistics and incompatibilities among datasets in PDF extraction, as well as the precision level of QCD coupling constant α_S . The uncertainties are extracted by comparing 30 alternative samples with varied PDF to the nominal PDF. Additionally, two variations of α_S from the nominal value are available for the samples with nominal PDF. The PDF + α_S uncertainties are derived separately for quark-induced samples and gluon-induced samples. The values of the uncertainties in each STXS bin are shown in Figure 5.18. The effects of PDF + α_S variations on m_J shape have been studied and found to be negligible.

A summary of the signal modeling uncertainties considered in the $VH(b\bar{b})$ boosted analysis is shown in Table 5.14. In the analysis, merged STXS bins instead of the maximum splitting are used. The uncertainties have been derived to correspond to the actual STXS bins in the analysis.

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

p_T^V	Δ_0^{qq}			Δ_1^{qq}	
	0 jet	1 jet	≥ 2 jets	1 jet	≥ 2 jets
[250, 400] GeV	-5.4%	5.3%	5.3%	-5.0%	6.7%
[400, + ∞] GeV	-6.8%	5.5%	5.5%	-5.7%	6.7%
p_T^V	Δ_0^{gg}			Δ_1^{gg}	
	0 jet	1 jet	≥ 2 jets	1 jet	≥ 2 jets
[250, 400] GeV	-100%	28%	28%	-38%	28%
[400, + ∞] GeV	-100%	30%	30%	-66%	30%

Table 5.13: Values of QCD scale uncertainties for nuisance parameters in STXS bins split in number of additional calo jets [145].

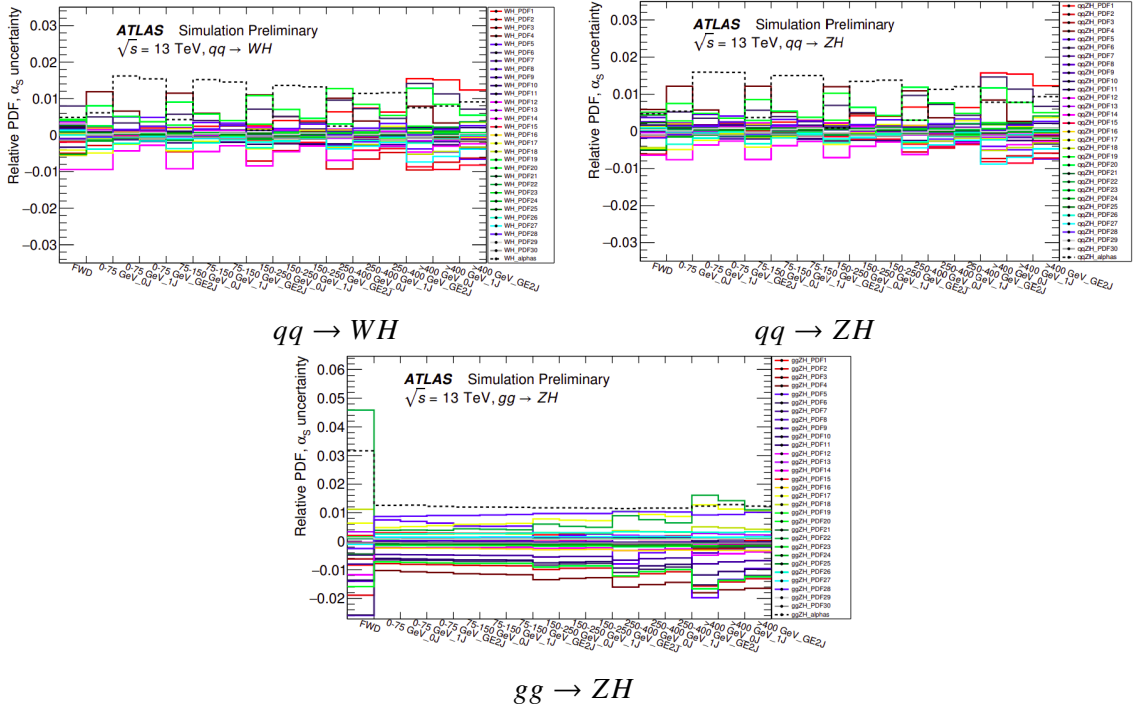


Figure 5.18: PDF + α_s uncertainties across STXS bins for PDF (colored) and α_s (dashed black) variations. The x -axis labeling indicates the STXS bins split in p_T^V and number of additional small- R jets: 0, 1 and ≥ 2 jets, labeled as 0J, 1J, GE2J [145].

5.3 Systematic uncertainties

Name	Sources	Affected region	Effect	Value
SysTheoryBRbb	BR variation	$H \rightarrow b\bar{b}$, all regions	norm.	1.6%
SysTheoryUEPSAcc	PS variation	$H \rightarrow b\bar{b}$	norm.	Tab. 5.11
SysTheoryUEPSShape	PS variation	$H \rightarrow b\bar{b}$	shape	-
SysQCDScaleDeltaY, 75, 150, 250, 400	QCD scale variation (p_T^V)	$qq \rightarrow VH, gg \rightarrow ZH$	norm.	Tab. 5.12
SysQCDScaleDelta1,2	QCD scale variation (n_{jets})	$qq \rightarrow VH, gg \rightarrow ZH$	norm.	Tab. 5.13
SysTheoryPDF_{1-30}	PDF variation	$qq \rightarrow VH, gg \rightarrow ZH$	norm.	Fig. 5.18
SysTheoryalphas	α_S variation	$qq \rightarrow VH, gg \rightarrow ZH$	norm.	Fig. 5.18
SysVHNLOEWK	NLO EW correction	$H \rightarrow b\bar{b}$, all regions	shape, norm.	-

Table 5.14: Nuisance parameters for signal modeling systematic uncertainties. The first column is the name of the uncertainty, the second column is the source of the uncertainty. The third column lists the samples that are affected by these systematic uncertainties. The last column shows the effects of these uncertainties in the final fit.

V+jets modeling systematic uncertainties

The V+jets background is the dominant background in all the channels. The nominal V+jets background samples are produced using SHERPA 2.2.1 as described in Section 5.1.2. The flavor composition of the jets are: $V + bb$, $V + bc$, $V + bl$, $V + cc$, $V + cl$ and $V + ll$, where b refers to b -jet, c refers to c -jet, l refers to light jet that can be u , d , s or gluon. The events with at least one b -jet or c -jet are classified as V+ heavy flavor (HF). The V+HF component (bb , bc , bl , cc) represents more than 90% of the background, therefore the normalization factors of V+HF background are free floated in the fit. The normalization factors are considered separately for each component of the Z+HF and W+HF processes.

The systematic uncertainties arising from alternative variations considered in the modeling of V+jets are the following.

- Variations of the renormalization scale μ_R and the factorization scale μ_F in the matrix element (ME) calculation by a factor of 2 or 0.5 compared to the normalized scale. These variations are implemented as internal weight in the nominal samples. The effects of these systematic uncertainties are negligible.
- Variations from alternative PDF sets. These systematic uncertainties are also implemented as internal weight in the nominal SHERPA samples. The impact of these variations are negligible in the phase space of the analysis in general.
- Variations from alternative MC event generator. The alternative samples are generated using MADGRAPH for the matrix element and PYTHIA 8 for parton shower.

The acceptance uncertainties considered in the modeling of V+jets background are all relative acceptance uncertainties because the main component of this background is free floated in the fit. All the effects from variations of μ_F , μ_R , PDF and alternative MC generator are considered in the acceptance uncertainties. These relative acceptance uncertainties are:

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

V+jets acceptance uncertainties				
Boson	W		Z	
Channel	0L	1L	0L	2L
CR-to-SR	90%	40%	40%	-
LP-to-HP	18%		18%	-
Med-to-High p_T^V	30%	10%	10%	
Channel Extrap.	20%	-	16%	-
$V + bc$ flavor comp.	30%			
$V + bl$ flavor comp.	30%			
$V + cc$ flavor comp.	20%			
$V + cl$ norm.	30%			
$V + ll$ norm.	30%			

Table 5.15: Summary of V+jets acceptance uncertainties; flavor composition uncertainties for $V + bc$, $V + bl$, $V + cc$; normalization uncertainties for $V + cl$ and $V + ll$.

- CR-to-SR extrapolation uncertainties applied in the control regions in 0L and 1L channel as the V+jets contribution is much lower in the CRs. The difference of CR-to-SR acceptance between the nominal SHERPA samples and the alternative MADGRAPH samples is up to 90% for W+HF process in 0L channel, and 40% for Z+HF process in 0L channel and W+HF process in 1L channel.
- LP-to-HP extrapolation uncertainty applied in high purity SRs in 0L and 1L channel. The uncertainty is 18% for V+HF processes in both 0L and 1L channels.
- Med-to-high p_T^V extrapolation uncertainties applied in high p_T^V regions. The uncertainties are estimated to be 30% on W+HF process in 0L channel and 10% in other regions.
- The uncertainties for extrapolation among channels are applied in 0L channel because it has more complicated background composition than other channels. The values of the uncertainties are 20% for Z+jets and 18% for W+jets.

The flavor composition uncertainties are applied to $V + bc$, $V + bl$ and $V + cc$ processes, which is 30% for $V + bc$ and $V + bl$ processes, 20% for $V + cc$ samples. Additionally, a 30% uncertainty is applied to the normalization of $V + cl$ and $V + ll$ processes as their normalization factors are not floating in the fit. A summary of the acceptance uncertainties is shown in Table 5.15.

The m_j shape uncertainties are extracted by comparing the nominal m_j distributions to the alternative distributions with varied μ_R . The ratio between the alternative distribution and the nominal distribution

5.3 Systematic uncertainties

Name	Sources	Affected region	Effect	Value
norm_Zbb	Z+HF norm.	Z+HF, all regions	norm.	float
SysZbbNorm_L0	Z+HF channel extrap.	Z+HF, applied to 0L	norm.	16%
SysZclNorm	Z + cl norm.	Z + cl , all regions	norm.	30%
SysZllNorm	Z + ll norm.	Z + ll , all regions	norm.	30%
SysZbcZbbRatio	Z + bc flavor comp.	Z + bc , applied to 0L, 2L	norm.	30%
SysZblZbbRatio	Z + bl flavor comp.	Z + bl , applied to 0L, 2L	norm.	30%
SysZccZbbRatio	Z + cc flavor comp.	Z + cc , applied to 0L, 2L	norm.	20%
SysRatioMedHighPtv_BMin400_Zhf	med-to-high p_T^V	Z+HF, applied to 0L 2L high p_T^V	norm.	10%
SysRatioHPLP_DSRnoaddbjetsr_J0_Zhf	LP-to-HP	Z+HF, applied to 0L HP SR	norm.	18%
SysRatioSRCR_DSRnoaddbjetsr_Zhf	CR-to-SR	Z+HF, applied to 0L CR	norm.	40%
SysZMbbBoosted	μ_R var. on m_J	Z+jets, all regions	shape	-

Table 5.16: Nuisance parameters for Z+jets modeling systematic uncertainties. The first column is the name of the uncertainty, the second column is the source of the uncertainty. The third column lists the samples that are affected by these systematic uncertainties. The last column shows the effects of these uncertainties in the final fit.

Name	Sources	Affected region	Effect	Value
norm_Wbb	W+HF norm.	W+HF, all regions	norm.	float
SysWbbNorm_L0	W+HF channel extrap.	Z+HF, applied to 0L	norm.	20%
SysWclNorm	W + cl norm.	W + cl , all regions	norm.	30%
SysWllNorm	W + ll norm.	W + ll , all regions	norm.	30%
SysWbcWbbRatio	W + bc flavor comp.	W + bc , applied to 0L, 1L	norm.	30%
SysWblWbbRatio	W + bl flavor comp.	W + bl , applied to 0L, 1L	norm.	30%
SysWccWbbRatio	W + cc flavor comp.	W + cc , applied to 0L, 1L	norm.	20%
SysRatioMedHighPtv0L_BMin400_Whf	med-to-high p_T^V	W+HF, applied to 0L high p_T^V	norm.	30%
SysRatioMedHighPtv1L_BMin400_Whf	med-to-high p_T^V	W+HF, applied to 1L high p_T^V	norm.	10%
SysRatioHPLP_DSRnoaddbjetsr_J0_Whf	LP-to-HP	Z+HF, applied to 0L, 1L HP SR	norm.	18%
SysRatioSRCR_DSRnoaddbjetsr_Whf	CR-to-SR	W+HF, applied to 0L, 1L CR	norm.	90%, 40%
SysWMbbBoosted	μ_R var. on m_J	W+jets, all regions	shape	-

Table 5.17: Nuisance parameters for W+jets modeling systematic uncertainties. The med-to-high acceptance uncertainty in 1L channel (SysRatioMedHighPtv1L_BMin400_Whf) is applied decorrelated between HP SR, LP SR and CR.

is fitted using a function ($e^{p_0+p_1x} + p_2$). The parameters are fitted separately in each analysis region. The up and down variations of the shape uncertainties are considered to be asymmetric.

A summary of the nuisance parameters for the modeling uncertainties of Z+jets samples is shown in Table 5.16, and Table 5.17 is for the modeling uncertainties of W+jets samples.

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

$t\bar{t}$ modeling systematic uncertainties

The $t\bar{t}$ background is the other dominant background in 0L channel and 1L channel. The nominal $t\bar{t}$ samples are generated using POWHEG generator, and interfaced to PYTHIA 8 for modeling of the parton shower. Control regions for $t\bar{t}$ process are defined by requiring at least one additional b -tagged track jets not matched to the leading large- R jet. The normalization factors of the $t\bar{t}$ background are also free floated separately in 0L channel and 1L channel in the fit because it is the dominate background. The $t\bar{t}$ background in 2L channel is very minimal. A 20% uncertainty is applied to the normalization of $t\bar{t}$ process in 2L channel.

Several alternative $t\bar{t}$ samples have been considered to estimate the modeling uncertainties. The variations are the following.

- POWHEG + HERWIG 7: this alternative sample is generated using POWHEG, which is the same configuration as the nominal sample, while the parton shower is simulated using HERWIG 7 instead of PYTHIA 8.
- MADGRAPH + PYTHIA 8: this alternative sample is generated using MADGRAPH at NLO precision level while the simulation of parton shower is the same as the nominal sample.
- Initial state radiation (ISR) variations: variations of the renormalization scale μ_R and the factorization scale μ_F in the matrix element (ME) calculation by a factor of 2 or 0.5. These variations are implemented as the internal weight in the nominal $t\bar{t}$ samples.
- Final state radiation (FSR) variations: variations of the renormalization scale $\mu_{R,FSR}$ in the simulation of parton shower with PYTHIA 8. Several factors are available compared to the nominal value, including 2.0, 1.75, 1.5, 1.25, 0.875, 0.75, 0.625, 0.5.

All the alternative simulated samples are required to have at least one W boson decays leptonically and at least one large- R truth jet with $p_T > 150$ GeV. This filter removes events that cannot enter the analysis regions to improve the efficiency of the simulation.

The acceptance uncertainties of $t\bar{t}$ modeling are relative acceptance uncertainties as the normalization factors are free floated. The most significant sources of the uncertainties are the variations of the matrix element calculation, followed by the variations of the parton shower simulation. The effects of ISR and FSR variations are subdominant. The acceptance uncertainties estimated for $t\bar{t}$ samples are:

- LP-to-HP acceptance uncertainty applied to high purity SR in 0L channel and 1L channel. The nuisance parameters for 0L and 1L are kept uncorrelated, and the value of uncertainty is estimated to be 18%.
- CR-to-SR acceptance uncertainty applied to SRs in 0L and 1L channel. The value of the uncertainty is estimated to be 8%.
- Med-to-high p_T^V acceptance uncertainty applied to high p_T^V regions. The value of this uncertainty is set to be 20%, which is higher than the one actually derived. It is set to higher value due to the top mis-modeling known in other analyses.

5.3 Systematic uncertainties

Name	Sources	Affected region	Effect	Value
norm_ttbar_L0	$t\bar{t}$ norm.	$t\bar{t}$ 0L, all regions	norm.	float
norm_ttbar_L1	$t\bar{t}$ norm.	$t\bar{t}$ 1L, all regions	norm.	float
SysttbarNorm_L2	$t\bar{t}$ norm.	$t\bar{t}$ 2L, all regions	norm.	20%
SysRatioTtbarVRjets	3pVR-to-2VR extroa.	$t\bar{t}$ 2VR, 0L and 1L	norm.	20%
SysRatioMedHighPtv_BMin400_L0_ttbar	med-to-high p_T^V	0L high p_T^V regions	norm.	20%
SysRatioMedHighPtv_BMin400_L1_ttbar	med-to-high p_T^V	1L high p_T^V regions	norm.	20%
SysRatioHPLP_DSRnoaddbjetsr_L0_J0_ttbar	LP-to-HP	0L HP SR	norm.	18%
SysRatioHPLP_DSRnoaddbjetsr_L1_J0_ttbar	LP-to-HP	1L HP SR	norm.	18%
SysRatioSRCR_DSRnoaddbjetsr_L0_ttbar	CR-to-SR	0L SR	norm.	6%
SysRatioSRCR_DSRnoaddbjetsr_L1_ttbar	CR-to-SR	1L SR	norm.	6%
SysTtbarMJISR	ISR var. on m_J	all regions	shape	-
SysTtbarMJPS	PS var. on m_J	all regions	shape	-

Table 5.18: Nuisance parameters for $t\bar{t}$ modeling systematic uncertainties.

- Extrapolation uncertainty applied to top CRs with exactly two matched track jets. The $t\bar{t}$ events can have two track jets matched to the leading large- R jet (2VR) or three or more matched track jets (3pVR). This uncertainty is applied to cover the extrapolation uncertainty between "3pVR" events and "2VR" events. Its value is 20% and it is fully correlated among channels and p_T^V regions.

The shape uncertainties are derived by comparing the nominal m_J distributions to the alternative m_J distributions. The variations considered for the shape uncertainties are:

- An ISR variation with effects of up to 20% on m_J shapes in both 0L channel and 1L channel.
- Parton shower variations with effects of up to 10% on m_J shapes in both 0L channel and 1L channel.

Finally, a summary of the modeling uncertainties considered in $t\bar{t}$ samples is shown in Table 5.18.

Single top modeling systematic uncertainties

The single top processes are divided into three channels: s -channel, t -channel and Wt -channel. The nominal single top samples are generated using POWHEG + PYTHIA 8 as described in Section 5.1.2. The single top background is a sub-dominant background in 0L channel and 1L channel. The fraction of single top process from s -channel and t -channel is even smaller (less than 1%) in the analysis regions. Therefore the alternative samples are generated considering only the Wt -channel. Similar to the nominal single top samples, the samples are generated both applying a dilepton filter and inclusively.

The variations considered in the modeling of single top samples are the same as the ones in the modeling of $t\bar{t}$ samples, including variations of parton shower (PS), matrix element (ME), ISR and FSR. The single top process in Wt -channel at NLO has interference with $t\bar{t}$ process. The interference

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

Name	Sources	Affected region	Effect	Value
stopWtNorm	Wt -channel norm.	Wt -channel, all regions	norm.	20%
stopsNorm	s -channel norm.	s -channel, all regions	norm.	4.6%
stoptNorm	t -channel norm.	t -channel, all regions	norm.	4.4%
SysStopNorm_L0	1L-to-0L	0L, all regions	norm.	20%
SysRatioMedHighPtv_BMin400_Stop	med-to-high p_T^V	0L, 1L high p_T^V regions	norm.	20%
SysRatioHPLP_DSRnoaddbjetsr_L1_J0_ttbar	LP-to-HP	0L, 1L HP SR	norm.	25%
SysRatioSRCR_DSRnoaddbjetsr_L0_ttbar	SR-to-CR	0L, 1L CR	norm.	30%
SysStopWtDSDR	DS/DR var. on m_J	Wt -channel, all regions	shape, norm.	-

Table 5.19: Nuisance parameters for single top modeling systematic uncertainties.

is removed using the diagram removal (DR) scheme that removes the resonant $t\bar{t}$ Feynman diagrams or the diagram subtraction (DS) scheme that implements a subtraction term to locally cancel the $t\bar{t}$ contribution. The DR scheme was implemented in the nominal single top samples, while the alternative samples are generated using both the DR and the DS scheme.

The majority of the single top events (80%) are in the SRs, so the top CRs are not suitable for single top background studies. The shape of the single top background and the flavor composition of the large- R jet is found to be consistent in 0L channel and 1L channel, which means the modeling uncertainties could be treated in a correlated way. Therefore, all the single top modeling uncertainties are evaluated in 1L channel and then applied to both 0L channel and 1L channel.

The normalization uncertainties are derived for cross section of each single top process. The total normalization uncertainty is estimated to be 4.4%, 4.6% and 6.2% for t -channel, s -channel and Wt -channel respectively. For t -channel and s -channel, only a single normalization uncertainty is applied to each of the process because the fraction of these background is smaller than 1% of the total top background in all the analysis regions. The uncertainty for Wt -channel is negligible compared to the uncertainties arise from the variations of alternative samples.

The acceptance uncertainties of single top modeling are evaluated by comparing the nominal samples and the alternative samples generated using DR scheme. The uncertainty from the discrepancy between DR scheme and DS scheme is considered separately. The following acceptance uncertainties are considered in single top modeling:

- Normalization uncertainty to cover the overall difference between the nominal samples and the alternative samples. This uncertainty is estimated to be about 20%. The cross section uncertainty is covered by this uncertainty for Wt -channel.
- LP-to-HP acceptance uncertainty applied to HP SR. The value of this uncertainty is estimated to be 25%.
- Med-to-high p_T^V acceptance uncertainty applied to high p_T^V regions, with a value of 20%.
- CR-to-SR acceptance uncertainty applied to CRs because the fraction of single top events in CR is much smaller than in SR. The uncertainty is estimated to be 30%.

5.3 Systematic uncertainties

Additionally, an acceptance uncertainty of 20% is applied to 0L channel. This uncertainty accounts for the difference between 0L and 1L since all the other uncertainties are derived in 1L channel. The variations caused by the discrepancy between DS scheme and DR scheme is also included.

The shape uncertainties are estimated by comparing the m_j distributions between the nominal samples and the alternative samples. The dominant contribution to the shape uncertainties is the discrepancy between the DS scheme and the DR scheme, while the effects from all the other sources are neglected. The variations from alternative samples with DS scheme is treated as a shape uncertainty and a normalization uncertainty in each region. These uncertainties are extracted from the shape variation of the nominal sample and the alternative samples, then multiplied by a normalization factor given by the yield ratio of the nominal sample and the alternative samples.

A summary of the single top modeling uncertainties is shown in Table 5.19.

Diboson modeling systematic uncertainties

The Diboson background is a sub-dominant background in all the channels. The final states include ZZ , WZ and WW , where one of the vector boson decays hadronically, the other one decays leptonically. The different diboson processes contribute to the analysis regions through the following: $ZZ \rightarrow \nu\nu b\bar{b}$ for 0L channel; $ZZ \rightarrow ll b\bar{b}$ for 2L channel; $WZ \rightarrow l\nu b\bar{b}$ for 1L channel. Additionally, WZ and WW processes can have small contributions due to the mis-tagging of a jet from the W decay or the failure to reconstruct one of the leptons, such as $WZ \rightarrow q\bar{q}\nu\nu(ll)$, $WW \rightarrow l\nu q\bar{q}$. As described in Section 5.1.2, the quark-induced diboson nominal samples are generated using SHERPA 2.2.1, the gluon-induced diboson nominal samples are generated using SHERPA 2.2.2.

The modeling uncertainties are studied for WZ and ZZ process, not for WW process because its contribution is very small (less than 2%). The alternative variations considered in the modeling of the diboson processes are similar to the variations considered in the modeling of $V+$ jets.

- Variations of the renormalization scale μ_R and the factorization scale μ_F in the matrix element (ME) calculation by a factor of 2 or 0.5. These variations are implemented as internal weight in the nominal samples. The effects of these systematic uncertainties are negligible.
- Variations from alternative PDF sets. These systematic uncertainties are also implemented as internal weight in the nominal SHERPA samples. The impact of these variations are negligible in the phase space of the analysis in general.
- Variations from alternative MC event generator. The alternative samples are generated with POWHEG and interfaced to PYTHIA 8 for modeling of the parton shower.

A normalization uncertainty of 16% and 10% is applied to WZ and ZZ process, respectively. For WW process, a normalization uncertainty of 25% is applied. The acceptance uncertainties are extracted from comparisons of the nominal samples and the alternative samples. Effects from all the variations are considered. The following uncertainties are derived for diboson processes.

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

Process	WZ			ZZ		
Channel	0L	1L	2L	0L	1L	2L
Normalization	16%			10%		
LP-to-HP	18%			18%		-
Med-to-high	10%			6%	18%	
Channel Extra.	8%	-	-	30%	18%	-

Table 5.20: Acceptance uncertainties for diboson process. All the acceptance uncertainties except the channel extrapolation uncertainties are correlated between channels.

- LP-to-HP acceptance uncertainty applied to HP SRs. This uncertainty is evaluated in 1L channel for WZ process and applied to both 0L channel and 1L channel, with a value of 18%. For ZZ process, the uncertainty is evaluated in both 0L channel and 1L channels, and also estimated to be 18%.
- Med-to-high p_T^V acceptance uncertainty applied to high p_T^V regions. This uncertainty is evaluated in 1L channel for WZ process, with a value of 10%. The uncertainty for ZZ process is estimated to be 6% in 0L channel and 18% in 1L and 2L channels.
- Channel extrapolation uncertainty. An uncertainty of 8% is applied to 0L channel for WZ process since all the other uncertainties are derived in 1L channel. For ZZ process, the uncertainties are derived 2L channel, so an uncertainty of 30% and 18% is applied in 0L channel and 1L channel, respectively.

These normalization and acceptance uncertainties are summarized in Table 5.20. The values are estimated conservatively and all the acceptance uncertainties are correlated except the channel extrapolation uncertainties.

The shape uncertainties on m_J are derived by comparing variations of μ_R and the alternative POWHEG + PYTHIA 8 samples. The impact from μ_F and PDF variations are negligible so they are not considered in the shape uncertainties. A summary of all the modeling uncertainties is shown in Table 5.21.

Multijet estimation systematic uncertainties

The QCD multijet background is a reducible background that can be suppressed. The multijet process has a very large cross section compared to the other processes, but the majority of the multijet events are rejected in the event selection. The MC simulation of this background is infeasible due to the limited amount of events after the event selection, so the background is modeled from data using data-driven techniques.

The multijet background in 0L channel is suppressed by the anti-QCD cuts described in Section 5.2:

5.3 Systematic uncertainties

ZZ				
Name	Sources	Affected region	Effect	Value
SysVZNorm	ZZ norm.	ZZ, all regions	norm.	10%
SysZZNorm_L0	0L/2L accept.	0L	norm.	30%
SysZZNorm_L1	1L/2L accept.	1L	norm.	18%
SysRatioMedHighPtv_VZ_BMin400	med-to-high	high p_T^V region	norm.	6% - 18%
SysRatioHPLP_VZ_DSRnoaddbjetsr_J0	LP-to-HP	HP SR in 0L, 1L	norm.	18%
SysVVMURMJBoosted	μ_R variation on m_J	all regions	shape	-
SysVVP8MJBoosted	alt. MC variation on m_J	all regions	shape	-
WZ				
SysVZNorm	WZ norm.	WZ, all regions	norm.	16%
SysRatioMedHighPtv_VZ_BMin400	med-to-high	high p_T^V region	norm.	10%
SysRatioHPLP_VZ_DSRnoaddbjetsr_J0	LP-to-HP	HP SR in 0L, 1L	norm.	18%
SysVVMURMJBoosted	μ_R variation on m_J	all regions	shape	-
SysVVP8MJBoosted	alt. MC variation on m_J	all regions	shape	-
WW				
SysWWNorm	WW norm.	WW, all regions	norm.	25%

Table 5.21: Modeling uncertainties for diboson process.

- $\min [\Delta\phi(\mathbf{E}_T^{\text{miss}}, \text{small-}R \text{ jets})] > 30^\circ$,
- $\Delta\phi(\mathbf{E}_T^{\text{miss}}, \text{large-}R \text{ jets}) > 120^\circ$,
- $\Delta\phi(\mathbf{E}_T^{\text{miss}}, \mathbf{p}_T^{\text{miss}}) < 90^\circ$.

These cut values are decided from the distributions of the variables. The multijet background is reduced to negligible after applying all the cuts in 0L channel.

The multijet background events in 2L channel need to have two non-prompt leptons, since the events with one prompt and one non-prompt lepton are already included in the $t\bar{t}$ and $V+$ jets samples. The electrons are not required to have opposite sign, while the muons need to have opposite sign in the baseline event selection. The multijet background in 2L channel is estimated based on fact that the contribution of opposite sign and same sign lepton pairs is expected to be symmetrical in good approximation. The same sign muon pairs is negligible in data, which means the multijet background in muon sub-channel is also negligible. The multijet background in electron sub-channel is also estimated by selecting the same sign electron pairs, the small amount of simulated events and data with same sign electron pairs are consistent within the large statistical uncertainty. The results confirm the multijet background in 2L channel is negligible.

The multijet background contributes to 1L channel due to the semileptonic decay of the heavy flavor hadrons. It is estimated using data-driven method in a multijet enriched control region. The control region is defined using inverted lepton isolation cuts.

- Electron: LooseTrackOnly, $E_T^{\text{cone0.2}} > \max(3.5 \text{ GeV}, 0.015 p_T^l [\text{GeV}])$.
- Muon: LooseTrackOnly, $p_T^{\text{cone0.2}} > 1.25 \text{ GeV}$.

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

The shape of multijet background in the signal regions is assumed to be the same as in the control regions. The shape is extracted from the distributions of the W boson transverse mass m_{T}^W in the control regions because it provides a good discrimination between the multijet and the electroweak backgrounds. A template fit to the m_{T}^W distribution is performed to extract multijet yields in the SRs after subtracting the electroweak and the top backgrounds. The fraction of multijet background in the electron sub-channel is about 2.3%, while in muon sub-channel it is negligible. A normalization uncertainty of 55% is applied to the multijet background in 1L channel.

5.4 Statistical analysis

The VH signal of the analysis is measured from the signal strength μ . It is obtained from a binned maximum profile likelihood fit [146] to the data with m_j distributions, while the fitting can have multiple parameters of interest (POI) for different signal strengths. The significance of the signal is obtained from a log-likelihood ratio test. This section describes the statistical procedure of the $VH(b\bar{b})$ boosted analysis.

5.4.1 The Likelihood function

The maximum profile likelihood fit is done simultaneously in all the analysis regions to extract the VH signal contribution. The diboson (VZ) background contribution is also extracted as a validation of the fit. The likelihood is defined as the product of the Poisson probabilities in all the bins, multiplied by the Gaussian probabilities arising from systematic uncertainties and the limited statistics in MC samples:

$$\begin{aligned}
 L(\mu, \theta, \gamma) &= L_{\text{pois}}(\mu, \theta, \gamma) \cdot L_{\text{syst}}(\theta) \cdot L_{\text{stat}}(\gamma) \\
 &= \prod_{i=1}^{n_{\text{bins}}} \text{Pois}(N_i | \mu \cdot s_i \cdot \nu_s(\theta) + \gamma_i b_i \cdot \nu_b(\theta)) \\
 &\quad \cdot \prod_{j=1}^{n_{\text{syst}}} \text{Gauss}(\theta_j | 0, 1) \cdot \prod_{i=1}^{n_{\text{bins}}} \text{Gauss}(\beta_i | \gamma_i \beta_i, \sqrt{\gamma_i \beta_i}).
 \end{aligned} \tag{5.5}$$

The first term of the likelihood is a set of Poisson functions for the probability of observing N events given λ expected events, where $\text{Pois}(N|\lambda) = e^{-\lambda} \lambda^N / N!$. The number of expected events in bin i is expressed as $\mu \cdot s_i \cdot \nu_s(\theta) + \gamma_i b_i \cdot \nu_b(\theta)$, where μ is the signal strength, s_i is the yields of simulated signal samples in bin i , and b_i is the yields of simulated background samples in bin i . The functions $\nu_s(\theta)$, $\nu_b(\theta)$ parameterize the effects of the systematic uncertainties on the signal and background yields. They are described in more detail below. The signal strength can be understood as the ratio of the observed cross section for the signal process and its SM prediction.

$$(\sigma \cdot \mathcal{B})_{\text{obs}} = \mu \cdot (\sigma \cdot \mathcal{B})_{\text{exp}}. \tag{5.6}$$

The signal strength is the POI determined in the fit. For the $VH(b\bar{b})$ boosted analysis, a simultaneous fit with two POIs associated to the cross sections of the VH and VZ processes is implemented. More POIs are used in various compatibility checks and the STXS measurement.

The effects of systematic uncertainties are parameterized as a set of nuisance parameters. These systematic uncertainties can modify the signal and background yields. The nuisance parameters are defined as described in Section 5.3, then implemented in the likelihood as a set of Gaussian penalty terms $L_{\text{syst}}(\theta)$. Each term is of the form $\text{Gauss}(\theta|0, 1) = e^{-\theta^2/2} / \sqrt{2\pi}$, where θ is the nuisance parameter for a given systematic uncertainty. The effect of the systematic uncertainty on the yields is parameterized as an exponential response function $\nu(\theta) = (1 + \epsilon)^\theta$, where ϵ is the value of the systematic uncertainty. In this case, the value of the uncertainty is derived as the effect on yields when the systematic uncertainty varies up and down by 1σ , corresponding to $\theta = \pm 1$. Aside from

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

the nuisance parameters with Gaussian constraint, several nuisance parameters (main background normalization factors) are freely floated in the fit. The values of these uncertainties are solely determined from constraints in data.

Additionally, the statistics in the MC samples is limited, so the statistical uncertainty of the expected yields needs to be addressed. The third term of the likelihood treats the uncertainties due to the finite sample size using the Beeston-Barlow technique [147]. These uncertainties are represented with nuisance parameters γ . The parameter γ_i modifies the expected background yields in bin i from b_i to $\gamma_i \cdot b_i$. The background yields b_i in the simulated samples is approximated as being drawn from a Poisson term with mean $\gamma_i \cdot b_i$. The constraints on γ are treated using Gaussian penalty terms in the $VH(b\bar{b})$ boosted analysis,

$$\prod_{i=1}^{n_{\text{bins}}} \text{Gauss} \left(\beta_i | \gamma_i \beta_i, \sqrt{\gamma_i \beta_i} \right), \quad (5.7)$$

where $\beta_i = 1/\sigma_i^2$ and σ_i is the relative statistical uncertainty on the background yields in bin i . Only the background yields are affected by these nuisance parameters because the statistical uncertainties for signal samples are small with respect to background samples.

In summary, the full likelihood is the product of the three terms with a form as written in Eq. 5.5. The POIs are μ_{VH} and μ_{VZ} for signal strengths of the VH signal process and the VZ process, respectively. The likelihood is maximized with respect to μ , θ and γ . The maximum of L determines the observed values of the signal strengths.

5.4.2 The likelihood ratio test

The profile log-likelihood ratio test [146] is used to test the background only or background and signal hypotheses. The test statistic is the log-likelihood ratio defined as

$$q(\mu) = -2 \ln \frac{L(\mu, \hat{\theta}_\mu, \hat{\gamma}_\mu)}{L_{\text{max}}}, \quad (5.8)$$

where L_{max} is the unconditional maximum of L over all possible μ , θ and γ ; $L(\mu, \hat{\theta}_\mu, \hat{\gamma}_\mu)$ is the conditional maximum of L when μ is fixed to a given value. The background only hypothesis is evaluated at $\mu = 0$. A variable referred to as p_0 is defined using the test statistic $q(0)$. The p_0 value is computed as the probability to obtain a $q(0)$ with a value larger than the observed value under the background only hypothesis.

$$p_0 = \int_{q(0)}^{+\infty} f(q|\mu=0) dq, \quad (5.9)$$

where $f(q|\mu=0)$ is probability density function of the test statistics q under background only hypothesis. If the data are compatible with the background only hypothesis, the asymptotic approximation of $q(0)$ would follow a χ^2 distribution with one degree of freedom when the experiment could be repeated enough times [148], which means $q(0)$ is likely to have a value close to 0. Differently, if the data has more events, $q(0)$ would have much larger values. The p_0 value is used to quantify the compatibility of the background-only hypothesis with the data.

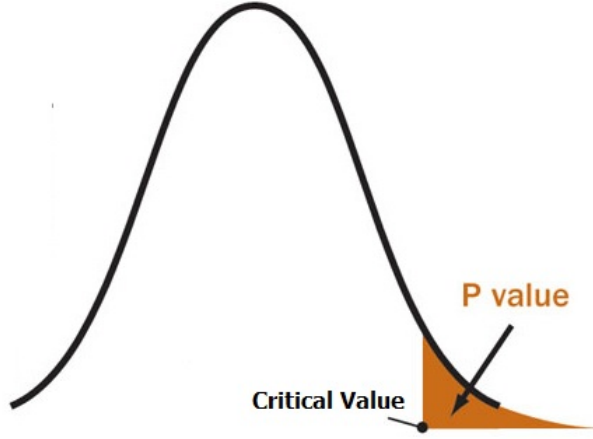


Figure 5.19: Relationship between the p -value and the significance. The black curve represents a normal distribution, the integral of the orange region equals to the p -value. The critical value (black points) represents the significance expressed in standard deviations.

The p_0 value can be expressed as the significance, defined as

$$Z = \Phi^{-1}(1 - p_0), \quad (5.10)$$

where Φ^{-1} is the inverse of the cumulative distribution function of Gaussian statistics. The value of significance is expressed in terms of standard deviations. Figure 5.19 shows the relationship between the p -value and the significance.

5.4.3 Fit inputs

The inputs to the fit are distributions of m_J in all the signal regions and control regions. The signal and background samples considered in the fit are described in Section 5.1, including the following processes.

- VH signal: $qq \rightarrow ZH \rightarrow \nu\nu b\bar{b}$, $qq \rightarrow WH \rightarrow l\nu b\bar{b}$, $qq \rightarrow ZH \rightarrow ll b\bar{b}$ and $gg \rightarrow ZH$.
- W +jets and Z +jets: split into different flavor compositions as V +HF (bb , bc , cc , bl), $V + cl$ and $V + ll$.
- Top quark pair production $t\bar{t}$.
- Single top production: s -channel, t -channel and Wt -channel.
- Diboson: WW , WZ , ZZ .
- Multijet (only in 1L channel).

The m_J templates in the fit are determined from simulation of these processes, except for the multijet background, which is extracted from data. The likelihood is built from one-dimensional m_J histograms by multiplying all the bins. The binning of the histograms are summarized in Table 5.22. The underflow and overflow are included in the first and the last bin of the histogram, respectively.

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

Channel	$p_T^V \in [250, 400]$ GeV		$p_T^V \in [400, +\infty]$ GeV	
	HP	LP	HP	LP
0L SR	[50, 210] GeV in 10 GeV	[50, 220] GeV in 10 GeV	[50, 240] GeV in 10 GeV	[50, 280] GeV in 10 GeV
0L CR	[50, 230] GeV in 20 GeV		[50, 250] GeV in 20 GeV	
1L SR	[50, 210] GeV in 10 GeV	[50, 220] GeV in 10 GeV	[50, 240] GeV in 10 GeV	[50, 280] GeV in 10 GeV
1L CR	[50, 230] GeV in 20 GeV		[50, 260] GeV in 20 GeV	
2L SR	[50, 200] GeV in 10 GeV		[50, 240] GeV in 10 GeV	

Table 5.22: Binning and range of the fit in $VH(b\bar{b})$ boosted analysis.

The choice of the different binning and range in different regions is made to maximize the resolution of the fit, while considering the following aspects:

- Avoid empty bins in the templates;
- Minimize empty bins in the data;
- Statistical uncertainty in the MC samples smaller than 20%.

5.4.4 Treatment of the nuisance parameters

The systematic uncertainties considered in the fit have been described in Section 5.3. These systematic uncertainties are treated as nuisance parameters. The impact of each systematic uncertainty on the m_J template is quantified using two additional histograms corresponding to the nuisance parameter varies 1σ up and down. The change of the yield in each bin (ϵ) is extracted from the additional histograms and then applied in the fit. Most of the systematic uncertainties have both an up and a down variation derived. Some of the uncertainties like the jet resolution are one-sided uncertainties, so the effect is symmetrized with respect to the nominal histogram to obtain the variation in the other direction.

Variations caused by certain uncertainties like the large- R jet energy scale uncertainties affect the mass and p_T of every individual jet. It can lead to bin migrations of events in the m_J distributions. In such cases, the difference between the nominal histogram and the variations might be subject to large statistical fluctuations. An iterative smoothing procedure is applied to prevent this unwanted effect. Bins are merged and grouped until the statistical uncertainty of the nominal distribution for each of the merged bins is less than 5%.

Some of the nuisance parameters considered in the fit have very small effects on the m_J template. These nuisance parameters are therefore removed to improve the stability of the fit. This procedure is referred to as pruning and applied to all the analysis regions. The pruning is performed as the following:

- Neglect the normalization uncertainty for a given sample in a region if the variation is less than 0.5% in all the bins.
- Neglect the shape uncertainty for a given sample in a region if the variation is less than 0.5% in all bins after the overall normalization is taken into account.

- Neglect the normalization and shape uncertainties for a given sample in a region if the signal contribution is less than 2% of the total backgrounds and the variations are each less than 0.5% of the total backgrounds.

The following part of this section presents several tools that are important to the validation of the fit result, including the nuisance parameter pulls and constraints, the ranking, the correlation matrices and the breakdown of the systematic uncertainties.

Nuisance parameter pulls and constraints

In the maximum profile likelihood fit, the likelihood is maximized with respect to the signal strengths μ and the nuisance parameters (NP) θ . The fit was performed first with the Asimov dataset, which is a representative dataset created from the simulated signal and background samples with $\mu_{VH} = 1$. The Asimov data is used to test the fit performance without using the actual data. The value of the nuisance parameters is not changed when fit to Asimov data, it can only be constrained. The observed signal strengths are extracted from fitting to the actual dataset. The value of the nuisance parameters is changed after fitting to the actual data. The pull is defined as the difference between the post fit value and the nominal value of the nuisance parameter.

$$\text{pull}(\theta) = \frac{\hat{\theta} - \theta_0}{\sigma_{\theta_0}} = \hat{\theta}, \quad (5.11)$$

where $\theta_0 = 0$, $\sigma_{\theta_0} = 1$, and $\hat{\theta}$ is the value obtained from the maximum profile likelihood fit. In a successful fit, the pull of all the nuisance parameters should not be too large.

The constraints on the nuisance parameters can be estimated from the likelihood as a function of θ around the maximal likelihood with $\hat{\theta}$. The estimation is done with the same method used to calculate the uncertainty of μ . If the estimated uncertainty is smaller than unity, the nuisance parameter is constrained by the fit. The pulls and constraints of the most important nuisance parameters are shown in Figure 5.20. The completed plots for pulls can be found in Appendix B. The pulls are obtained from unconditional fit to the Asimov data (red) and the actual data (black), where "unconditional" means μ is not fixed in the fit. On the contrary, conditional fit means μ is fixed to a given value in the fit. The pulls and constraints of the nuisance parameters can also be studied from a conditional fit with $\mu = 1$. This way allows to study the stability of the fit without using the POI.

The names of the nuisance parameters in the plots correspond to the systematic uncertainty names described in Section 5.3. Figure 5.20 (a) shows the NP pulls of large- R jets uncertainties. Most of the NPs are not pulled very much. The NPs related to the large- R jet mass resolution (FJ_JMR) for $t\bar{t}$ process are constrained better because of the presence of a characteristic mass peak. It allows to precisely pin down the uncertainties related to the shape of the m_J distribution. The NP related to the large- R jet mass resolution for diboson process is pulled down due to the disagreement between data and MC samples. Figure 5.20 (b) shows the pulls of the NPs related to b -tagging uncertainties. These NPs are constrained well and the pulls are small. The NP pulls for $t\bar{t}$ modeling uncertainties are shown in Figure 5.20 (c). The normalization factors are pulled a little and constrained reasonably well because the $t\bar{t}$ processes have dedicated top control regions in 0L channel and 1L channel. The NP related to the acceptance uncertainty between HP SR and LP SR in 0L channel (R_Ptv_400_L0_Ttbar) is

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

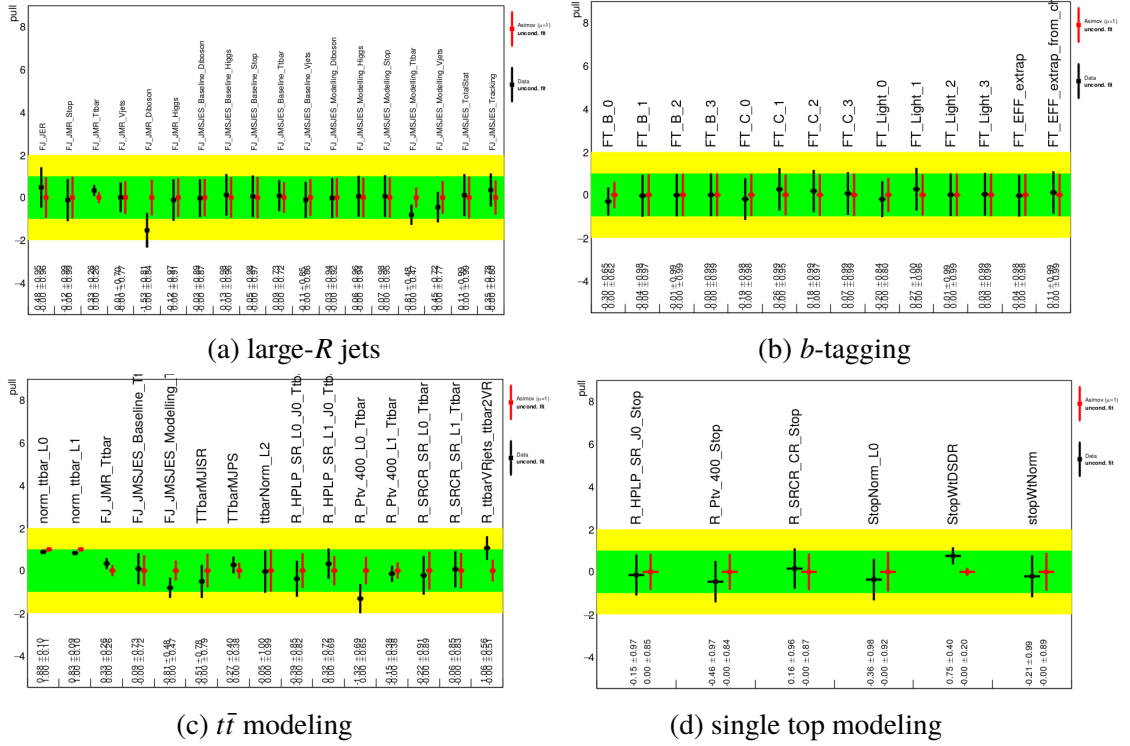


Figure 5.20: NP pulls and constraints for the most important systematic uncertainties: (a) large- R jets related systematic uncertainties; (b) b -tagging related systematic uncertainties; (c) $t\bar{t}$ modeling uncertainties; (d) single top modeling uncertainties, obtained from unconditional fit to the Asimov data (red) and the actual data (black).

pulled down because of the under-fluctuation of data in high p_T^V top CR. The pulls for NPs related to single top modeling uncertainties are shown in Figure 5.20 (d). The NP related to DS-DR variations is pulled by around 0.7σ because the fit tends to prefer the DS scheme in the alternative samples over the DR scheme in the nominal samples. The other NPs are pulled very little.

Correlation matrices

The correlation matrix shows the correlation between every pairs of the POIs and the NPs. It is obtained from fit to both the asimov data and the actual data. Figure 5.21 shows the correlation matrix obtained from the unconditional fit to the Asimov data, and Figure 5.22 shows the correlation matrix obtained from the unconditional fit to the actual data. Only the NPs that have at least one correlation larger than 0.25 with another NP are shown.

The correlation between the VH signal strength and VZ signal strength is 9% in the fit to Asimov data and 11% in the fit to the actual data. The correlation matrices look rather diagonal, and the correlations obtained from fit to the Asimov data and the actual data are similar. However, several blocks can be identified. The first one to mention concerns the floating normalization and the b -tagging uncertainties. The NP related to b -tagging uncertainty of c -jet (FT_C_1) is highly correlated to the $t\bar{t}$ normalization factors (norm_ttbar_L0/1) because the c -jet can be mis-tagged as a b -jet in about

5.4 Statistical analysis

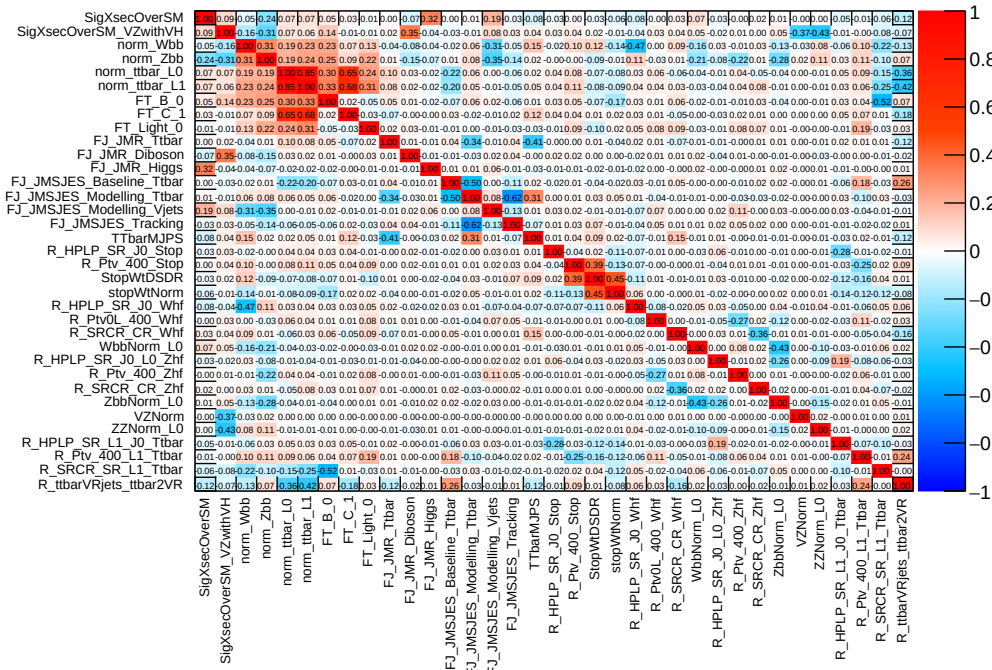


Figure 5.21: Correlation matrix from unconditional fit to the Asimov data. Only NPs that have at least one correlation larger than 0.25 with another NP are shown.

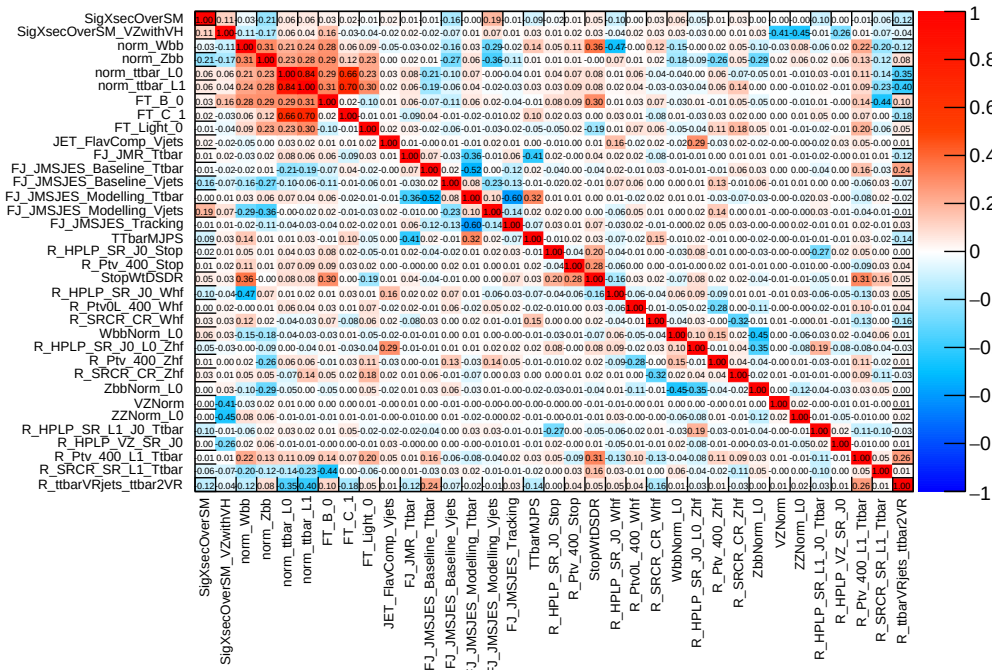


Figure 5.22: Correlation matrix from unconditional fit to the actual data. Only NPs that have at least one correlation larger than 0.25 with another NP are shown.

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

70% of the large- R jets in the $t\bar{t}$ events. The NP related to b -tagging uncertainty for b -jet (FT_B_0) is correlated to the normalization factors of $t\bar{t}$, W +HF, and Z +HF processes. The floating normalization factors themselves are also correlated. Another group of correlated NPs are related to the large- R jet uncertainties for $t\bar{t}$ samples. The JMR NP shows a high anti-correlation with the JES and JMS NPs. This correlation is expected because all these NPs have roughly the same effect on the top mass peak.

In summary, most of the correlations between different NPs are small. The high correlations observed in NPs related to the floating normalization, the b -tagging uncertainties and the large- R jet uncertainties are understandable.

Nuisance parameter ranking

The maximum profile likelihood fit simultaneously extracts the signal strength μ and the nuisance parameters θ . The fit adjusts the estimation of the NPs from the pre-fit nominal value θ_0 to the post-fit value $\hat{\theta}$. An important validation of the fit is to quantify the impact that a NP has on the signal strength. The impact of a NP on the signal strength is defined as:

$$\text{impact} = \hat{\mu}(\hat{\theta} \pm \Delta_{\theta}) - \hat{\mu}, \quad (5.12)$$

where $\hat{\mu}$ is the signal strength that maximize the likelihood, $\hat{\mu}(\hat{\theta} + \Delta_{\theta})$ is the signal strength that maximize the likelihood when the value of θ varies by $\pm\Delta_{\theta}$ around the post-fit value. The impact quantifies how much the signal strength varies when the value of a NP changes. Figure 5.23 shows the ranking of the NPs based on their impact on the signal strength.

The impact in the plots are calculated by varying the value of each NP by $\pm\sigma$. The top 15 NPs with the highest impact are shown. The dashed and the white boxes show the impact of the NPs on μ with $1\sigma_{\hat{\theta}}$ up and down variations, where $\sigma_{\hat{\theta}}$ is the post-fit uncertainty of the NP. The yellow bands show the impact of the NPs on signal strength with 1σ up and down variations, where σ is the pre-fit uncertainty with a value of 1. Both the two kinds of impacts are displayed in the top x -axis. The black points with error bars show the pulls and constraints of the NPs, and they are displayed in the bottom x -axis. The open red circles with error bars represent the floating normalizations and their uncertainties.

The ranking of the NPs agrees with the correlation matrices. As expected, these highest ranking NPs also show the highest correlations to the signal strength.

Breakdown of the systematic uncertainties

The impact of different systematic uncertainties can be studied in combination. Systematic uncertainties from similar sources are grouped together, and their combined impact are studied. This method can tell the uncertainties that limit the precision of the measurement the most, which provides directions of improvements in future analyses. The impact of a group of uncertainties is obtained by comparing

5.4 Statistical analysis

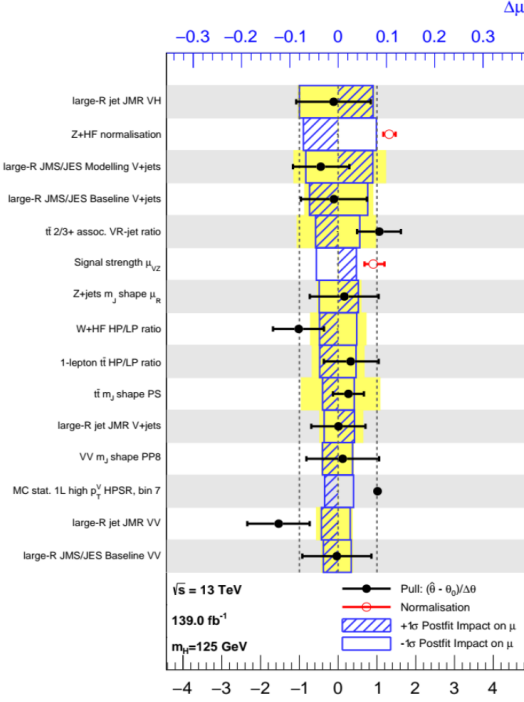


Figure 5.23: Ranking plots for the impact of NPs on signal strength μ_{VH} . The black points with error bars are the pulls and the constraints. The open red circles with error bars are the pulls and constraints of the floating normalizations with normalized value at 1. The boxes are the NP impacts on μ , ranked in decreasing order. The yellow bands show the NP impacts with ± 1 variation.

the uncertainty of the signal strength obtained from the nominal fit and the uncertainty of the signal strength obtained from a fit with the NPs in the group fixed to $\hat{\theta}$.

$$\text{impact} = \sqrt{\sigma_{\mu}^2 - \sigma_{\mu, \theta=\hat{\theta}}^2}, \quad (5.13)$$

where σ_{μ} is the uncertainty of the signal strength obtained from the nominal fit with all the NPs and the POIs floated in the fit; $\sigma_{\mu, \theta=\hat{\theta}}$ is the uncertainty of the signal strength obtained from a fit with the NPs in the group fixed to their best-fit value $\hat{\theta}$. For example, all the NPs are fixed to their best-fit value when study the impact of the total statistical uncertainties. The impact of the floating normalizations can be calculated by fixing the normalization factors to their best-fit values, while other NPs are still floated in the fit.

5.5 Results of boosted $VH(H \rightarrow b\bar{b})$ production

The results of the $VH(b\bar{b})$ boosted analysis are obtained from the binned maximum profile likelihood fit and the profile log-likelihood ratio test using m_J as the final discriminant. This section presents the fitted signal strength and the cross section measurements. The cross sections measured within the STXS framework are further interpreted using an effective field theory (EFT) to study anomalous couplings.

5.5.1 Post-fit m_J distributions

The distributions of m_J in the signal regions and control regions are all used as inputs to the fit. The signal strengths and the NPs are changed from their nominal values to maximize the likelihood. The post-fit yields are also changed with the best-fit values of the signal strengths and the NPs,

$$\hat{\mu} \cdot s_i \cdot \nu_s(\hat{\theta}) + \hat{\gamma}_i \cdot b_i \cdot \nu_b(\hat{\theta}). \quad (5.14)$$

As defined in Section 5.4.1, $\hat{\mu}$ is the best-fit value of the signal strength, $\hat{\theta}$ and $\hat{\gamma}_i$ are the best-fit values of the NPs, s_i is the pre-fit signal yields and b_i is the pre-fit background yields, ν_s and ν_b are parameterizations of the effects of the NPs on the signal and background yields. Figure 5.24, Figure 5.25 and Figure 5.26 show the post-fit m_J distributions in 0L channel, 1L channel and 2L channel, respectively. The data events are shown as the black dots, while the post-fit signal and background events are shown as the histograms. The signal strength, the NPs and the normalization factors are all set to their best-fit values. The agreement between data and simulated samples is good in all the signal regions and control regions.

5.5.2 Signal strength and significance

The parameters of interest (POIs) in the default fit are the VH signal strength μ_{VH} and the VZ signal strength μ_{VZ} . The signal strengths are extracted simultaneously in the fit to maximize the likelihood. For a Higgs boson mass of 125 GeV, the measured VH signal strengths when all three channels are combined is:

$$\mu_{VH} = 0.72^{+0.39}_{-0.36} = 0.72^{+0.29}_{-0.28}(\text{stat.})^{+0.26}_{-0.22}(\text{syst.}). \quad (5.15)$$

The total statistical uncertainties are calculated with all the NPs fixed to their best-fit value. The total systematic uncertainties are calculated as shown in Eq. 5.13. The observed excess with respect to the background-only hypothesis has an observed significance of 2.1 standard deviations calculated from unconditional fit to the actual data, compared to an expected significance of 2.7 standard deviations calculated from conditional fit to the data at $\mu_{VH} = \mu_{VZ} = 1$.

The measured VZ signal strength is

$$\mu_{VZ} = 0.91^{+0.29}_{-0.23} = 0.91 \pm 0.15(\text{stat.})^{+0.24}_{-0.17}(\text{syst.}). \quad (5.16)$$

The VZ process has an observed significance of 5.4 standard deviations, compared to an expected significance of 5.7 standard deviations. Figure 5.27 shows the distribution of m_J summed over all

5.5 Results of boosted $VH(H \rightarrow b\bar{b})$ production

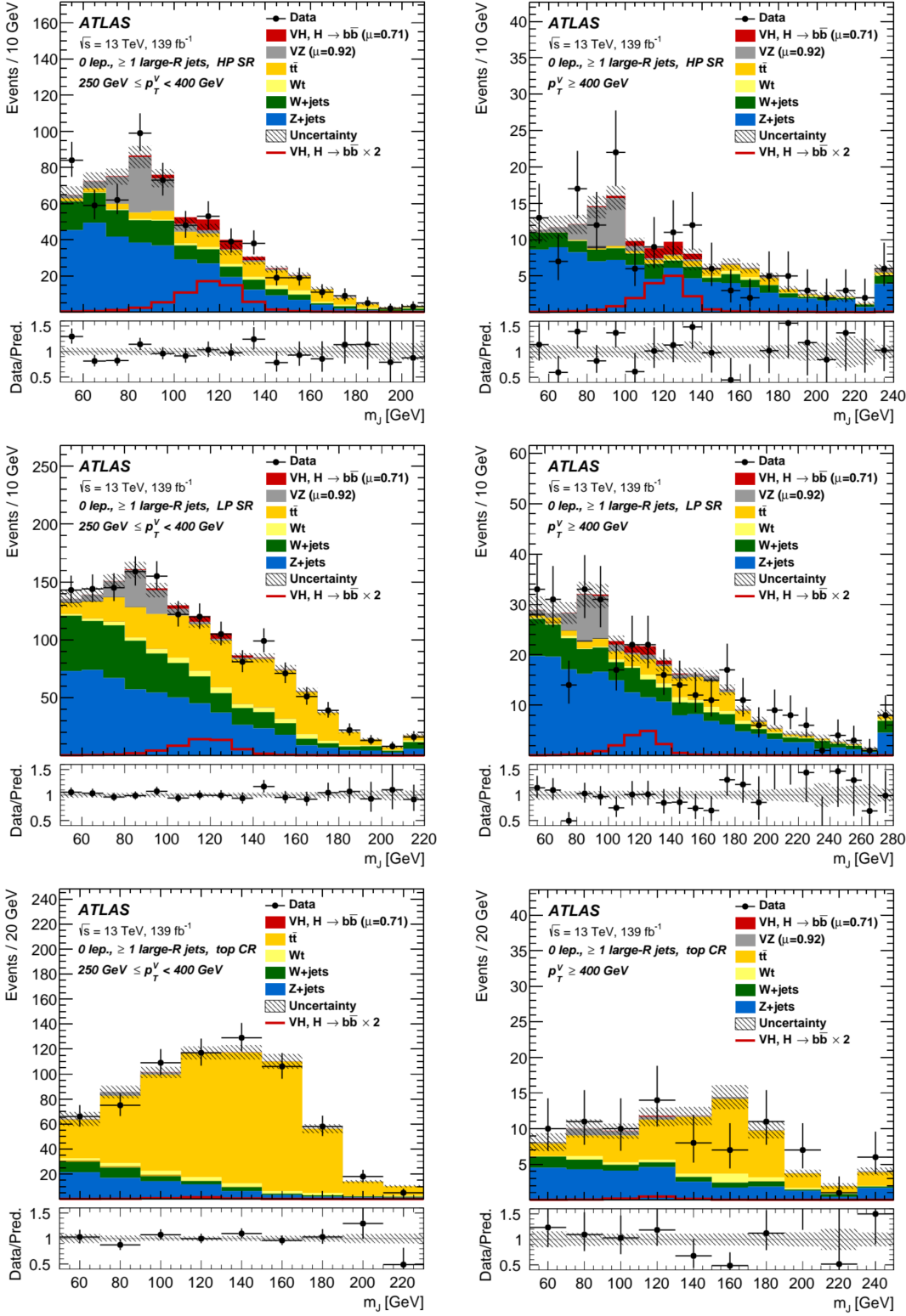


Figure 5.24: Post-fit m_J distributions in 0L channel in the HP SRs (top), LP SRs (middle) and CRs (bottom) for $250 < p_T^V < 400$ GeV (left) and $p_T^V > 400$ GeV (right) regions [140]. The black dots are data. The simulated events are shown as the histograms, with the signal strength, the NPs and the normalization factors all set to their best-fit values.

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

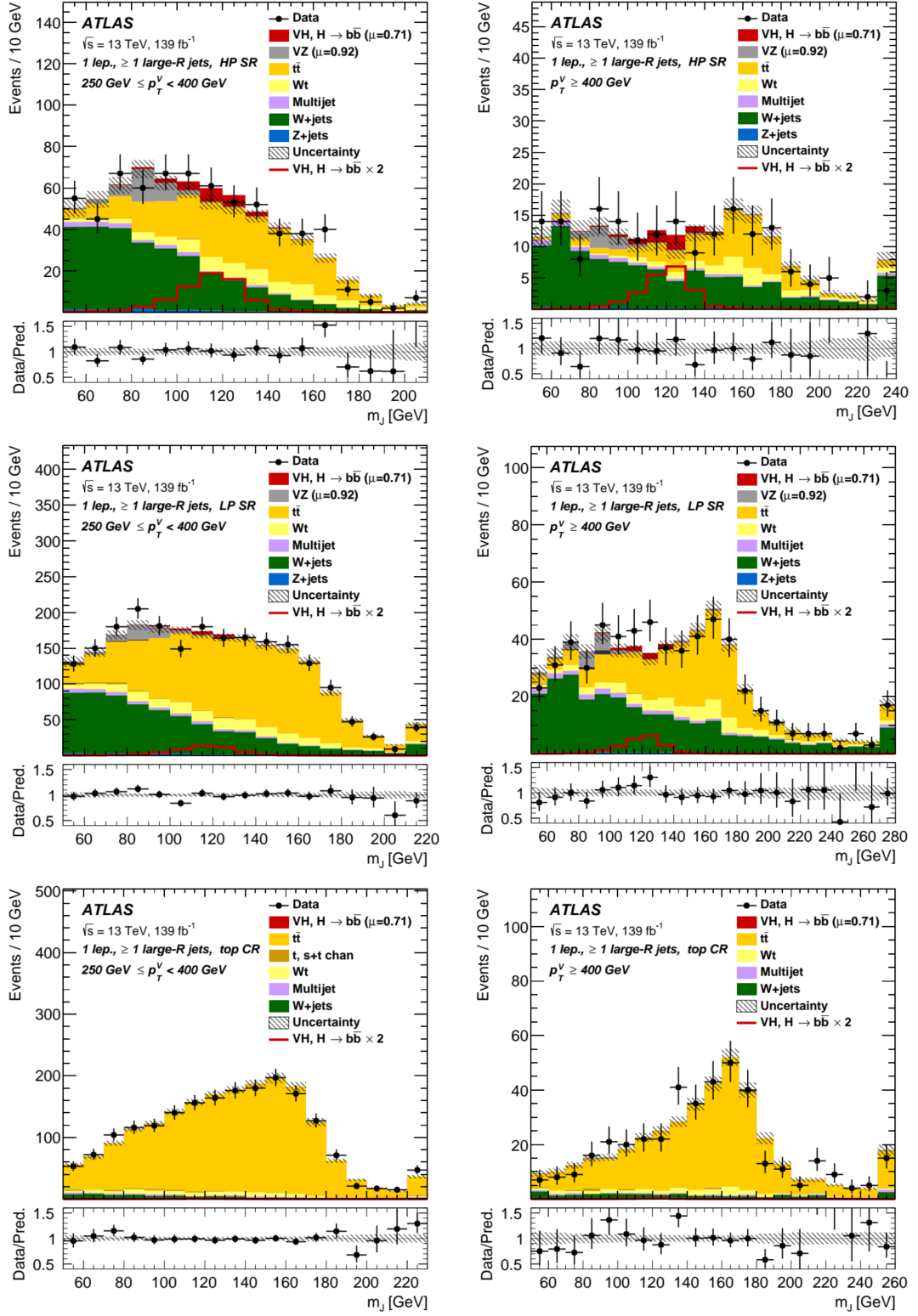


Figure 5.25: Post-fit m_J distributions in 1L channel in the HP SRs (top), LP SRs (middle) and CRs (bottom) for $250 < p_T^V < 400$ GeV (left) and $p_T^V > 400$ GeV (right) regions [140]. The black dots are data. The simulated events are shown as the histograms, with the signal strength, the NPs and the normalization factors all set to their best-fit values.

5.5 Results of boosted $VH(H \rightarrow b\bar{b})$ production

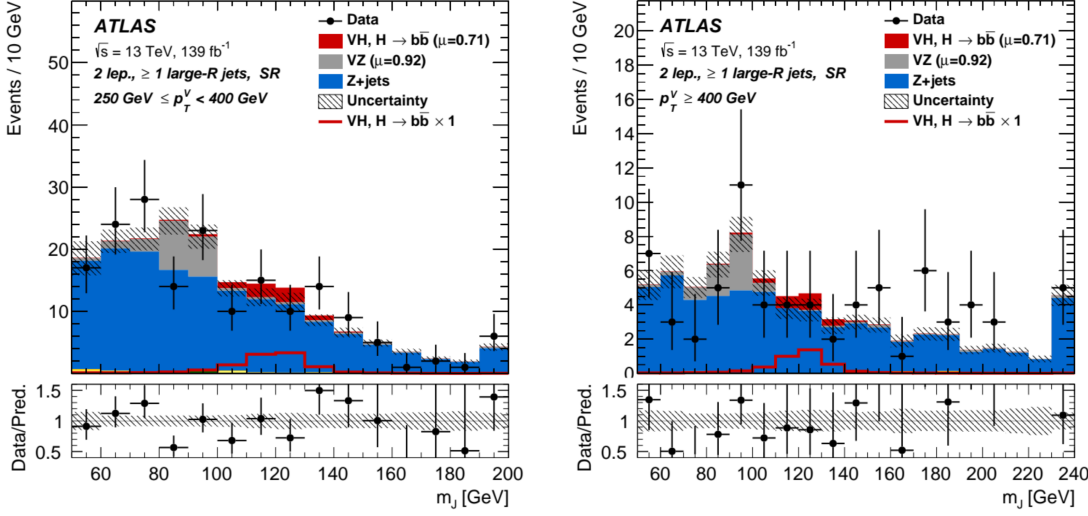


Figure 5.26: Post-fit m_J distributions in 2L channel for $250 < p_T^V < 400 \text{ GeV}$ (left) and $p_T^V \geq 400 \text{ GeV}$ (right) regions [140]. The black dots are data. The simulated events are shown as the histograms, with the signal strength, the NPs and the normalization factors all set to their best-fit values.

the signal regions after subtracting all the backgrounds except for the VZ process. The events are weighted by S/B , which is the respective values of the fitted VH signal and background yields ratio.

Both the measured VH and VZ signal strengths are compatible with the SM predictions. In the result of the VH signal strength measurement, the impact of the systematic uncertainties is at the same level as the total statistical uncertainties. Table 5.23 shows the breakdown of the impact of different sources of uncertainties on the measured signal strength. The impact of the total statistical uncertainties includes the impact from the statistical uncertainty itself and the floating normalizations. The statistical component is dominated by the limited size of the samples and it is more than twice as large as the impact from the normalizations. The systematic uncertainties can be further broken down into experimental uncertainties and theoretical uncertainties. The theoretical and modeling uncertainties for different processes have similar impact on the signal strength. Among the experimental uncertainties, the large- R jet uncertainties have the most impact on the signal strength, in particular the m_J resolution. The large- R jet induced uncertainties also have the dominant impact among all the systematic uncertainties. Therefore, the analysis can benefit from large- R jet collection with better resolution.

5.5.3 Cross section measurements

The result is also presented as a cross section measurement within the Simplified Template Cross Section (STXS) framework described in Section 1.3.1. The fit in the STXS framework has similar procedure as the default one. However, the POIs are cross sections in the STXS bins instead of the signal strengths. The definition of the STXS bins used in $VH(b\bar{b})$ boosted analysis is referred to as

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

Source of uncertainty	Impact on error	Avg. impact
Total	+0.388/−0.356	0.372
Statistical	+0.286/−0.280	0.283
↔ Data stat. only	+0.251/−0.245	0.248
↔ Floating normalizations	+0.096/−0.092	0.094
Systematic	+0.261/−0.219	0.240
Experimental uncertainties		
Lepton	+0.013/−0.007	0.010
E_T^{miss}	+0.008/−0.007	0.007
Small- R jets	+0.041/−0.034	0.038
Large- R jets	+0.161/−0.105	0.133
FTAG: b -jet	+0.028/−0.004	0.016
FTAG: c -jet	+0.012/−0.011	0.012
FTAG: light flavor jet	+0.009/−0.007	0.008
FTAG: extrapolation	+0.004/−0.005	0.004
Pile-up	+0.001/−0.002	0.001
Luminosity	+0.019/−0.007	0.013
Theoretical and modeling uncertainties		
Signal	+0.057/−0.019	0.038
Backgrounds	+0.106/−0.095	0.100
↔ single top	+0.031/−0.024	0.027
↔ $t\bar{t}$	+0.037/−0.032	0.035
↔ W +jets	+0.059/−0.057	0.058
↔ Z +jets	+0.049/−0.047	0.048
↔ Diboson	+0.035/−0.029	0.032
↔ Multijet	+0.009/−0.009	0.009
MC statistical	+0.091/−0.092	0.092

Table 5.23: Breakdown of the impact of the systematic uncertainties on the measured signal strength μ_{VH} . The sum in quadrature of the systematic uncertainties attached to the categories differs from the total systematic uncertainty due to correlations.

5.5 Results of boosted $VH(H \rightarrow b\bar{b})$ production

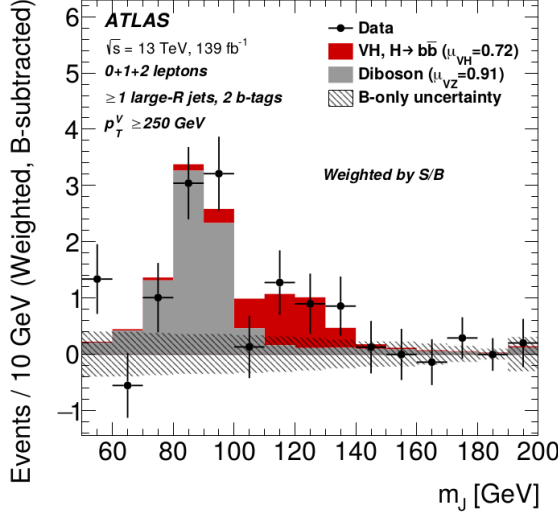


Figure 5.27: m_J distribution summed over all the signal regions after subtracting all the backgrounds except for WZ and ZZ processes [140]. The VH signal and the VZ process are shown as their post-fit distributions with the signal strengths and the NPs set to their best-fit values. The total statistical and systematic uncertainties of all the other backgrounds are shown as the hatched band.

the main splitting:

$$\begin{aligned} ZH, p_T^V &\in [250, 400] \text{ GeV}, [400, +\infty] \text{ GeV}, \\ WH, p_T^V &\in [250, 400] \text{ GeV}, [400, +\infty] \text{ GeV}. \end{aligned}$$

No splitting on the number of additional small- R jets is applied. Each STXS bin has a POI corresponding to the cross section in the region, multiplied by the branching ratio $\mathcal{B}_{H \rightarrow b\bar{b}}$. The expected signal distribution in each STXS bin is estimated from the simulated signal samples using truth information, in particular the truth p_T^V . The VZ process is still floating in the fit and its signal strength is extracted simultaneously. The systematic uncertainties are identical to those defined in Section 5.3.

The measured VH cross section in the STXS bins and their SM predictions are shown in Figure 5.28. The black dots are the measured cross sections, with the black error bars showing the total uncertainties and the green error bars showing the statistical uncertainties. The red lines are the SM predicted cross sections. The main uncertainty of the cross section measurement is statistical, especially in the high p_T^V regions. The values of the measured and the predicted cross sections are summarized in Table 5.24. All the measurements are in agreement with the SM predictions.

5.5.4 Effective Field Theory interpretations

The STXS results are further interpreted in an effective field theory (EFT) approach, introduced in Section 1.3.2. In the EFT approach, the SM is considered as a low energy approximation of a more generic model and higher dimensional operators are added to the SM Lagrangian. The cross

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

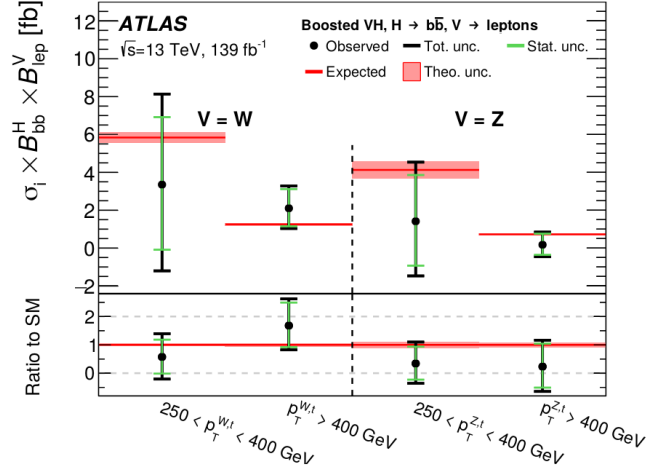


Figure 5.28: Measured VH signal cross section multiplied by $\mathcal{B}_{H \rightarrow b\bar{b}} \cdot \mathcal{B}_{V \rightarrow \text{leptons}}$ in the STXS bins [140].

STXS bin	SM prediction [fb]	measured	Tot.	Stat.	Syst. [fb]
$WH, p_T^{W, \text{true}} \in [250, 400] \text{ GeV}$	5.83 ± 0.26	3.3	+4.8 -4.6	+3.6 -3.4	+3.2 -3.0
$WH, p_T^{W, \text{true}} \in [400, \infty] \text{ GeV}$	1.25 ± 0.06	2.1	+1.2 -1.1	+1.0 -0.9	+0.6 -0.5
$ZH, p_T^{Z, \text{true}} \in [250, 400] \text{ GeV}$	4.12 ± 0.45	1.4	+3.1 -2.9	+2.4 -2.3	+1.9 -1.7
$ZH, p_T^{Z, \text{true}} \in [400, \infty] \text{ GeV}$	0.72 ± 0.05	0.2	+0.7 -0.6	+0.6 -0.5	+0.3 -0.3

Table 5.24: Measured and predicted VH cross sections in STXS bins, multiplied by the branching ratio of $H \rightarrow b\bar{b}$ and $V \rightarrow \text{leptons}$ [140].

section and the branching ratio predicted by the EFT model can be related to the Wilson coefficients as shown in Eq. 1.57. The Wilson coefficients are zero in the SM. The constraints on these Wilson coefficients can be extracted from fitting to the STXS measurement, using the procedure described in Ref. [63]. In total, 17 operators have effects on $VH(b\bar{b})$ process at LO, as listed in Table 1.3. Due to the limited number of the STXS bins, not all the Wilson coefficients of these operators can be extracted simultaneously. Therefore, four of the Wilson coefficients (c_{HWB} , c_{HW} , c_{Hu} , c_{Hq3}) that the analysis is most sensitive to are studied. Additionally, the coefficient $|c_{dH}|$, which affects $H \rightarrow b\bar{b}$ decay width directly, is also studied. Each of the five coefficients is varied individually, with the other coefficients set to their SM value of zero.

The result of the constraints on the five Wilson coefficients is shown in Figure 5.29. The linear terms come from the interference between the SM and the BSM processes, while the linear and quadratic terms include pure BSM processes. The confidence intervals are shown at 68% confidence level (CL) with solid lines and 95% CL with dashed line. The measured coefficients are compatible with the SM value of zero at 68% CL.

The other approach extracts a set of four linear combinations of all the Wilson coefficients simultaneously. These combinations, referred to as eigenvector combinations, are linear combinations of

5.5 Results of boosted $VH(H \rightarrow b\bar{b})$ production

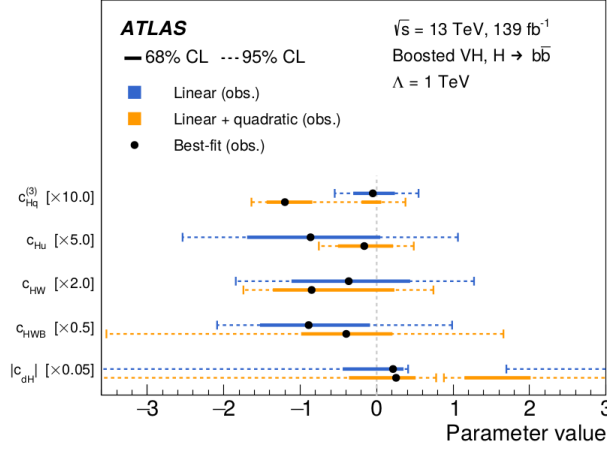


Figure 5.29: Constraints on the Wilson coefficients c_{HWB} , c_{HW} , c_{Hu} , c_{Hq3} and $|c_{dH}|$. The black dots are the observed best-fit values, the solid lines are confidence intervals at 68% CL, the dashed lines are confidence intervals at 95% CL [140].

Coefficient	Eigenvector combination
c_{EA}	$0.99c_{Hq}^{(3)} + 0.11c_{Hu}$
c_{EB}	$0.82c_{Hu} - 0.49c_{Hq}^{(1)} - 0.24c_{Hd} - 0.13c_{Hq}^{(3)}$
c_{EC}	$0.67\mathcal{I}_{BR} + 0.66c_{HW} + 0.18c_{Hq}^{(1)} - 0.16c_{Hl}^{(3)} + 0.14c_{HWB} + 0.12c_{ll}^{(1)}$
c_{ED}	$0.70c_{Hq} + 0.52c_{HWB} + 0.27c_{Hu} - 0.27c_{HW} - 0.24c_{Hd} + 0.13c_{HB}$

Table 5.25: Linear combinations of the main Wilson coefficients according to the principal component decomposition eigenvectors.

the Wilson coefficients based on the eigenvectors of the principal component decomposition. The combinations are shown in Table 5.25, expressed in terms of the main Wilson coefficients and an additional independent parameter $\mathcal{I}_{BR}$ for $H \rightarrow b\bar{b}$ branching ratio. The fit result of these linear combinations is shown in Table 5.26. No significant deviation from the SM prediction is observed.

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

Coefficient	Expected	Observed
c_{EA}	$0.000^{+0.030}_{-0.027}$	$-0.010^{+0.027}_{-0.025}$
c_{EB}	$0.00^{+0.20}_{-0.19}$	$-0.21^{+0.19}_{-0.20}$
c_{EC}	$0.00^{+0.71}_{-0.67}$	$-0.62^{+0.70}_{-0.66}$
c_{ED}	$0.0^{+2.8}_{-2.7}$	$0.4^{+2.8}_{-2.7}$

Table 5.26: Expected and observed values of the coefficients corresponding to the eigenvector combinations, with 68% CL [140].

5.6 Future improvements of $H \rightarrow b\bar{b}$ measurements

The $VH(b\bar{b})$ resolved and boosted analyses have been carried out independently with full Run 2 dataset. The next step naturally comes to the combination of the two analysis. The $VH(b\bar{b})$ legacy analysis aims to design one coherent, harmonized approach to extract the results of the stand-alone $VH(b\bar{b})$ resolved and boosted analyses. The phase space of the two analyses significantly overlaps and the combination is not straight forward. To combine the two analyses, one of the method is to select the more suitable topology for every single event, which is also the most complicated combination. Simpler approach uses p_T^V cuts to decide the phase space for resolved and boosted analyses. The easiest combination is to use the resolved analysis strategy for $p_T^V < 400$ GeV, and the boosted analysis strategy for $p_T^V > 400$ GeV. The $VH(b\bar{b})$ legacy analysis benefits from the improved performance like the improved small- R jet resolution, better b -tagging, as well as several improvements in the analysis strategy. This section describes the optimization of the analysis strategy that I have personally contributed to.

5.6.1 Selection of Higgs boson candidate

As described in Section 5.2.1, the events are required to have at least two track jets being ghost-associated to the leading large- R jet, and the two leading track jets have to be b -tagged at 70% efficiency in the stand-alone $VH(b\bar{b})$ boosted analysis. This tagging strategy was optimized when the final discriminant is the large- R jet mass. However, this strategy is no longer the optimal one with the introduction of the multivariate analysis (MVA). The tagging strategies that have been studied in the legacy analysis are

- Leading2: the two leading track jets matched to the large- R jet have to be b -tagged.
- Leading3: exactly two track jets in the three leading track jets matched to the large- R jet have to be b -tagged. If the event only has two matched track jets, then both of the two jets have to be b -tagged.

The "Leading2" strategy is the same as the tagging strategy used in the stand-alone $VH(b\bar{b})$ boosted analysis. The signal regions in 0L channel and 1L channel still require no b -tagged track jets outside the leading large- R jet, and the top control regions are required to have at least one additional b -tagged track jet.

The b -tagging algorithm applied in the $VH(b\bar{b})$ legacy analysis is the DL1 [135, 136], which is a tagger based on a deep neural network (DNN). The DL1 tagger has an multidimensional output corresponding to the probabilities for a jet to be a b -jet, a c -jet or a light-flavor jet. The final DL1 b -tagging discriminant is defined as:

$$\text{DL1r} = \ln \left(\frac{p_b}{f_c \cdot p_c + (1 - f_c) \cdot p_{\text{light}}} \right), \quad (5.17)$$

where p_b , p_c and p_{light} are the b -jet, c -jet and light-flavor jet probabilities, respectively; f_c is the effective c -jet fraction in the sample used in the training. The distribution of the DL1 b -tagging discriminant is shown in Figure 5.30. Similar to the MV2 discriminant, the DL1 discriminant

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

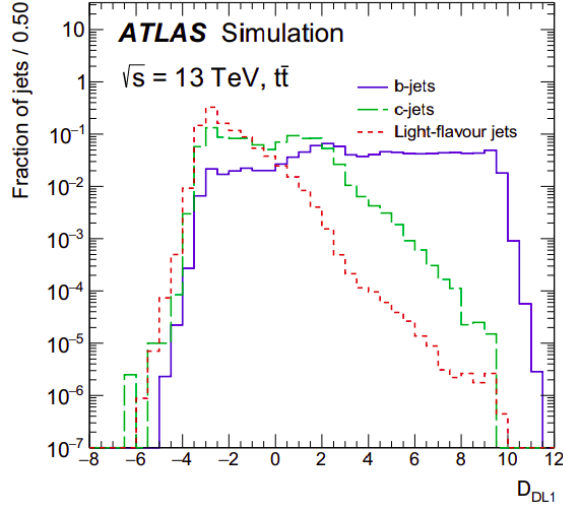


Figure 5.30: Distribution of the b -tagging discriminant from the DL1 algorithm for b -jets, c -jets and light-flavor jets in simulated $t\bar{t}$ events [135].

Efficiency	Cut value	Rejections		
		c -jet	τ -jet	light jet
60%	> 2.74	27	220	1300
70%	> 2.02	9.4	43	390
77%	> 1.45	4.9	14	130
85%	> 0.46	2.6	3.9	29

Table 5.27: Background jets rejection rates for different DL1 b -tagging efficiency working points, computed from simulated $t\bar{t}$ samples [135].

distributions are also divided into five bins, with boundaries defined at b -tagging efficiency for 85%, 77%, 70%, 60%, bounded by the trivial 100% and 0% selections. Table 5.27 summarizes the background jet rejection rates for different working points.

Different b -tagging efficiencies are tested with both the "Leading2" and the "Leading3" tagging strategies. The distributions of m_J in signal regions with the "Leading2" and the "Leading3" tagging strategy at different b -tagging WPs are shown in Figure 5.31. The "Leading3" tagging strategy combined with the b -tagging WP at 85% allows the most events passing the event selection, which is a benefit for the MVA training. Table 5.28 summarizes the significance evaluated from distributions of m_J and MVA discriminant with different b -tagging WP applied to track jets matched to the large- R jet for "Leading2" strategy. The values of the significance are calculated using the formula

$$Z = \sqrt{\sum_{i \in \text{bins}} 2((s_i + b_i) \cdot \ln(1 + s_i/b_i) - s_i)}, \quad (5.18)$$

5.6 Future improvements of $H \rightarrow b\bar{b}$ measurements

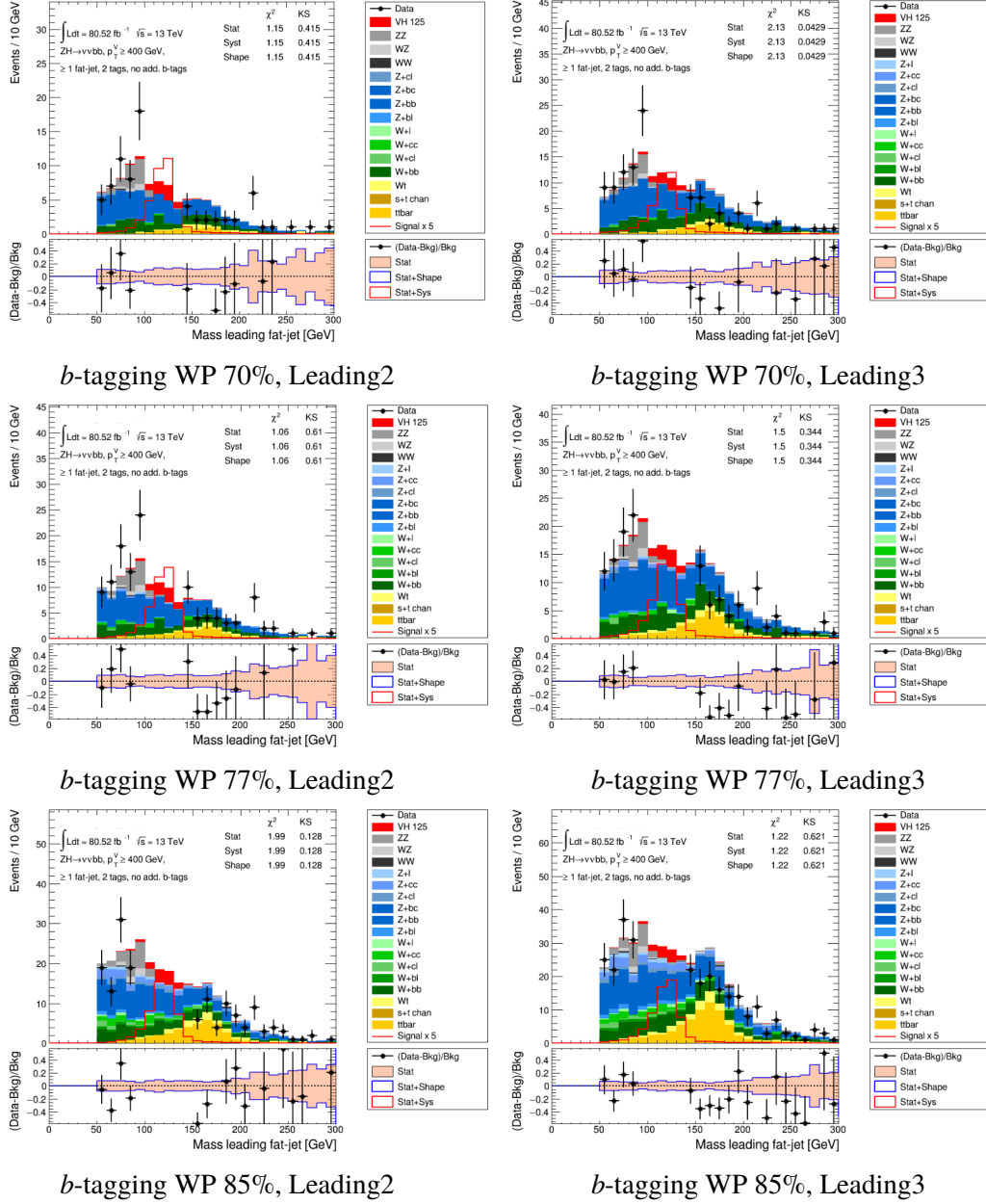


Figure 5.31: Distributions of m_l in high purity signal regions with the "Leading2" (left) and the "Leading3" (right) tagging strategy for b -tagging WPs at 70%, 77% and 85%.

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

Leading2 strategy				
b-tagging WP	$p_T^V > 400$ GeV HP SR		$p_T^V > 400$ GeV LP SR	
	m_J	MVA	m_J	MVA
70%	1.22	1.64	0.82	1.09
77%	1.27	1.87	0.83	1.19
85%	1.32	1.93	0.77	1.26

Table 5.28: Significance evaluated from distributions of m_J and MVA discriminant with different b -tagging WPs for track jets matched to the leading fat jet. All results are obtained with "Leading2" strategy.

b -tagging WP at 85%		
	Leading2	Leading3
HP SR	2.01	2.04
LP SR	1.32	1.35
inclu. SR	2.42	2.48

Table 5.29: Significance calculated from BDT output distributions for Leading2 and Leading3 strategies in $p_T^V > 400$ GeV HP SR, LP SR or in one inclusive $p_T^V > 400$ GeV SR, 0L channel.

where s_i and b_i are signal and background yields in each bin of the distribution. For the results in Table 5.28, the b -tagging WP applied to track jets not matched to the leading large- R jet are all set to 70%. The significance improved over 40% with the MVA output as the final discriminant. The b -tagging efficiency at 85% together with the MVA give the largest overall significance among all the working points for "Leading2" strategy. The b -tagging WP applied to track jets not matched to the leading large- R jet is also studied. With the b -tagging efficiency for matched track jets fixed at 85%, the b -tagging efficiency for unmatched track jets has very little impact on the significance.

The "Leading3" strategy is also studied with different b -tagging efficiencies. The significance evaluated using MVA discriminant is constantly a little larger with the "Leading3" strategy compared to the "Leading2" strategy, as the "Leading3" strategy provides more information to the MVA. Table 5.29 shows the comparison of the two strategies when the b -tagging efficiency is set to 85%. All the significance values in the table are calculated using the formula from the MVA discriminant distributions. The small difference between Table 5.29 and Table 5.28 comes from the optimization of the MVA training.

Finally, the default Higgs tagging strategy for boosted topology in the $VH(b\bar{b})$ legacy analysis requires:

- At least two track jets matched to the leading large- R jet through ghost association.
- Exactly two track jets in the three leading matched track jets being b -tagged at 85% efficiency.

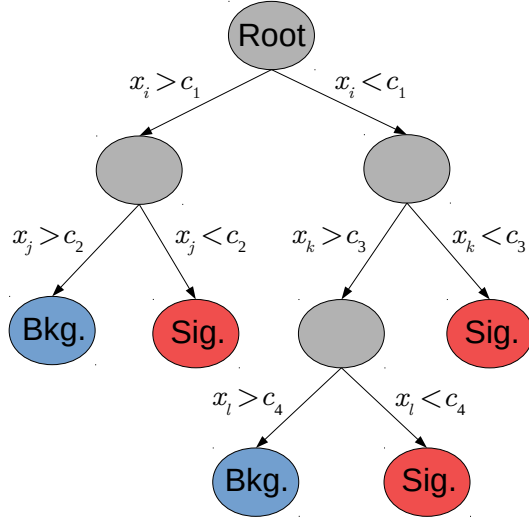


Figure 5.32: Sketch of a decision tree with three layers. The decision starts from the root node on variable x_i , the samples are divided to two categories. The procedure is repeated at every node, until each event is decided as signal-like or background-like.

5.6.2 Multivariate discriminants

The multivariate analysis (MVA) is introduced to boosted topology in the $VH(b\bar{b})$ legacy analysis. A set of boosted decision trees (BDT) are trained to discriminant the VH signal from the background processes. The BDT is a collection of decision trees that are combined with weights. The MVA is provided with a training sample with simulated signal and background events, each event has a set of variables. The structure of a decision tree is illustrated in Figure 5.32. Each decision tree is constructed with several nodes. The events are split based on one of the input variables at each node. The split at each node is chosen to maximize information gain or minimize entropy. The information gain is based on the decrease of entropy after the split. The entropy is a measurement of purity in the sample. It is maximized for equal mix of signal and background events; minimized for nodes with only signal events or only background events. The terminal nodes, also known as "leaves", represent a class label or a probability. In this case, it is signal or background. Technically, a decision tree can grow as much as it could. However, the resulting BDT would be extremely over-trained, which means the training would perform significantly worse if applied to an independent testing sample. The configuration of the BDT needs to achieve a balance between the separation power and the over-training.

The BDT combines many decision trees into a strong classifier through techniques like adaptive boosting (AdaBoost). The decision trees are created iteratively. After each iteration of the training process, the misclassified events in the previously trained decision tree are given higher weights. Then all the events with higher weights are given more importance in the next iteration. The initial event weight is $w_{i,1} = 1/N$, where N is the total number of events. After the m -th iteration, the event weight

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

is updated to

$$w_{i,m+1} = w_{i,m} e^{\alpha_m I(y_i \neq h(x_i))}, \quad (5.19)$$

where y_i is the desired output for event i , $h(x_i)$ is the output from the weak learner (one decision tree), $I(y_i \neq h(x_i))$ is the misclassification rate of the decision tree, α_m is the weight of the tree defined as:

$$\alpha_m = \ln \frac{1 - \epsilon_m}{\epsilon_m}, \quad (5.20)$$

ϵ_m is the weighted misclassification rate calculated as $\epsilon_m = \sum_i w_{i,m} \cdot I(y_i \neq h(x_i)) / \sum_i w_{i,m}$. The final BDT discriminant becomes

$$\hat{h}(x_i) = \text{sign} \left[\sum_{m=1}^M \alpha_m h_m(x_i) \right]. \quad (5.21)$$

The BDTs are trained in merged signal regions in $p_T^V > 400$ GeV and $250 \text{ GeV} < p_T^V < 400$ GeV. The high purity SR and the low purity SR in each p_T^V region are merged as the number of additional small- R jet is used as an input variable to the MVA.

The MVA is trained after applying the baseline event selection and the Higgs tagging selection. The input variables to the training are:

- large- R jet mass m_J .
- Missing transverse momentum E_T^{miss} .
- $p_T^{\text{track jet 1}}, p_T^{\text{track jet 2}}, p_T^{\text{track jet 3}}$: transverse momentum of the three leading track jets matched to the leading large- R jet. If an event only has two matched track jets, p_T of the third track jet is set to -1.
- DL1(b -jet 1), DL1(b -jet 2): binned DL1 b -tagging discriminant for the two b -tagged track jets. The discriminant distributions are divided into five bins as described in the previous section.
- $\Delta\phi(E_T^{\text{miss}}, \text{large-}R \text{ jet})$: angular distance between E_T^{miss} and the large- R jet.
- $\Delta R(b\text{-jet 1}, b\text{-jet 2})$: distance between the two b -tagged track jets.
- number of small- R jets not matched to the leading large- R jet.
- number of track jets matched to the leading large- R jet.

The selection of these input variables is based on the variables used in event selection and categorization. The large- R jet mass is the final discriminant of the $VH(b\bar{b})$ boosted analysis. E_T^{miss} , track jet p_T , $\Delta\phi(E_T^{\text{miss}}, \text{large-}R \text{ jet})$ and $\Delta R(b\text{-jet 1}, b\text{-jet 2})$ have been proved useful in the baseline event selection of the $VH(b\bar{b})$ boosted and resolved analyses. Number of small- R jet not matched to the leading large- R jet and number of track jets matched to the leading large- R jet are used in the event categorization of the $VH(b\bar{b})$ boosted analysis. The binned DL1 b -tagging discriminants have been tested in the training. The BDT trained with these two variables is able to improve the significance of the analysis by more than 10%.

The MVA parameters have been optimized in a wide range to maximize the significance and keep the over-training small at the same time. The optimized configuration is found as the following:

5.6 Future improvements of $H \rightarrow b\bar{b}$ measurements

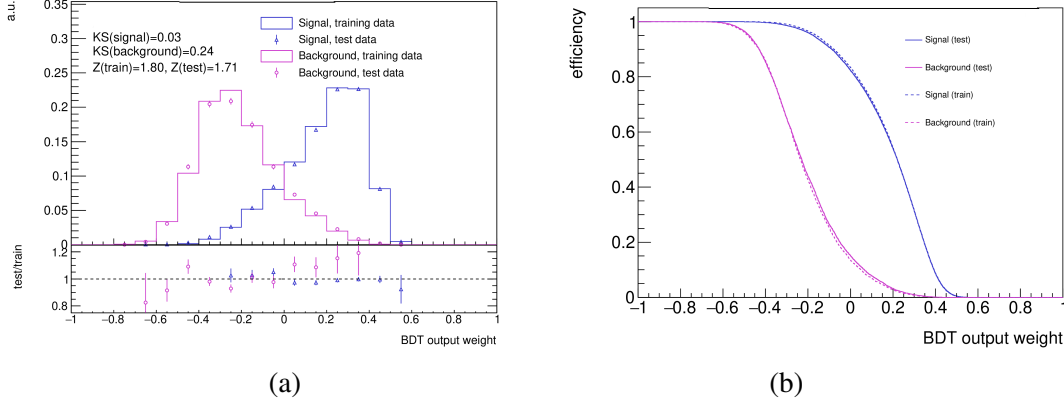


Figure 5.33: Results of MVA training with the optimized configuration, (a) BDT output distribution, (b) signal and background efficiency as a function of the cut value on the BDT output.

- MaxDepth = 3, a decision tree can have three layers at most.
- nCuts = 60, the maximum number of cuts can be used to build a tree.
- MinNodeSize = 2, minimum 0.2% of data be used to make a new node.
- nTrees = 800, number of trees in the training.
- AdaBoostBeta = 0.35, learning rate, the step to adjust weight after each iteration.

Figure 5.33 shows the distributions of the BDT output and the signal (background) efficiency. The training has the optimized configuration listed above. The core of the training algorithms is based on adaptive boosting in the TMVA framework [149]. The preliminary result of fitting with BDT output in 0L channel is shown in Figure 5.34. The binning of the fit is determined from the binning algorithm referred to as "transformations". The transformation function used in the fit of the preliminary result is defined as

$$Z = z_s n_s / N_s + z_b n_b / N_b, \quad (5.22)$$

where $z_s = 3$, $z_b = 2$ are parameters with values that can be adjusted, $N_{s(b)}$ is the total yields of signal (background) events in the histogram, $n_{s(b)}$ is the number of signal (background) events in the interval $I(k, l)$. The binning process starts from the rightmost bin of the input histogram, the range of the interval $I(k, l)$ is increased by adding one bin after the other. The value of Z is calculated at each step. Once the condition $Z(I(k, l)) > 1$ is fulfilled, the second bin boundary is defined.

The significance from fitting with BDT output is 2.01 in $p_T^V > 400$ GeV, 0L channel. The significance from fitting with m_J in the same region using the exact same event selection is 0.99. The implementation of MVA for $VH(b\bar{b})$ boosted is expected to bring in significant improvement to the analysis.

5 Measurement of boosted $VH(H \rightarrow b\bar{b})$ process

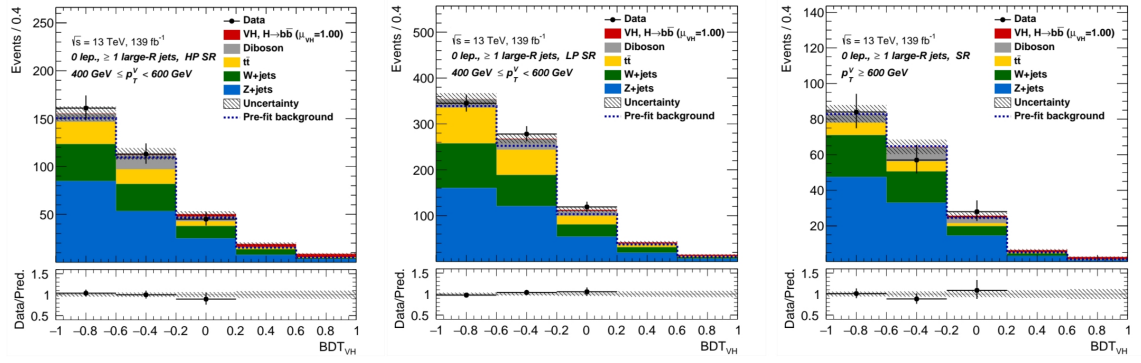


Figure 5.34: Preliminary result of fitting with BDT output in 0L channel.

6 Conclusions

The discovery of the Higgs boson in 2012 marks an outstanding achievement in particle physics. The properties of the Higgs boson has been measured more and more precisely since then. This thesis presents the measurements of a Higgs boson decaying into a pair of b -quarks, produced in association with a vector boson decaying into a pair of leptons. The $H \rightarrow b\bar{b}$ decay is an important channel to study the coupling between the Higgs boson and the fermions. It was observed seven years after the discovery of the Higgs boson because this channel has an all-hadronic final state through the gluon-gluon fusion production mode. The most sensitive production mode to study $H \rightarrow b\bar{b}$ decay is the VH production. The leptons from the decay of the vector boson enable sufficient rejection of the multijet background.

The observation of the $VH(H \rightarrow b\bar{b})$ process is carried out in two analysis strategies with pp collision data at center-of-mass energy of 13 TeV collected by the ATLAS detector during 2015 to 2018, corresponding to an integrated luminosity of 139 fb^{-1} . The $VH(b\bar{b})$ resolved analysis targets low to medium p_T phase space, while the $VH(b\bar{b})$ boosted analysis targets medium to high p_T phase space. This thesis focuses on the $VH(b\bar{b})$ boosted analysis. The decay products of the Higgs boson are reconstructed as a large- R jet. For a Higgs boson mass of 125 GeV, the significance of the VH process was observed (expected) to be 2.1 (2.7) standard deviation. The results are compatible with the Standard Model predictions. The analysis performs a simultaneous fit to study the signal-like VZ background. The VZ process is observed with a significance of 5.4 standard deviations, compared to an expected significance of 5.7 standard deviations. The cross section of the VH signal has been measured in the STXS bins and interpreted in an EFT approach. All the measurements are compatible with the SM.

The analysis can be improved from several aspects. The total statistical uncertainty is at the same level as the total systematic uncertainties with full Run 2 dataset, therefore the increased luminosity of Run 3 will significantly improve the sensitivity of the analysis. The large- R jet uncertainties have the most important contribution to the uncertainty of the signal strength. The improved large- R jet reconstructed from unified flow objects will enable more precise measurement of the signal strength. The analysis will also benefit from the improved energy resolution of the small- R jet collections and the flavor tagging performance.

The analysis strategy has already been improved in the $VH(b\bar{b})$ legacy analysis. One of the important improvement in the legacy analysis is the use of machine learning technique. The boosted decision tree output will be used as the final discriminant in the fit. It is expected to bring in significant improvement to the analysis on the basis of the large- R jet mass. The Higgs candidate selection strategy is also improved together with the multivariate analysis. Another improvement is the combination of $VH(b\bar{b})$ resolved and $VH(b\bar{b})$ boosted analyses. It can be implemented in several ways. Other than the simple combination, the more complicated combination could give extra improvement to the analysis.

6 Conclusions

The $H \rightarrow b\bar{b}$ decay channel has been observed by the CMS experiment in Run 2 [58, 150]. The main contribution to the observation is from VH production. The analysis strategies are similar to the $VH(b\bar{b})$ resolved and boosted analyses at the ATLAS experiment. The CMS $VH(b\bar{b})$ resolved analysis also used two small- R ($R = 0.4$) jets to reconstruct the decay products of the Higgs boson. The VH signal was observed with a significance of 4.8 standard deviations, and expected with 4.9 standard deviations using data collected in Run 1 and early Run 2. The CMS $VH(b\bar{b})$ boosted analysis used $R = 0.8$ large- R jets and different b -tagging algorithms. The ATLAS and the CMS analyses have the same STXS bins. The resolved and the boosted analysis has been combined by using the resolved strategy in $p_{\text{T}}^V < 250$ GeV, and the boosted strategy in $p_{\text{T}}^V > 250$ GeV. The inclusive signal strength measured by the CMS experiment using full Run 2 data is $0.58^{+0.19}_{-0.18}$ [151], with an observed significance of 3.3 standard deviations. Future combination of results from the ATLAS and the CMS experiment could improve the measurement further.

The Standard Model is still far from being a complete theory. The $H \rightarrow b\bar{b}$ decay channel provides a great way to study the Yukawa nature of the coupling between the Higgs boson and the fermions [152, 153]. With the upgrades of LHC Run 3 and the High-luminosity Large Hadron Collider (HL-LHC) project in the future, the precision measurements in $H \rightarrow b\bar{b}$ channel have potential to observe anomalous couplings.

Acknowledgements

I would like to express my deepest appreciation to my supervisor, Prof. Yanwen Liu, for the guidance of my Ph.D study and research. I could not have undertaken this journey without his continuous support.

I am thankful to the USTC ATLAS group for providing me the opportunity to participate in analyses at the ATLAS experiment. Many thanks to Changqiao Li for his patience in replying to my questions, working together and sharing ideas.

I am also thankful to the ATLAS $\text{jet}/E_{\text{T}}^{\text{miss}}$ group and the jet definition group, in particular the group conveners: Christopher Young, Chris Malena Delitzsch, Steven Schramm, Jennifer Roloff and Matt Leblanc for the excellent jet performance studies and the amazing working environment. Special thanks to Pierre-Antoine Delsart for the guidance and discussion that allowed me to acquire expertise with the jet reconstruction.

Thanks should also go to every member in the $VH(b\bar{b})$ analysis team for the collaboration to make this analysis possible, the discussion and the idea sharing that allowed me to get an overview with every part of the analysis. I am also thankful to the analysis contacts for the very welcoming atmosphere in the group.

I would like to acknowledge all my colleagues and friends for working together, and for all the fun we have had in the past few years.

Lastly, I would be remiss in not mentioning my parents for supporting me spiritually throughout my Ph.D, and for always standing behind every decision I made.

Publications and conference talks

Published journal articles

1. ATLAS Collaboration, *Measurement of the associated production of a Higgs boson decaying into b -quarks with a vector boson at high transverse momentum in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Phys. Lett. B 816 (2021) 136204.
2. ATLAS Collaboration, *Optimisation of large-radius jet reconstruction for the ATLAS detector in 13 TeV proton-proton collisions*, Eur. Phys. J. C 81 (2021) 4, 334.

ATLAS conference notes

1. ATLAS Collaboration, *Performance of large-radius jet definitions with the ATLAS detector in 13 TeV proton-proton collisions*, ATLAS-CONF-2020-021, <https://cds.cern.ch/record/2723736>.
2. ATLAS Collaboration, *Measurement of the associated production of a Higgs boson decaying to b quarks with a vector boson at high transverse momentum in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, ATLAS-CONF-2020-007, <https://cds.cern.ch/record/2715063>.

Conference talks

1. Tagging boosted decays using large-radius jets with ATLAS.
BOOST2021: 13th International Workshop on Boosted Object.
2. Performance of large radius jets reconstruction at ATLAS.
China LHC Physics workshop (CLHCP) 2020.

Appendix

A MC calibration of $R = 0.4$ jets in low- μ environment

Measurements that probe the strong interaction, constrain parton distribution functions and measurements with heavy ion data have special interests for data collected in low pile-up environment. Jets reconstructed in low- μ data need to have a dedicated MC JES calibration. Therefore the MC calibration was derived for small- R jets reconstructed from EM energy scale topo-clusters (EMTopo) and PFlow objects (EMPFlow). The binning used in small- R JES calibration is slightly different, the η bins are from -4.5 to 4.5 with bin width of 0.1 and the E bins are from 20 GeV to 5500 GeV with various bin width. The MC samples are QCD di-jet processes generated by Pythia 8.

The initial JES calibration derived for Run 2 data is for EMTopo jets. The average out of time pile-up $\langle\mu\rangle$ is 2. The reconstruction of the jets uses the same calorimeter cell noise threshold as the standard LHC Run 2 data, corresponding to $\langle\mu\rangle = 40$. No pile-up correction is applied before MC calibration. The energy response distributions of uncalibrated and calibrated EMTopo jets are shown in Figure A.1. The calibrated energy response is close to unity in most of the regions. The only non-closure occurs when p_T gets very small (close to 20 GeV), so the recommended low p_T threshold for EMTopo jets is above 20 GeV.

With the improved performance of small- R jets reconstructed from PFlow objects, the default jet collection in the ATLAS experiment has been updated. The MC JES calibration for EMPFlow jets in low- μ environment is also derived. The calorimeter noise threshold used in the reconstruction was the

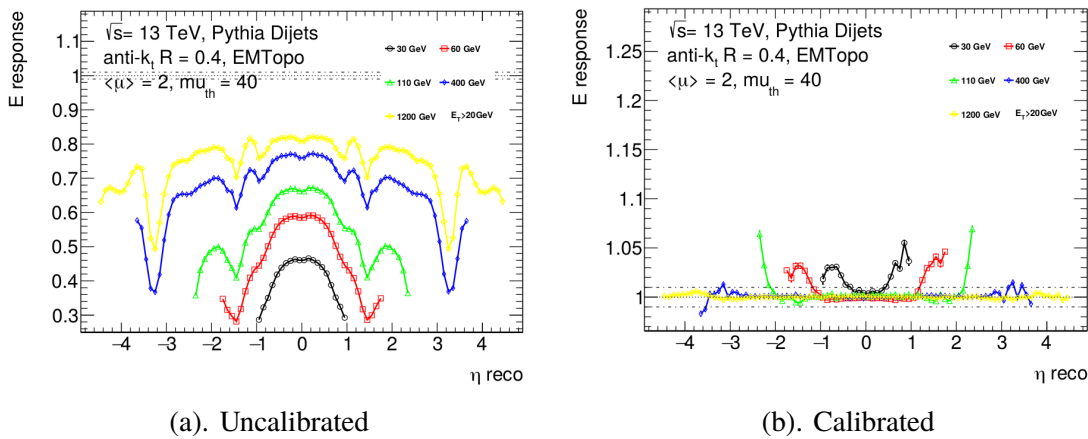


Figure A.1: Energy response for EMTopo jets in low- μ environment, in bins of η at different E^{true} slices. The jet energy is reconstructed with calorimeter noise threshold corresponding to $\langle\mu\rangle = 40$.

A MC calibration of $R = 0.4$ jets in low- μ environment

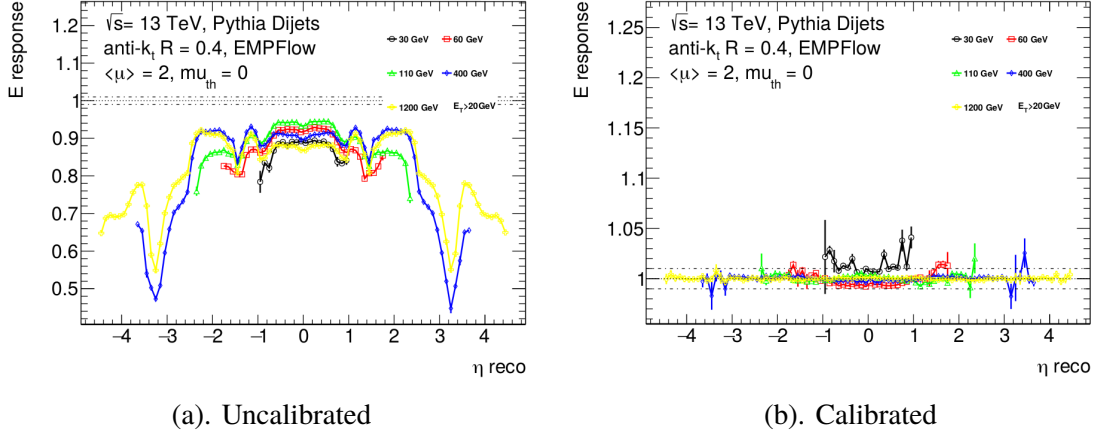


Figure A.2: Energy response for EMPFlow jets in low- μ environment, in bins of η at different E^{true} slices. The jet energy is reconstructed with calorimeter noise threshold corresponding to low- μ environment.

same as the standard Run 2 data, denoted as $th(\mu = 40)$. Then it was upgraded to the noise threshold corresponding to low- μ data, denoted as $th(\mu = 0)$. Figure A.2 shows the energy response distributions of EMPFlow jets reconstructed with $th(\mu = 0)$ in low- μ environment. The JES factors and calibration performance of EMPFlow jets reconstructed with $th(\mu = 40)$ are very similar. The uncertainties can be limited within 4% in $p_T > 20$ GeV, and within 1% in most of the studied regions.

The advantage of using $th(\mu = 0)$ in the reconstruction is to improve the resolution of jet energy. Figure A.3 shows the estimated energy resolution from the calibrated energy response histograms generated in the JES calibration. The energy resolution is estimated as 68% interquartile range of the calibrated energy response histograms in (E^{true}, η) bins. The bin centers are taken as the jet kinematic variables of the estimated resolution. Jets reconstructed with $th(\mu = 0)$ has better energy resolution than jets reconstructed with $th(\mu = 40)$ constantly. EMPFlow jets reconstructed with $th(\mu = 0)$ becomes the default jet collection for low pile-up data.

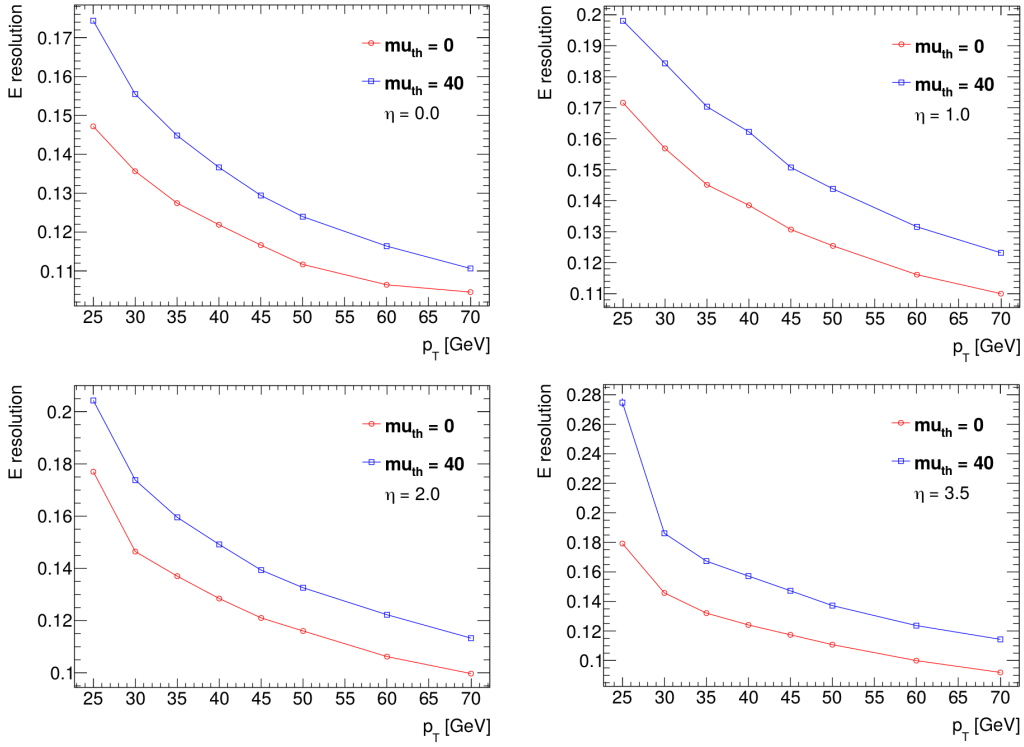


Figure A.3: Energy resolution distributions of EMPFlow jets reconstructed with calorimeter noise threshold corresponding to standard Run 2 environment: $\text{th}(\mu = 40)$, and low- μ environment: $\text{th}(\mu = 0)$. The resolution values are estimated as 68% interquartile range of $R = E^{\text{calib}}/E^{\text{true}}$ histograms in (E^{true}, η) bins.

B Nuisance parameter pull plots

The NP pulls shown in this appendix are obtained from unconditional fit to the Asimov dataset (red) and the actual dataset (black) in $VH(bb)$ boosted analysis. These plots together with the plots presented in Section 5.4.4 contain the completed NP pulls. Figure B.1 shows the pulls for NPs related to experimental systematic uncertainties, and Figure B.2 shows the pulls for NPs related to modeling systematic uncertainties.

Figure B.1 (a) shows the NP pulls for the uncertainties related to small- R jet. Most of these NPs are not pulled very much. The NPs related to jet energy resolution (JER) are pulled a little. The uncertainties have very small impact on the analysis because the small- R jets are only used in the event selection. The NP pulls for uncertainties related to leptons, E_T^{miss} and luminosity are shown in the rest plots in Figure B.1. All these NPs are constrained well and pulled very little.

Figure B.2 (a) and Figure B.2 (b) show the NP pulls for V +jets modeling uncertainties. The floating normalization factors for Z +HF (norm_Zbb) and W +HF (norm_Wbb) are constrained to 1.32 and

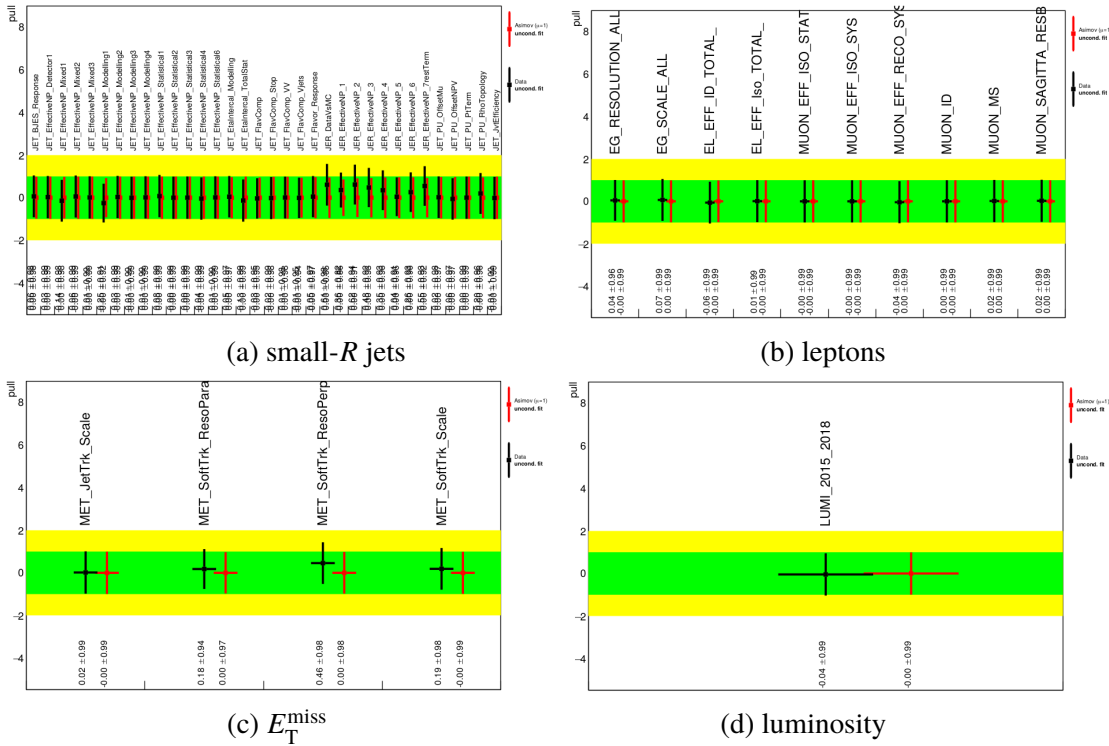


Figure B.1: Pulls and constraints for NPs related to experimental systematic uncertainties.

B Nuisance parameter pull plots

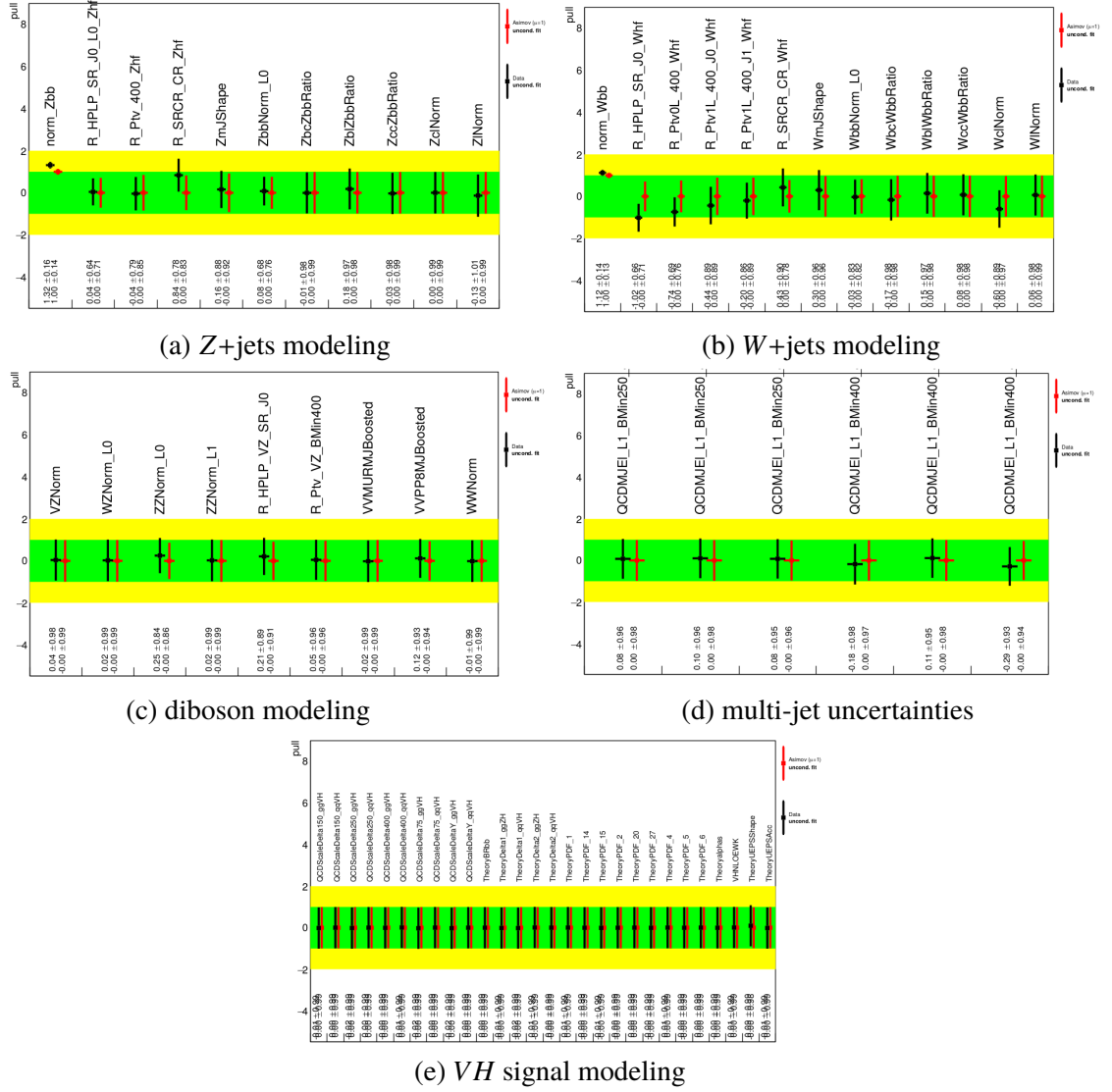


Figure B.2: Pulls and constraints for NPs related to modeling systematic uncertainties and QCD multi-jet.

1.12, respectively. The NPs are not pulled very much. The acceptance uncertainty between signal region and top control region for Z +jets is pulled up. The acceptance uncertainty between HP and LP signal region for W +jets is pulled down. These pulls occurred because the fit tends to adjust the MC predictions to match the data. The pulls of modeling uncertainties for diboson, multijet and signal are shown in Figure B.2 (c), B.2 (d) and B.2 (e), respectively. All the NPs are pulled very little.

Bibliography

- [1] P. D. Group et al., *Review of Particle Physics*, *Progress of Theoretical and Experimental Physics* **2020** (2020), 083C01, ISSN: 2050-3911, eprint: <https://academic.oup.com/ptep/article-pdf/2020/8/083C01/34673722/ptaa104.pdf>, URL: <https://doi.org/10.1093/ptep/ptaa104> (cit. on pp. 15, 19).
- [2] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, *Phys. Lett. B* **716** (2012) 1, arXiv: [1207.7214 \[hep-ex\]](#) (cit. on p. 15).
- [3] CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, *Phys. Lett. B* **716** (2012) 30, arXiv: [1207.7235 \[hep-ex\]](#) (cit. on p. 15).
- [4] L. Evans and P. Bryant, *LHC Machine*, *JINST* **3** (2008) S08001 (cit. on pp. 15, 43).
- [5] 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** (2015) 191803, arXiv: [1503.07589 \[hep-ex\]](#) (cit. on pp. 15, 29).
- [6] M. K. Gaillard, P. D. Grannis and F. J. Sciulli, *The Standard model of particle physics*, *Rev. Mod. Phys.* **71** (1999) S96, arXiv: [hep-ph/9812285 \[hep-ph\]](#) (cit. on p. 19).
- [7] M. Gell-Mann, *The Eightfold Way: A Theory of strong interaction symmetry*, (1961) (cit. on pp. 19, 23).
- [8] E. Noether, *Invariant Variation Problems*, *Gott. Nachr.* **1918** (1918) 235, [Transp. Theory Statist. Phys.1,186(1971)], arXiv: [physics/0503066 \[physics\]](#) (cit. on p. 19).
- [9] The LHCb Collaboration, *Observation of structure in the J/ψ -pair mass spectrum*, (), arXiv: [2006.16957 \[hep-ex\]](#) (cit. on p. 19).
- [10] The LHCb Collaboration, *Observation of J/ψ resonances consistent with pentaquark states in $\Lambda_b^0 \rightarrow J/\psi K^- p$ decays*, *Phys. Rev. Lett.* **072001** (2015) 115, arXiv: [1507.03414 \[hep-ex\]](#) (cit. on p. 19).
- [11] S. Weinberg, *A Model of Leptons*, *Phys. Rev. Lett.* **19** (1967) 1264 (cit. on p. 19).
- [12] F. Englert and R. Brout, *Broken Symmetry and the Mass of Gauge Vector Mesons*, *Phys. Rev. Lett.* **13** (1964) 321, [157(1964)] (cit. on pp. 20, 26).
- [13] P. W. Higgs, *Broken Symmetries and the Masses of Gauge Bosons*, *Phys. Rev. Lett.* **13** (1964) 508, [160(1964)] (cit. on pp. 20, 26).

BIBLIOGRAPHY

- [14] Wikipedia contributors, *The Standard Model* — *Wikipedia, The Free Encyclopedia*, [Online; accessed 07-Dec-2021], 2021, URL: https://en.wikipedia.org/wiki/Standard_Model (cit. on p. 20).
- [15] M. E. Peskin and D. V. Schroeder, *An Introduction to quantum field theory*, Reading, USA: Addison-Wesley, 1995, ISBN: 9780201503975, URL: <http://www.slac.stanford.edu/~mpeskin/QFT.html> (cit. on p. 20).
- [16] P. A. M. Dirac, *Quantum theory of emission and absorption of radiation*, *Proc. Roy. Soc. Lond. A* **114** (1927) 243 (cit. on p. 21).
- [17] P. A. M. Dirac, *The quantum theory of the electron*, *Proc. Roy. Soc. Lond. A* **117** (1928) 610 (cit. on p. 21).
- [18] W. Heisenberg and W. Pauli, *Zur Quantendynamik der Wellenfelder*, *Zeitschrift für Physik* **56** (1929) 1, ISSN: 0044-3328, URL: <https://doi.org/10.1007/BF01340129> (cit. on p. 21).
- [19] W. Heisenberg and W. Pauli, *Zur Quantentheorie der Wellenfelder. II*, *Zeitschrift für Physik* **59** (1930) 168, ISSN: 0044-3328, URL: <https://doi.org/10.1007/BF01341423> (cit. on p. 21).
- [20] C.-N. Yang and R. L. Mills, *Conservation of Isotopic Spin and Isotopic Gauge Invariance*, *Phys. Rev.* **96** (1954) 191, [,150(1954)] (cit. on p. 22).
- [21] C. N. Yang and R. L. Mills, *Conservation of Isotopic Spin and Isotopic Gauge Invariance*, *Phys. Rev.* **96** (1 1954) 191, URL: <https://link.aps.org/doi/10.1103/PhysRev.96.191> (cit. on p. 22).
- [22] Wikipedia contributors, *Antisymmetric tensor* — *Wikipedia, The Free Encyclopedia*, 2022, URL: https://en.wikipedia.org/wiki/Antisymmetric_tensor (cit. on p. 23).
- [23] M. Gell-Mann, *The interpretation of the new particles as displaced charge multiplets*, *Nuovo Cim.* **4** (1956) 848 (cit. on p. 23).
- [24] M. Y. Han and Y. Nambu, *Three Triplet Model with Double SU(3) Symmetry*, *Phys. Rev.* **139** (1965) B1006, [,187(1965)] (cit. on p. 23).
- [25] O. W. Greenberg, *Spin and Unitary Spin Independence in a Paraquark Model of Baryons and Mesons*, *Phys. Rev. Lett.* **13** (1964) 598 (cit. on p. 23).
- [26] H. Fritzsch, M. Gell-Mann and H. Leutwyler, *Advantages of the Color Octet Gluon Picture*, *Phys. Lett. B* **47** (1973) 365 (cit. on p. 23).
- [27] S. L. Glashow, *Partial Symmetries of Weak Interactions*, *Nucl. Phys.* **22** (1961) 579 (cit. on p. 24).
- [28] S. Weinberg, *Implications of dynamical symmetry breaking*, *Phys. Rev. D* **13** (1976) 974 (cit. on p. 24).
- [29] S. Weinberg, *Implications of dynamical symmetry breaking: an addendum*, *Phys. Rev. D* **19** (1979) 1277 (cit. on p. 24).
- [30] N. Cabibbo, *Unitary Symmetry and Leptonic Decays*, *Phys. Rev. Lett.* **10** (1963) 531, [,648(1963)] (cit. on p. 25).

BIBLIOGRAPHY

- [31] ATLAS Collaboration, *Measurement of the Higgs boson mass in the $H \rightarrow ZZ^* \rightarrow 4\ell$ and $H \rightarrow \gamma\gamma$ channels with $\sqrt{s} = 13$ TeV pp collisions using the ATLAS detector*, *Phys. Lett. B* **784** (2018) 345, arXiv: [1806.00242 \[hep-ex\]](#) (cit. on p. 29).
- [32] CMS Collaboration, *A measurement of the Higgs boson mass in the diphoton decay channel*, *Phys. Lett. B* **805** (2020) 135425, arXiv: [2002.06398 \[hep-ex\]](#) (cit. on p. 29).
- [33] ATLAS Collaboration, *Evidence for the spin-0 nature of the Higgs boson using ATLAS data*, *Phys. Lett. B* **726** (2013) 120, arXiv: [1307.1432 \[hep-ex\]](#) (cit. on p. 30).
- [34] R. D. Ball et al., *Parton distributions for the LHC run II*, *JHEP* **04** (2015) 040, arXiv: [1410.8849 \[hep-ph\]](#) (cit. on pp. 30, 101).
- [35] LHC Higgs Working Group, *LHC Higgs Working Group: Higgs cross sections and decay branching ratios*, URL: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHWG#Higgs_cross_sections_and_decay_b (cit. on pp. 31, 32).
- [36] D. de Florian et al., *Handbook of LHC Higgs cross sections: 4. Deciphering the nature of the Higgs sector*, (2016), arXiv: [1610.07922 \[hep-ph\]](#) (cit. on pp. 31, 32, 101).
- [37] ATLAS Collaboration, *Observation and measurement of Higgs boson decays to WW^* with the ATLAS detector*, *Phys. Rev. D* **92** (2015) 012006, arXiv: [1412.2641 \[hep-ex\]](#) (cit. on p. 33).
- [38] CMS Collaboration, *Measurement of Higgs boson production and properties in the WW decay channel with leptonic final states*, *JHEP* **01** (2014) 096, arXiv: [1312.1129 \[hep-ex\]](#) (cit. on p. 33).
- [39] ATLAS Collaboration, *Measurements of gluon-gluon fusion and vector-boson fusion Higgs boson production cross-sections in the $H \rightarrow WW^* \rightarrow e\nu\mu\nu$ decay channel in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, *Phys. Lett. B* **789** (2019) 508, arXiv: [1808.09054 \[hep-ex\]](#) (cit. on p. 33).
- [40] CMS Collaboration, *Measurement of the inclusive and differential Higgs boson production cross sections in the leptonic WW decay mode at $\sqrt{s} = 13$ TeV*, (2020), arXiv: [2007.01984 \[hep-ex\]](#) (cit. on p. 34).
- [41] ATLAS Collaboration, *Cross-section measurements of the Higgs boson decaying into a pair of τ -leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, *Phys. Rev. D* **99** (2019) 072001, arXiv: [1811.08856 \[hep-ex\]](#) (cit. on p. 33).
- [42] CMS Collaboration, *Measurement of the inclusive and differential Higgs boson production cross sections in the decay mode to a pair of τ leptons in pp collisions at $\sqrt{s} = 13$ TeV*, (2021), arXiv: [2107.11486 \[hep-ex\]](#) (cit. on p. 33).
- [43] ATLAS Collaboration, *Measurements of Higgs boson production and couplings in the four-lepton channel in pp collisions at center-of-mass energies of 7 and 8 TeV with the ATLAS detector*, *Phys. Rev. D* **91** (2015) 012006, arXiv: [1408.5191 \[hep-ex\]](#) (cit. on p. 34).

BIBLIOGRAPHY

- [44] ATLAS Collaboration, *Measurement of Higgs boson production in the diphoton decay channel in pp collisions at center-of-mass energies of 7 and 8 TeV with the ATLAS detector*, *Phys. Rev. D* **90** (2014) 112015, arXiv: [1408.7084 \[hep-ex\]](#) (cit. on p. 34).
- [45] CMS Collaboration, *Measurement of differential and integrated fiducial cross sections for Higgs boson production in the four-lepton decay channel in pp collisions at $\sqrt{s} = 7$ and 8 TeV*, *JHEP* **04** (2016) 005, arXiv: [1512.08377 \[hep-ex\]](#) (cit. on p. 34).
- [46] CMS Collaboration, *Observation of the diphoton decay of the Higgs boson and measurement of its properties*, *Eur. Phys. J. C* **74** (2014) 3076, arXiv: [1407.0558 \[hep-ex\]](#) (cit. on p. 34).
- [47] ATLAS Collaboration, *Measurements of Higgs boson properties in the diphoton decay channel with 36fb^{-1} of pp collision data at $\sqrt{s} = 13$ TeV with the ATLAS detector*, *Phys. Rev. D* **98** (2018) 052005, arXiv: [1802.04146 \[hep-ex\]](#) (cit. on p. 35).
- [48] ATLAS Collaboration, *Measurements of the Higgs boson inclusive and differential fiducial cross sections in the 4ℓ decay channel at $\sqrt{s} = 13$ TeV*, *Eur. Phys. J.* (2020), arXiv: [2004.03969 \[hep-ex\]](#) (cit. on p. 35).
- [49] CMS Collaboration, *Measurements of Higgs boson production cross sections and couplings in the diphoton decay channel at $\sqrt{s} = 13$ TeV*, *JHEP* **07** (2021) 027, arXiv: [2103.06956 \[hep-ex\]](#) (cit. on p. 35).
- [50] CMS Collaboration, *Measurements of production cross sections of the Higgs boson in the four-lepton final state in proton–proton collisions at $\sqrt{s} = 13$ TeV*, *EPJC* **81** (2021) 488, arXiv: [2103.04956 \[hep-ex\]](#) (cit. on p. 35).
- [51] ATLAS Collaboration, *Direct constraint on the Higgs-charm coupling from a search for Higgs boson decays to charm quarks with the ATLAS detector*, ATLAS-CONF-2021-021, 2021, URL: <https://cdsweb.cern.ch/record/2771724> (cit. on p. 34).
- [52] CMS Collaboration, *A search for the standard model Higgs boson decaying to charm quarks*, *JHEP* **03** (2020) 131, arXiv: [1912.01662 \[hep-ex\]](#) (cit. on p. 34).
- [53] ATLAS Collaboration, *A search for the dimuon decay of the Standard Model Higgs boson with the ATLAS detector*, (2020), arXiv: [2007.07830 \[hep-ex\]](#) (cit. on p. 36).
- [54] CMS Collaboration, *Evidence for Higgs boson decay to a pair of muons*, (2021), arXiv: [2009.04363 \[hep-ex\]](#) (cit. on p. 36).
- [55] ATLAS Collaboration, *Evidence for Higgs boson decays to a low-mass dilepton system and a photon in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, *Phys. Lett. B* **819** (2021), arXiv: [2103.10322 \[hep-ex\]](#) (cit. on p. 36).
- [56] CMS Collaboration, *Search for the decay of a Higgs boson in the $\ell\ell\gamma$ channel in proton–proton collisions at $\sqrt{s} = 13$ TeV*, *JHEP* **11** (2018) 152, arXiv: [1806.05996 \[hep-ex\]](#) (cit. on p. 36).
- [57] ATLAS Collaboration, *Evidence for the $H \rightarrow b\bar{b}$ decay with the ATLAS detector*, *JHEP* **12** (2017) 024, arXiv: [1708.03299 \[hep-ex\]](#) (cit. on p. 36).
- [58] CMS Collaboration, *Observation of Higgs Boson Decay to Bottom Quarks*, *Phys. Rev. Lett.* **121** (2018) 121801, arXiv: [1808.08242 \[hep-ex\]](#) (cit. on pp. 36, 162).

BIBLIOGRAPHY

- [59] ATLAS Collaboration, *Measurements of WH and ZH production in the $H \rightarrow b\bar{b}$ decay channel in pp collisions at 13 TeV with the ATLAS detector*, (2020), arXiv: [2007.02873 \[hep-ex\]](#) (cit. on p. 36).
- [60] ATLAS Collaboration, *Methodology for EFT interpretation of Higgs boson simplified template cross-section results in ATLAS*, ATL-PHYS-PUB-2019-042, 2019, URL: <https://cds.cern.ch/record/2694284> (cit. on pp. 37, 42).
- [61] B. Grzadkowski, M. Iskrzyński, M. Misiak, J. Rosiek, *Dimension-six terms in the Standard Model Lagrangian*, *JHEP* **1010** (2010) 085, arXiv: [1008.4884v2 \[hep-ph\]](#) (cit. on pp. 37, 40).
- [62] N. Berger at al., *Simplified Template Cross Sections - Stage 1.1*, (2019), arXiv: [1906.02754 \[hep-ph\]](#) (cit. on p. 38).
- [63] Hays, Chris (University of Oxford (GB)) ; Sanz Gonzalez, Veronica (CERN) ; Zemaityte, Gabija (University of Oxford (GB)), *Constraining EFT parameters using simplified template cross sections*, LHCHXSWG-2019-004, 2019, URL: <https://cds.cern.ch/record/2673969> (cit. on pp. 41, 150).
- [64] E. Mobs, *The CERN accelerator complex. Complexe des accélérateurs du CERN*, (2016), General Photo, URL: <https://cds.cern.ch/record/2197559> (cit. on p. 43).
- [65] ATLAS Collaboration, *Luminosity Public Results in Run 2*, 2016, URL: <https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2> (visited on 10/08/2018) (cit. on p. 45).
- [66] ATLAS Collaboration, *ATLAS Central Solenoid: Magnet Project Technical Design Report, Volume 4*, ATLAS-TDR-9; CERN-LHCC-97-021, CERN, 1997, URL: <https://cds.cern.ch/record/331067> (cit. on p. 47).
- [67] ATLAS Collaboration, *ATLAS Barrel Toroid: Magnet Project Technical Design Report, Volume 2*, ATLAS-TDR-7; CERN-LHCC-97-019, 1997, URL: <https://cds.cern.ch/record/331065> (cit. on p. 47).
- [68] ATLAS Collaboration, *ATLAS End-Cap Toroids: Magnet Project Technical Design Report, Volume 3*, ATLAS-TDR-8; CERN-LHCC-97-020, CERN, 1997, URL: <https://cds.cern.ch/record/331066> (cit. on p. 47).
- [69] *ATLAS magnet system: Technical Design Report, 1*, Technical design report. ATLAS, Geneva: CERN, 1997, URL: <https://cds.cern.ch/record/338080> (cit. on p. 47).
- [70] ATLAS Collaboration, *ATLAS Inner Detector: Technical Design Report, Volume 1*, ATLAS-TDR-4; CERN-LHCC-97-016, 1997, URL: <https://cds.cern.ch/record/331063> (cit. on p. 48).
- [71] ATLAS Collaboration, *ATLAS Inner Detector: Technical Design Report, Volume 2*, ATLAS-TDR-5, CERN-LHCC-97-017, 1997, URL: <https://cds.cern.ch/record/331064> (cit. on p. 48).

BIBLIOGRAPHY

- [72] ATLAS Collaboration, *The ATLAS Inner Detector commissioning and calibration*, *Eur. Phys. J. C* **70** (2010) 787, arXiv: [1004.5293 \[hep-ex\]](#) (cit. on p. 48).
- [73] ATLAS Collaboration, *ATLAS Pixel Detector: Technical Design Report*, ATLAS-TDR-11; CERN-LHCC-98-013, 1998, URL: <https://cds.cern.ch/record/381263> (cit. on p. 48).
- [74] B. Abbott et al., *Production and integration of the ATLAS Insertable B-Layer*, *JINST* **13** (2018) T05008, arXiv: [1803.00844 \[physics.ins-det\]](#) (cit. on p. 48).
- [75] ATLAS Collaboration, *Efficiency and Hit Spatial Resolution of ATLAS IBL Sensors in LHC Run 2 Collision Events*, ATL-INDET-PUB-2016-001, 2016, URL: <https://cds.cern.ch/record/2203893> (cit. on p. 48).
- [76] ATLAS Collaboration, *Alignment of the ATLAS Inner Detector with the initial LHC data at $\sqrt{s} = 13$ TeV*, ATL-PHYS-PUB-2015-031, 2015, URL: <https://cds.cern.ch/record/2038139> (cit. on pp. 48, 49).
- [77] ATLAS Collaboration, *Studies of the performance of the ATLAS detector using cosmic-ray muons*, *Eur. Phys. J. C* **71** (2011) 1593, arXiv: [1011.6665 \[hep-ex\]](#) (cit. on pp. 50, 55).
- [78] ATLAS Collaboration, *Track Reconstruction Performance of the ATLAS Inner Detector at $\sqrt{s} = 13$ TeV*, ATL-PHYS-PUB-2015-018, 2015, URL: <https://cds.cern.ch/record/2037683> (cit. on pp. 50, 62).
- [79] *ATLAS liquid-argon calorimeter: Technical Design Report*, Technical Design Report ATLAS, Geneva: CERN, 1996, URL: <https://cds.cern.ch/record/331061> (cit. on pp. 50, 52, 53).
- [80] *ATLAS tile calorimeter: Technical Design Report*, Technical Design Report ATLAS, Geneva: CERN, 1996, URL: <https://cds.cern.ch/record/331062> (cit. on pp. 50, 53).
- [81] ATLAS Collaboration, *Operation and performance of the ATLAS Tile Calorimeter in Run 1*, *Eur. Phys. J. C* **78** (2018) 987, arXiv: [1806.02129 \[hep-ex\]](#) (cit. on p. 51).
- [82] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, *JINST* **3** (2008) S08003 (cit. on p. 53).
- [83] ATLAS Collaboration, *ATLAS Muon Spectrometer: Technical Design Report*, ATLAS-TDR-10; CERN-LHCC-97-022, CERN, 1997, URL: <https://cds.cern.ch/record/331068> (cit. on p. 53).
- [84] ATLAS Collaboration, *ATLAS Muon Spectrometer Phase-II Upgrade: Technical Design Report*, ATLAS-TDR-026; CERN-LHCC-2017-017, 2017, URL: <https://cds.cern.ch/record/2285580> (cit. on pp. 55, 56).
- [85] ATLAS Collaboration, *Commissioning of the ATLAS Muon Spectrometer with cosmic rays*, *Eur. Phys. J. C* **70** (2010) 875, arXiv: [1006.4384 \[hep-ex\]](#) (cit. on p. 54).

- [86] ATLAS Collaboration, *Resolution of the ATLAS muon spectrometer monitored drift tubes in LHC Run 2*, *JINST* **14** (2019) P09011, arXiv: 1906.12226 [hep-ex] (cit. on p. 55).
- [87] ATLAS Collaboration, *Performance of the ATLAS trigger system in 2015*, *Eur. Phys. J. C* **77** (2017) 317, arXiv: 1611.09661 [hep-ex] (cit. on p. 56).
- [88] The ATLAS Collaboration, *The ATLAS Experiment TDAQ Public Plots*, URL: <https://twiki.cern.ch/twiki/bin/view/AtlasPublic/ApprovedPlotsDAQ> (cit. on p. 57).
- [89] T. Cornelissen, M. Elsing, S. Fleischmann, W. Liebig and E. Moyses, *Concepts, Design and Implementation of the ATLAS New Tracking (NEWT)*, (2007), ed. by A. Salzburger (cit. on p. 60).
- [90] R. Fruhwirth, *Application of Kalman filtering to track and vertex fitting*, *Nucl. Instrum. Meth. A* **262** (1987) 444 (cit. on p. 61).
- [91] ATLAS Collaboration, *Early Inner Detector Tracking Performance in the 2015 Data at $\sqrt{s} = 13$ TeV*, ATL-PHYS-PUB-2015-051, 2015, URL: <https://cds.cern.ch/record/2110140> (cit. on pp. 61, 62).
- [92] ATLAS Collaboration, *Vertex Reconstruction Performance of the ATLAS Detector at $\sqrt{s} = 13$ TeV*, ATL-PHYS-PUB-2015-026, 2015, URL: <https://cds.cern.ch/record/2037717> (cit. on pp. 63, 64).
- [93] ATLAS Collaboration, *Reconstruction of primary vertices at the ATLAS experiment in Run 1 proton–proton collisions at the LHC*, *Eur. Phys. J. C* **77** (2017) 332, arXiv: 1611.10235 [hep-ex] (cit. on pp. 63, 64).
- [94] ATLAS Collaboration, *Electron efficiency measurements with the ATLAS detector using 2012 LHC proton–proton collision data*, *Eur. Phys. J. C* **77** (2017) 195, arXiv: 1612.01456 [hep-ex] (cit. on pp. 63, 66).
- [95] ATLAS Collaboration, *Electron efficiency measurements with the ATLAS detector using the 2015 LHC proton–proton collision data*, ATLAS-CONF-2016-024, 2016, URL: <https://cds.cern.ch/record/2157687> (cit. on p. 65).
- [96] W. Lampl et al., *Calorimeter Clustering Algorithms: Description and Performance*, ATL-LARG-PUB-2008-002, 2008, URL: <https://cds.cern.ch/record/1099735> (cit. on p. 64).
- [97] ATLAS Collaboration, *Improved electron reconstruction in ATLAS using the Gaussian Sum Filter-based model for bremsstrahlung*, ATLAS-CONF-2012-047, 2012, URL: <https://cds.cern.ch/record/1449796> (cit. on p. 65).
- [98] ATLAS Collaboration, *Electron and photon performance measurements with the ATLAS detector using the 2015–2017 LHC proton–proton collision data*, *JINST* **14** (2019) P12006, arXiv: 1908.00005 [hep-ex] (cit. on pp. 66–68, 120).

BIBLIOGRAPHY

- [99] ATLAS Collaboration, *Muon reconstruction performance of the ATLAS detector in proton–proton collision data at $\sqrt{s} = 13$ TeV*, *Eur. Phys. J. C* **76** (2016) 292, arXiv: [1603.05598 \[hep-ex\]](#) (cit. on pp. 68, 120).
- [100] ATLAS Collaboration, *Muon reconstruction and identification efficiency in ATLAS using the full Run 2 pp collision data set at $\sqrt{s} = 13$ TeV*, *Eur. Phys. J. C* **81** (2021) 578, arXiv: [2012.00578 \[hep-ex\]](#) (cit. on pp. 68, 70, 120).
- [101] J. Illingworth and J. Kittler, *A survey of the hough transform*, *Computer Vision, Graphics, and Image Processing* **44** (1988) 87, ISSN: 0734-189X, URL: <http://www.sciencedirect.com/science/article/pii/S0734189X88800331> (cit. on p. 68).
- [102] ATLAS Collaboration, *Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1*, *Eur. Phys. J. C* **77** (2017) 490, arXiv: [1603.02934 \[hep-ex\]](#) (cit. on p. 71).
- [103] ATLAS Collaboration, *Improving jet substructure performance in ATLAS using Track-CaloClusters*, ATL-PHYS-PUB-2017-015, 2017, URL: <https://cds.cern.ch/record/2275636> (cit. on pp. 72, 73).
- [104] ATLAS Collaboration, *Jet reconstruction and performance using particle flow with the ATLAS Detector*, *Eur. Phys. J. C* **77** (2017) 466, arXiv: [1703.10485 \[hep-ex\]](#) (cit. on pp. 74, 75).
- [105] ATLAS Collaboration, *Optimisation of large-radius jet reconstruction for the ATLAS detector in 13 TeV proton-proton collisions*, *Eur. Phys. J. C* **81** (2021) 334, arXiv: [2009.04986 \[hep-ex\]](#) (cit. on p. 76).
- [106] ATLAS Collaboration, *Impact of Alternative Inputs and Grooming Methods on Large-R Jet Reconstruction in ATLAS*, ATL-PHYS-PUB-2017-020, 2017, URL: <https://cds.cern.ch/record/2297485> (cit. on p. 76).
- [107] M. Cacciari, G. P. Salam and G. Soyez, *The anti- k_t jet clustering algorithm*, *JHEP* **04** (2008) 063, arXiv: [0802.1189 \[hep-ph\]](#) (cit. on p. 77).
- [108] Stephen D. Ellis, Davision E. Soper, *Successive Combination Jet Algorithm For Hadron Collisions*, *Phys.Rev.D* (1993), arXiv: [9305266 \[hep-ph\]](#) (cit. on p. 77).
- [109] Yu.L. Dokshitzer, G.D. Leder, S. Moretti, B.R. Webber, *Better Jet Clustering Algorithms*, (1997), arXiv: [9707323 \[hep-ph\]](#) (cit. on p. 77).
- [110] M. Wobisch, T. Wengler, *Hadronization Corrections to Jet Cross Sections in Deep-Inelastic Scattering*, (1999), arXiv: [9907280 \[hep-ph\]](#) (cit. on p. 77).
- [111] ATLAS Collaboration, *Jet energy scale measurements and their systematic uncertainties in proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, *Phys. Rev. D* **96** (2017) 072002, arXiv: [1703.09665 \[hep-ex\]](#) (cit. on p. 78).

- [112] ATLAS Collaboration, *Performance of jet substructure techniques for large- R jets in proton–proton collisions at $\sqrt{s} = 7$ TeV using the ATLAS detector*, *JHEP* **09** (2013) 076, arXiv: [1306.4945 \[hep-ex\]](#) (cit. on p. 79).
- [113] ATLAS Collaboration, *Jet global sequential corrections with the ATLAS detector in proton–proton collisions at $\sqrt{s} = 8$ TeV*, ATLAS-CONF-2015-002, 2015, URL: <https://cds.cern.ch/record/2001682> (cit. on p. 81).
- [114] ATLAS Collaboration, *Jet Calibration and Systematic Uncertainties for Jets Reconstructed in the ATLAS Detector at $\sqrt{s} = 13$ TeV*, ATL-PHYS-PUB-2015-015, 2015, URL: <https://cds.cern.ch/record/2037613> (cit. on p. 81).
- [115] ATLAS Collaboration, *In situ calibration of large-radius jet energy and mass in 13 TeV proton–proton collisions with the ATLAS detector*, *Eur. Phys. J. C* **79** (2019) 135, arXiv: [1807.09477 \[hep-ex\]](#) (cit. on pp. 81, 82, 120).
- [116] ATLAS Collaboration, *Performance of missing transverse momentum reconstruction with the ATLAS detector using proton–proton collisions at $\sqrt{s} = 13$ TeV*, *Eur. Phys. J. C* **78** (2018) 903, arXiv: [1802.08168 \[hep-ex\]](#) (cit. on pp. 83, 84, 121).
- [117] Wikipedia contributors, *Polynomial regression — Wikipedia, The Free Encyclopedia*, 2022, URL: https://en.wikipedia.org/wiki/Polynomial_regression (cit. on p. 86).
- [118] Wikipedia contributors, *Kernel smoother — Wikipedia, The Free Encyclopedia*, 2022, URL: https://en.wikipedia.org/wiki/Kernel_smoother (cit. on p. 87).
- [119] ATLAS Collaboration, *Performance of the missing transverse momentum triggers for the ATLAS detector during Run-2 data taking*, *JHEP* **08** (2020) 080, arXiv: [2005.09554 \[hep-ex\]](#) (cit. on p. 98).
- [120] ATLAS Collaboration, *Trigger Menu in 2016*, ATL-DAQ-PUB-2017-001, 2017, URL: <https://cds.cern.ch/record/2242069> (cit. on p. 98).
- [121] ATLAS Collaboration, *Trigger Menu in 2017*, ATL-DAQ-PUB-2018-002, 2018, URL: <https://cds.cern.ch/record/2625986> (cit. on p. 98).
- [122] ATLAS Collaboration, *Trigger Menu in 2018*, ATL-DAQ-PUB-2019-001, 2019, URL: <https://cds.cern.ch/record/2693402> (cit. on p. 98).
- [123] ATLAS Collaboration, *Performance of electron and photon triggers in ATLAS during LHC Run 2*, *Eur. Phys. J. C* **80** (2020) 47, arXiv: [1909.00761 \[hep-ex\]](#) (cit. on pp. 98, 99).
- [124] P. Nason, *A new method for combining NLO QCD with shower Monte Carlo algorithms*, *JHEP* **11** (2004) 040, arXiv: [hep-ph/0409146](#) (cit. on p. 100).
- [125] S. Frixione, P. Nason and C. Oleari, *Matching NLO QCD computations with parton shower simulations: the POWHEG method*, *JHEP* **11** (2007) 070, arXiv: [0709.2092 \[hep-ph\]](#) (cit. on p. 100).
- [126] T. Sjöstrand, S. Mrenna and P. Skands, *A brief introduction to PYTHIA 8.1*, *Comput. Phys. Commun.* **178** (2008) 852, arXiv: [0710.3820 \[hep-ph\]](#) (cit. on p. 100).
- [127] T. Gleisberg et al., *Event generation with SHERPA 1.1*, *JHEP* **02** (2009) 007, arXiv: [0811.4622 \[hep-ph\]](#) (cit. on p. 100).

BIBLIOGRAPHY

- [128] *Sherpa Manual Version 2.2*,
URL: <https://sherpa.hepforge.org/doc/SHERPA-MC-2.2.5.html> (cit. on p. 100).
- [129] *The Herwig Event Generator*, URL: <https://herwig.hepforge.org> (cit. on p. 100).
- [130] M. Bahr, S. Gieseke, M. A. Gigg, D. Grellscheid, K. Hamilton, O. Latunde-Dada, S. Platzer, P. Richardson, M. H. Seymour, A. Sherstnev, J. Tully, B. R. Webber,
Herwig++ Physics and Manual, (2008), arXiv: [0803.0883 \[hep-ph\]](#) (cit. on p. 100).
- [131] J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer and T. Stelzer, *MadGraph 5 : Going Beyond*,
JHEP **06** (2011) 128, arXiv: [1106.0522 \[hep-ph\]](#) (cit. on p. 100).
- [132] K. Hamilton, P. Nason and G. Zanderighi, *MINLO: Multi-Scale Improved NLO*,
JHEP **10** (2012) 155, arXiv: [1206.3572 \[hep-ph\]](#) (cit. on p. 101).
- [133] ATLAS Collaboration, *Measurement of the transverse momentum distribution of Z/γ^* bosons in proton–proton collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector*,
Phys. Lett. B **705** (2011) 415, arXiv: [1107.2381 \[hep-ex\]](#) (cit. on p. 101).
- [134] M. Beneke, P. Falgari, S. Klein and C. Schwinn,
Hadronic top-quark pair production with NNLL threshold resummation,
Nucl. Phys. B **855** (2012) 695, arXiv: [1109.1536 \[hep-ph\]](#) (cit. on p. 103).
- [135] ATLAS Collaboration, *ATLAS b -jet identification performance and efficiency measurement with $t\bar{t}$ events in pp collisions at $\sqrt{s} = 13$ TeV*, *Eur. Phys. J. C* **79** (2019) 970,
arXiv: [1907.05120 \[hep-ex\]](#) (cit. on pp. 106, 108, 153, 154).
- [136] ATLAS Collaboration, *Optimisation and performance studies of the ATLAS b -tagging algorithms for the 2017-18 LHC run*, ATL-PHYS-PUB-2017-013, 2017,
URL: <https://cds.cern.ch/record/2273281> (cit. on pp. 106, 107, 153).
- [137] ATLAS Collaboration,
Secondary vertex finding for jet flavour identification with the ATLAS detector,
ATL-PHYS-PUB-2017-011, 2017, URL: <https://cds.cern.ch/record/2270366>
(cit. on p. 107).
- [138] G. S. Matteo Cacciari Gavin P. Salam, *The Catchment Area of Jets*, *JHEP* **04** (2008),
arXiv: [0802.1188 \[hep-ph\]](#) (cit. on p. 109).
- [139] ATLAS Collaboration, *Identification of boosted Higgs bosons decaying into b -quark pairs with the ATLAS detector at 13 TeV*, *Eur. Phys. J. C* **79** (2019) 836,
arXiv: [1906.11005 \[hep-ex\]](#) (cit. on p. 115).
- [140] ATLAS Collaboration, *Measurement of the associated production of a Higgs boson decaying into b -quarks with a vector boson at high transverse momentum in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, *Phys. Lett. B* (2021), arXiv: [2008.02508 \[hep-ex\]](#)
(cit. on pp. 115, 145–147, 149–152).
- [141] J. T. Andrew J. Larkoski Gavin P. Salam, *Energy Correlation Functions for Jet Substructure*,
JHEP **05** (2013), arXiv: [1305.0007 \[hep-ph\]](#) (cit. on p. 115).
- [142] G. Avoni et al.,
The new LUCID-2 detector for luminosity measurement and monitoring in ATLAS,
JINST **13** (2018) P07017 (cit. on p. 119).

BIBLIOGRAPHY

- [143] ATLAS Collaboration, *Luminosity determination in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector at the LHC*, ATLAS-CONF-2019-021, 2019, URL: <https://cds.cern.ch/record/2677054> (cit. on p. 119).
- [144] ATLAS Collaboration, *Performance of b-jet identification in the ATLAS experiment*, JINST **11** (2016) P04008, arXiv: [1512.01094 \[hep-ex\]](#) (cit. on p. 121).
- [145] ATLAS Collaboration, *Evaluation of theoretical uncertainties for simplified template cross section measurements of V-associated production of the Higgs boson*, ATL-PHYS-PUB-2018-035, 2018, URL: <https://cds.cern.ch/record/2649241> (cit. on pp. 123, 124).
- [146] L. Lyons, *Statistics for nuclear and particle physicists*, University of Cambridge, 1986, ISBN: 0 521 255406, URL: <https://www.cambridge.org/highereducation/books/statistics-for-nuclear-and-particle-physicists/9544B39F3244D9457BEC324CD34F1571#overview> (cit. on pp. 135, 136).
- [147] R. J. Barlow and C. Beeston, *Fitting using finite Monte Carlo samples*, Computer Physics Communications **77** (1993) (cit. on p. 136).
- [148] G. Cowan, K. Cranmer, E. Gross and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, Eur. Phys. J. C **71** (2011) 1554, arXiv: [1007.1727 \[physics.data-an\]](#) (cit. on p. 136), Erratum: Eur. Phys. J. C **73** (2013) 2501.
- [149] A. Hoecker et al., *TMVA - Toolkit for Multivariate Data Analysis*, 2007, arXiv: [physics/0703039 \[physics.data-an\]](#) (cit. on p. 159).
- [150] CMS Collaboration, *Inclusive Search for a Highly Boosted Higgs Boson Decaying to a Bottom Quark–Antiquark Pair*, Phys. Rev. Lett. **120** (2018) 071802, arXiv: [1709.05543 \[hep-ex\]](#) (cit. on p. 162).
- [151] CMS Collaboration, *Simplified template cross section measurements of Higgs boson produced in association with vector bosons in the $H \rightarrow b\bar{b}$ decay channel in proton-proton collisions at $\sqrt{s} = 13$ TeV*, (2022), URL: <https://cds.cern.ch/record/2827421> (cit. on p. 162).
- [152] ATLAS Collaboration, *A detailed map of Higgs boson interactions by the ATLAS experiment ten years after the discovery*, Nature **607** (2022), arXiv: [2207.00092 \[hep-ex\]](#) (cit. on p. 162).
- [153] CMS Collaboration, *A portrait of the Higgs boson by the CMS experiment ten years after the discovery*, Nature **607** (2022) 7917, 60, arXiv: [2207.00043 \[hep-ex\]](#) (cit. on p. 162).