Studies of Identification of Boosted Higgs Decaying Heavy Flavor Quarks in the LHC ATLAS Experiment
（LHC-ATLAS 実験における 重いフレーバークォークに崩壊する ブーストされたヒッグス粒子の識別に関する研究）

Thesis submitted to

The University of Tokyo

for the degree of

Master of Science

in

Department of Physics

by

Jian Wu

January, 2024

Acknowledgements

First of all, I would like to extend my deepest gratitude to my supervisor, Prof. Junichi Tanaka, for his continuous guidance and support throughout my research. His expertise in the field of particle physics and his patience in explaining complex concepts has been invaluable to my academic journey.

Special thanks also go to Sanmay Ganguly for his expertise in physics analysis and valuable advice about machine learning techniques. I would like to express my gratitude to some members of the ATLAS Flavor-Tagging Xbb sub-group, particularly to Arely Cortes Gonzelez, Jinchao Zheng, Adam Campbell Anderson and Samuel Van Stroud. Their experiences on the tagger development gave me a lot of inspiration and constructive suggestions on the detailed implementation of my research. I also express my profound gratitude for the contributions of other ICEPP members, particularly Masahiko Saito, Takuya Nobe and Masahiro Morinaga. They have provided me with a lot of valuable advice in my research to help me overcome many technical difficulties.

Additionally, I would like to thank the senior students and my friends in the Tanaka Lab: Tingyu Zhang, Jiaqi Zang, Marin Furukawa. Their willingness to share technical expertise extended beyond the confines of formal research discussions, encompassing aspects of daily life in Japan, offering a holistic support system that significantly contributed to my overall experience.

Furthermore, my gratitude extends to other friends outside the academic realm in Japan, for their support and encouragement during my studies, and for making my academic journey not only enriching but also enjoyable. Finally, my heartfelt thanks to my family for their unconditional love and unwavering support, sustained me through the challenges of academic rigor and also provided the emotional foundation necessary for my life.

Abstract

The Higgs boson is a crucial part of the Standard Model (SM) of particle physics, to explain the origin of the mass of elementary particles. The Large Hadron Collider (LHC) at CERN discovered the Higgs boson in 2012, and since then, detailed studies of the Higgs boson are being performed. Thus, for precise measurement of the Yukawa coupling of the Higgs boson to bottom and charm quarks, the identification of the Higgs boson, especially those generated with high transverse momentum, is a crucial part of the physics program. Recently, a novel $X \to b\bar{b}/c\bar{c}$ tagger is developed to tagging large-R jets containing boosted b-hadrons or c-hadrons, which may be of great importance for future Higgs studies.

This thesis covers the introduction, calibration, and improvement attempts of this novel $X \to b\bar{b}$ tagger, called GN2X. The GN2X tagger efficiency is determined using only MC, thus data-to-simulation efficiency correction factors (scale factors) need to be extracted. The in-situ calibration of the $X \to b\bar{b}$ tagger is performed to determine the scale factors using $Z(\to b\bar{b}) + \text{jets}$ and $Z(\to \ell\ell) + \text{jets}$ events for $p_T > 450\,\text{GeV}$ with data collected by the ATLAS experiment in Run 2 ($140\,\text{fb}^{-1}$). The dominant background, dijet events, is modeled by fitting the data directly using an exponentiated polynomial or polynomial function. The scale factors at 60% working point are measured to be 1.30 ± 0.50 for $450 < p_T < 500\,\text{GeV}$, 0.83 ± 0.27 for $500 < p_T < 600\,\text{GeV}$, and 0.87 ± 0.27 for $600 < p_T < 1000\,\text{GeV}$.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

In this chapter, the SM is introduced in Section 1.1, and the Higgs measurement at LHC is summarized and the role of the VH channel is briefly described in Section 1.2. In Section 1.3, the boosted Higgs is discussed, in particular, $H \to b\bar{b}/c\bar{c}$. Then, the purpose of this thesis is given. Section 1.4 explains the structure of this thesis.

1.1 The Standard Model

1.1.1 Elementary Particles

The Standard Model (SM) of particle physics describes elementary particles and interactions between them. In SM, there are two types of elementary particles, fermions and bosons, which are shown in Figure 1.1.

Quarks and leptons are fermions, which have 1/2 spin and build matter. There are 3 generations of quarks and leptons, and each particle has an antiparticle. The first generation of quarks and leptons are the lightest and most stable of three generations. The second and third generations are heavier and except neutrinos, they can decay into the first generation, which means that they can only be produced in high-energy environments.

Gauge bosons are bosons, which have 1 spin and mediate interactions. There are 3 forces in the SM: the electromagnetic force, the weak force and the strong force, which are mediated by photon (γ), weak bosons ($W^{\pm}, Z$) and gluons (g) respectively. Graviton (G) is the hypothetical gauge boson of gravity, which is not included in the SM.

The Higgs boson (H) is the only scalar boson in the SM. It has 0 spin is responsible for the mass of elementary particles.

1.1.2 Gauge Theory

The SM is a gauge theory with the symmetry group $SU(3)_C \times SU(2)_L \times U(1)_Y$. The $SU(3)_C$ group describes the strong interaction, which is mediated by gluons and acts on particles with color charge. The $SU(2)_L \times U(1)_Y$ group describes the electroweak interaction, which is mediated by the weak bosons and photons and acts on particles with weak isospin and weak hypercharge.

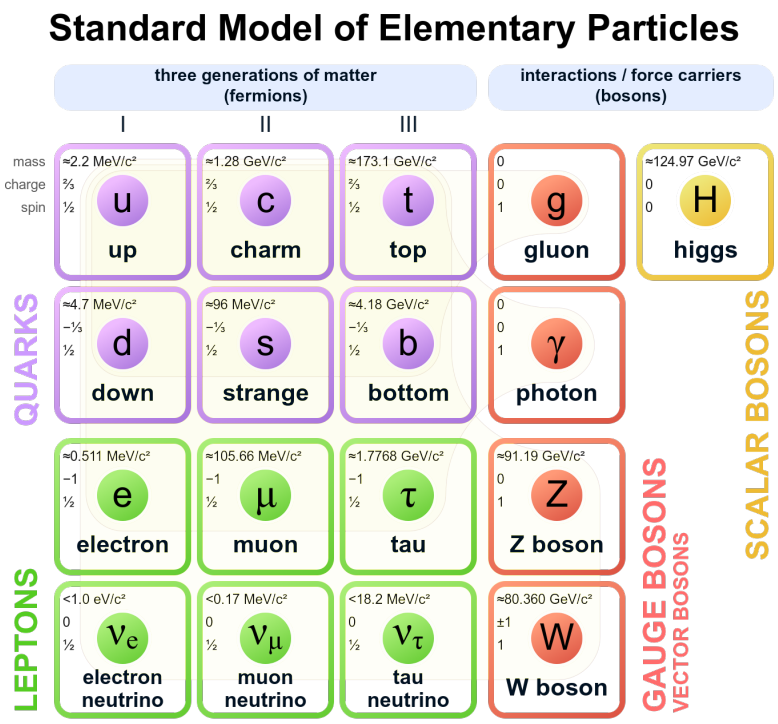


Figure 1.1 Elementary particles in the SM [1].

1.2 The Higgs Boson

1.2.1 Higgs Boson Measurement

The Higgs boson was the last missing piece of the SM, which was discovered by ATLAS and CMS experiments in 2012 [2-3]. The mass of the Higgs boson is measured to be 125.11 ± 0.11 GeV [4]. The total width is measured to be $4.5^{+3.3}_{-2.5}$ MeV [5].

For the discovery of Higgs, evidence for Higgs was present in three decay modes $H \to \gamma\gamma$, $H \to WW^{(*)} \to \ell\nu\ell\nu$ and $H \to ZZ^{(*)} \to 4\ell$ in both experiments was present [6-11]. Then, the $H \to \tau\tau$ was observed in 2016 [12] by CMS and in 2018 by ATLAS [13]. After that, Higgs production in association with a top quark-antiquark pair (ttH) [14-15] and the $H \to b\bar{b}$ was confirmed in 2018 [16-17]. The summary of the measured couplings to gauge bosons and fermions is shown in Figure 1.2. $H \to \mu\mu$ and $H \to c\bar{c}$ are not confirmed yet and the observation of such decays is one of the most important topics in Run 3 (2022-2025) and High-Luminosity LHC (HL-LHC, 2029-).

1.2.2 Higgs Production and Decay

The main Higgs production mechanisms at the LHC are as follows:

- gluon-gluon fusion (ggF)

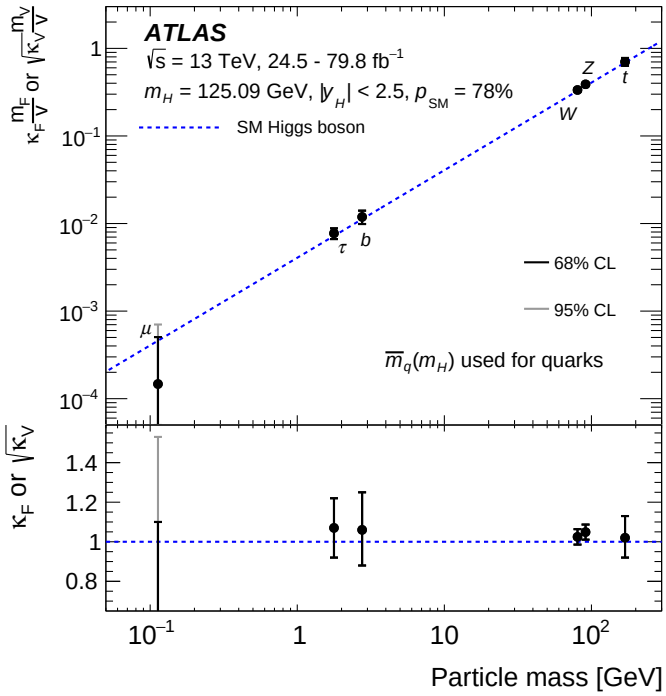


Figure 1.2 Reduced coupling strength modifiers $\kappa_{\mathrm{F}}\frac{m_{\mathrm{F}}}{v}$ for fermions ($\mathrm{F} = t, b, \tau, \mu$) and $\sqrt{\kappa_{\mathrm{V}}}\frac{m_{\mathrm{V}}}{v}$ for weak gauge bosons ($\mathrm{V} = W, Z$) as a function of their masses m_{F} and m_{V} [18].

- vector boson fusion (VBF)
- associated production with a vector boson (VH)
- associated production with a top quark pair ($\mathrm{t\bar{t}H}$)

The detailed Feynman diagrams of these processes are shown in Figure 1.3. The cross-sections for the production of an SM Higgs boson as a function of its mass are shown in Figure 1.4(a). Table 1.1 shows the Higgs boson production cross-sections and relative uncertainties for a Higgs boson mass of 125 GeV at $\sqrt{s} = 7$, 8, 13 and 13.6 TeV.

Table 1.1 Production cross-sections of the $m_H = 125$ GeV Higgs boson at the LHC [20].

$\sqrt{s}$ (TeV)	Production cross-section (pb) for $m_H = 125$ GeV					
	ggF	VBF	WH	ZH	$\mathrm{t\bar{t}H}$	Total
7	$16.9^{+5.5\%}_{-7.6\%}$	$1.24^{+2.2\%}_{-2.2\%}$	$0.58^{+2.2\%}_{-2.3\%}$	$0.34^{+3.1\%}_{-3.0\%}$	$0.09^{+5.6\%}_{-10.2\%}$	19.1
8	$21.4^{+5.4\%}_{-7.6\%}$	$1.60^{+2.1\%}_{-2.1\%}$	$0.70^{+2.1\%}_{-2.2\%}$	$0.42^{+3.4\%}_{-2.9\%}$	$0.13^{+5.9\%}_{-10.1\%}$	24.2
13	$48.6^{+5.6\%}_{-7.4\%}$	$3.78^{+2.1\%}_{-2.1\%}$	$1.37^{+2.0\%}_{-2.0\%}$	$0.88^{+4.1\%}_{-3.5\%}$	$0.50^{+6.8\%}_{-9.9\%}$	55.1
13.6	$54.7^{+5.6\%}_{-7.4\%}$	$4.28^{+2.1\%}_{-2.1\%}$	$1.51^{+1.8\%}_{-1.9\%}$	$0.99^{+4.1\%}_{-3.7\%}$	$0.61^{+6.9\%}_{-9.8\%}$	62.1

The Higgs boson branching ratios and production cross-sections are shown in Figure

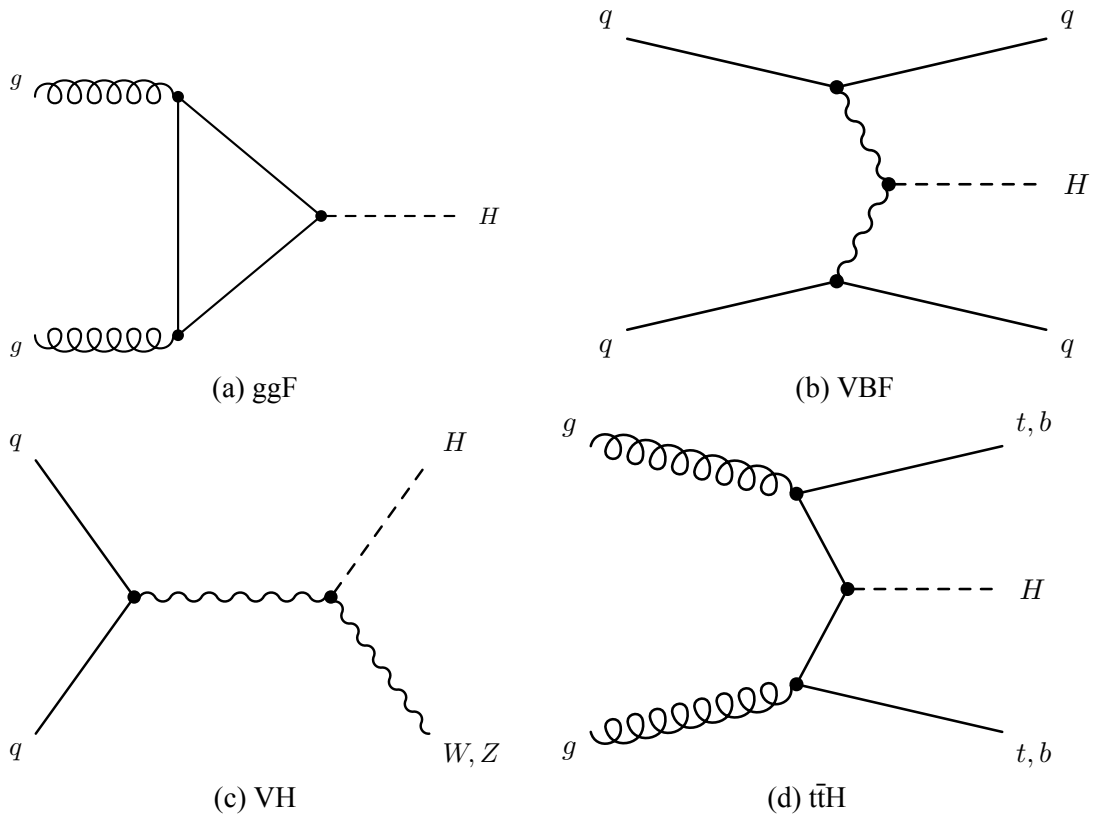


Figure 1.3 Feynman diagrams of the main Higgs boson production modes at the LHC [19].

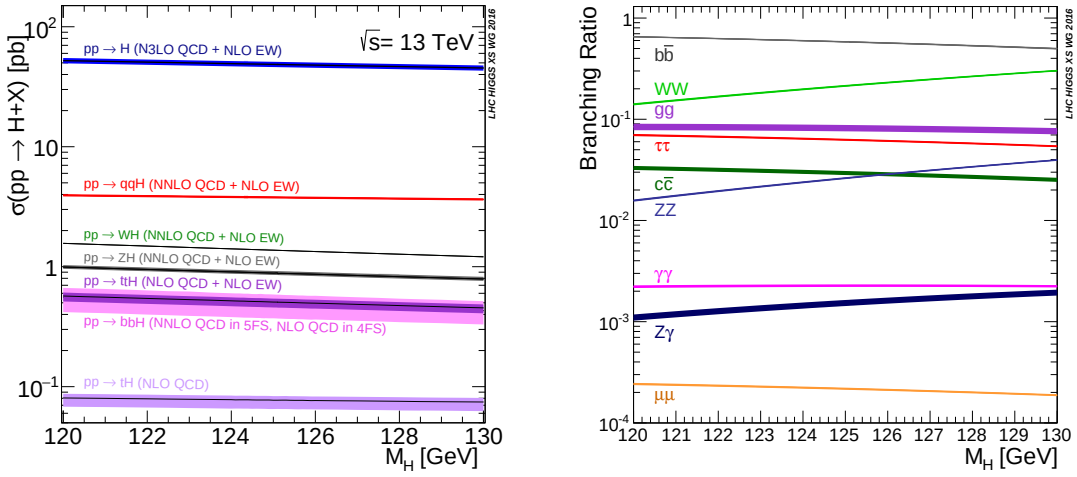
1.4(b). $H \rightarrow b\bar{b}$ is the dominant decay mode of the Higgs boson, which is 58.2% for $m_H = 125$ GeV. ggF and VBF are the 1st and 2nd largest cross-sections. However, they can not be used for $H \rightarrow b\bar{b}$ search, because of the large amount of multijet background. Therefore, the VH channel, the 3rd largest cross-section, is the most promising channel to observe the $H \rightarrow b\bar{b}$ decay. It has about 1 pb cross-section in the SM. In this channel, signal events can be selected efficiently using final state leptons from the vector boson decay.

1.2.3 VH Channel

In the VH channel, the Higgs boson is produced in association with a vector boson. There are three channels depending on number of reconstructed leptons:

- VH $\rightarrow \nu\nu$qq (V=Z, 0-lepton channel)
- VH $\rightarrow \ell\nu$qq (V=W, 1-lepton channel)
- VH $\rightarrow \ell\ell$qq (V=Z, 2-lepton channel)

Using the decay products of the gauge boson (V), we can trigger the events and also reduce the background. VH channel is the most sensitive Higgs production channel, and thus, it is one of the promising channels to observe charm Yukawa coupling using



(a) (b)

Figure 1.4 Higgs boson production cross-sections (left) and branching ratios (right) as a function of the Higgs boson mass [20].

a boosted event topology, which is explained in the next section.

1.3 Boosted Higgs

The LHC's high collision energies can lead to the generation of Higgs bosons with transverse momenta (p_T) significantly higher than their mass. There are two kinds of event topologies possible for the VH channel, the boosted and resolved topologies, as Figure 1.5 shows. The boosted Higgs bosons exhibit highly collimated decay products, and in cases of fully hadronic decays, they can be reconstructed as a single hadronic jet. For boosted $H \to b\bar{b}/c\bar{c}$ decays, it's reconstructed as a large-R jet and the b/c-quarks are reconstructed as subjets. On the contrary, with the resolved topology, the Higgs boson candidate was reconstructed from two b-tagged small-R jets in the event. Due to the highly collimated b-jets at large transverse momentum, with the resolved topology, the Higgs boson candidate reconstruction efficiency was highly degraded.

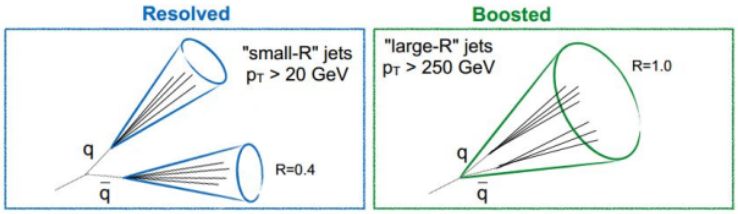


Figure 1.5 Boosted and resolved topologies.

In addition, during years of runs and upgrades, the LHC has been operated from $\sqrt{s} =$

7 to 13.6 TeV and the production of the Higgs boson at m_H = 125 GeV from pp collisions increased from 19.1 fb to 55.1 fb. A 400 fb^{-1} total integrated luminosity is expected to be collected by the end of Run 3 [1]. In the future, as the LHC gears up (HL-LHC), the instantaneous luminosity will increase, and more data of 3,000 fb^{-1} will be collected, enabling further observation and study of bottom or charm quark Yukawa couplings.

The expected number of events from WH/ZH can be evaluated using MC samples. By using ZH MC sample shown in Fig 1.6, the efficiency of the boosted Higgs ($p_T > 450$ GeV) is estimated as about 0.33 %., Because Z and W mass are similar, this value is assumed to be also applicable to WH. By using this value, the number of events can be calculated as follows:

$$N_{\text{0-lepton}} = \mathcal{L} \times \sigma \times \text{BR}(H \to bb/cc) \times \text{BR}(Z \to \nu\nu) \times \epsilon$$

$$N_{\text{1-lepton}} = \mathcal{L} \times \sigma \times \text{BR}(H \to bb/cc) \times \text{BR}(W \to \ell\nu) \times \epsilon$$

$$N_{\text{2-lepton}} = \mathcal{L} \times \sigma \times \text{BR}(H \to bb/cc) \times \text{BR}(Z \to \ell\ell) \times \epsilon$$

where ℓ means e or μ, $\mathcal{L}$ is the luminosity, σ is the cross-section and ϵ is the efficiency. Thus, a total of about 411 $VH, H \to b\bar{b}$ events and 20 $VH, H \to c\bar{c}$ events are expected to be collected in Run 3. In the HL-LHC, a total of about 3,078 $VH, H \to b\bar{b}$ events and 160 $VH, H \to c\bar{c}$ events are expected to be collected. However, to get these values, any selection except Higgs p_T is not taken into account, so they are just for the ideal case.

The identification of the boosted Higgs bosons (Xbb tagger) is performed with deep learning, which is explained in Section 5.2.5. Such a development is done based on the simulated data but it will be applied to data. So, one of the important tasks to use the boosted object identification tool is the calibration, that is, the tool is tuned to data properly. This thesis focuses on the calibration of a new Xbb tagger, GN2X, using $Z(\to b\bar{b})$+jet events. boosted $Z(\to b\bar{b}) +$ jets and $Z(\to ee/\mu\mu) +$ jets events. The GN2X [21] Xbb tagger has been developed for the coming Run 3 analysis including Run 2 data. The method used is based on the previous study [22] but uses a new Xbb tagger and MC simulated data for Run 3 analysis with a new ATLAS software release.

1 In Fig 2.2, the expected integrated luminosity is 450 fb^{-1} at the end of Run 3 but due to an issue with the LHC accelerator in 2023, data of 70 fb^{-1} was taken in 2022-2023. So, here we assume 400 fb^{-1} instead of 450 fb^{-1}.

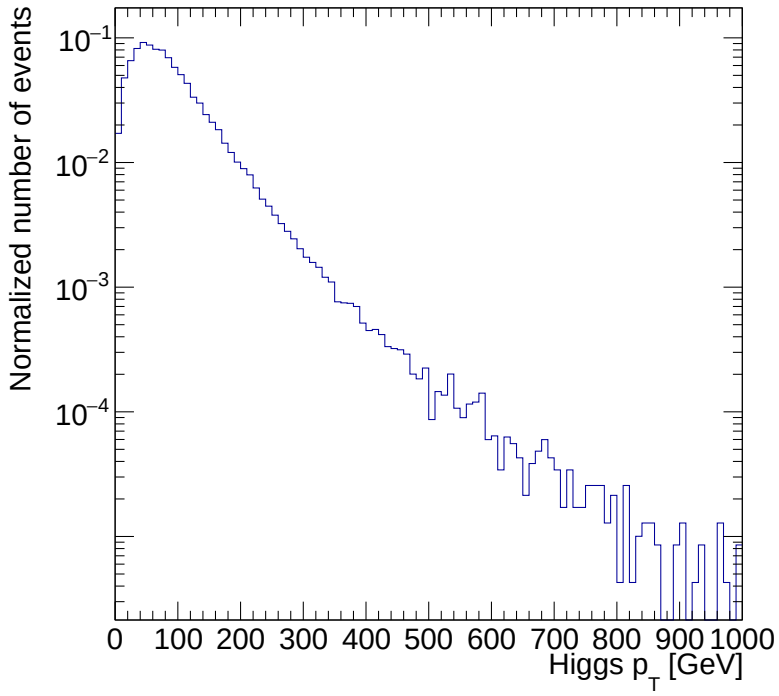


Figure 1.6 Normalized Higgs p_T distribution in a ZH MC sample at $\sqrt{s} = 13$ TeV and $m(H) = 125$ GeV.

1.4 Thesis Structure

This thesis introduces the LHC and ATLAS experiment in Chapter 2. Chapter 3 presents the data and MC samples used in this thesis. The objection reconstruction and the jet labeling used are introduced in Chapter 4. Chapter 5 shows the methodology of the calibration for the Xbb tagger. Then, the signal and background modeling are discussed in Chapter 6. The results are presented in Chapter 7. Finally, Chapter 8 concludes this thesis and gives an outlook for future work.

Chapter 2 LHC-ATLAS Experiment

The Large Hadron Collider (LHC) [23] is the world's largest and most powerful particle accelerator, which is located at the European Organization for Nuclear Research (CERN) in Geneva, Switzerland. The ATLAS (A Toroidal LHC ApparatuS) detector [24] is a general-purpose detector at the LHC. In this chapter, the LHC is briefly introduced in Section 2.1, and the ATLAS detector is described in Section 2.2.

2.1 Large Hadron Collider

The LHC is a circular proton-proton (pp) collider with a ring with a circumference of 27 km buried 100 m underground. It is designed to accelerate two counter-rotating beams of protons to a center-of-mass energy of 13.6 TeV. The two proton beams are accelerated in opposite directions in two separate beam pipes, which are kept at ultrahigh vacuum and are brought into collision at four interaction points (IP), where the detectors are located. The accelerator complex of the LHC is shown in Figure 2.1.

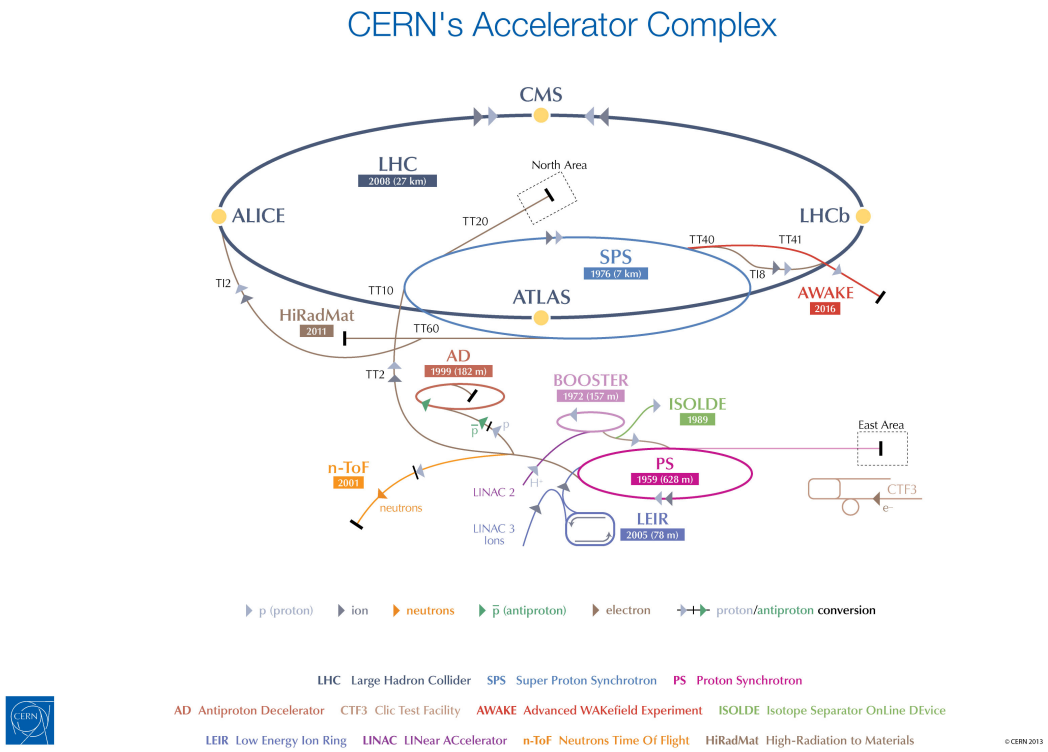


Figure 2.1 Cern's accelerator complex [25].

The luminosity of the LHC is defined by the following equation:

$$\mathcal{L} = \frac{N_b^2 n_b f_{\text{rev}} \gamma_r}{4\pi \epsilon_n \beta^*} F, \tag{2.1}$$

where N_b is the number of particles per bunch, n_b is the number of bunches per beam, f_{rev} is the revolution frequency, γ_r is the relativistic gamma factor, ϵ_n is the normalized transverse beam emittance, β^* is the beta function at the collision point, and F is the geometric luminosity reduction factor due to the crossing angle at the interaction point. The design and performance of the LHC up to 2022 are summarized in Table 2.1.

Table 2.1 Design and performance of the LHC up to 2022 [26][27].

Parameter	Design	2015	2016	2017	2018	2022
Energy (TeV)	7.0	6.5	6.5	6.5	6.5	6.8
Number of bunches	2808	2244	2220	2556-1868	2556	2748
Max. stored energy (MJ)	362	280	280	315	312	400
β^* (cm)	55	80	40	40→30	30→27→25	60
Bunch population, N_b (10^{11}p)	1.15	1.2	1.25	1.25	1.1	1.4
Normalized Emittance Stable Beams (μm)	3.75	2.6-3.5	1.8-2	1.8-2.2	1.8-2.2	1.8
Peak Luminosity ($10^{34}\text{cm}^{-2}\text{s}^{-1}$)	1.0	<0.6	1.5	2.0	2.1	1.9

The primary goal of the LHC is to search for the Higgs boson and new physics beyond the Standard Model (SM). The Higgs boson was discovered by the ATLAS and CMS experiments in 2012 [2-3].

The energy collision at 7 TeV started in 2010 (Run 1). After two years of operation, the collision energy was increased to 8 TeV in 2012 (Run 1). The LHC was shut down for two years for maintenance and upgrade. Run 2 started in 2015 and the beam energy of the LHC was increased to 6.5 TeV, which corresponds to a center-of-mass energy of 13 TeV. After Run 2, the LHC was shut down again for a three-year upgrade and started Run 3 in 2022 with 13.6 TeV.

In Run 1, the LHC delivered the collision data with a total integrated luminosity of 5.46 fb^{-1} at 7 TeV and 22.8 fb^{-1} at 8 TeV [28]. In Run 2, the LHC delivered a total integrated luminosity of 156 fb^{-1} at 13 TeV [29]. In Run 3 till 2023, the LHC delivered 70 fb^{-1} at 13.6 TeV [30].

To achieve such high luminosity, the LHC delivers as high as possible collision rates. As a direct consequence, an effect called pile-up which means multiple proton-proton interactions occur in the same or nearby bunch crossings, pollutes the final state of the collision events.

2.1.1 HL-LHC and Run 3

The LHC has been continuously upgraded to increase the collision's energy and luminosity. Figure 2.2 shows the operation and upgrade plan of the LHC. After Run 3, the LHC will enter into three years of the long shutdown 3 (LS3) for the installation of new upgrades for the High Luminosity phase of LHC (HL-LHC). The HL-LHC runs are planned to start from 2029 with a luminosity of 5 to 7.5 times the design one.

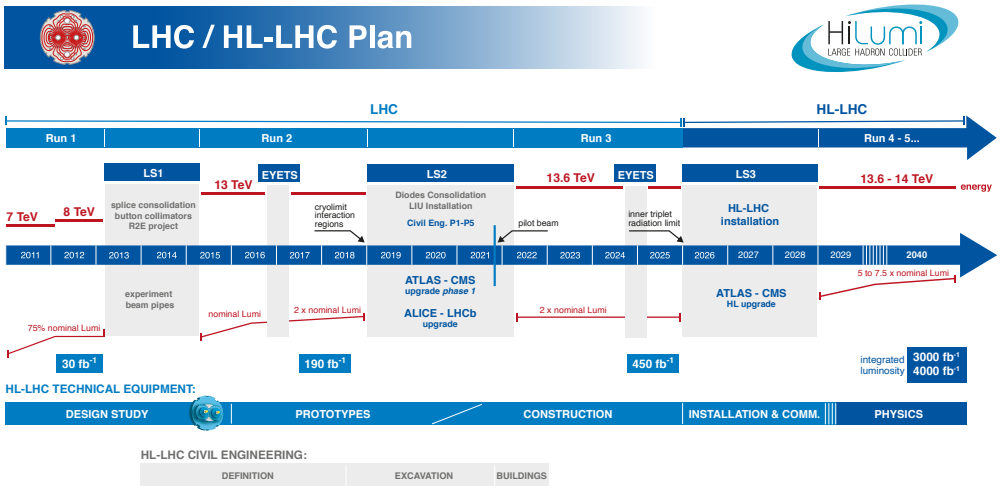


Figure 2.2 LHC/HL-LHC upgrade plan [31].

2.2 ATLAS Detector

The ATLAS experiment at the LHC features a multi-purpose particle detector characterized by a forward-backward symmetric cylindrical design with comprehensive coverage nearing 4π in solid angle. The detector comprises an inner tracking detector (ID) enveloped by a slender superconducting solenoid, alongside electromagnetic and hadronic calorimeters, a muon spectrometer (MS), and a trigger and data acquisition system (TDAQ). Notably, the ATLAS detector spans a length of 44 meters, boasts a diameter of 25 meters, and impressively weighs 7000 tons. The schematic view of the ATLAS detector is shown in Figure 2.3.

The nominal interaction point (IP) is located at the center of the detector, defining the origin of the ATLAS coordinate system right-handed. The x-axis points from the IP to the center of the LHC ring. The y-axis points upward. The z-axis is defined by the beam direction. Therefore the x-y plane is the transverse plane. Additionally, on this plane, the

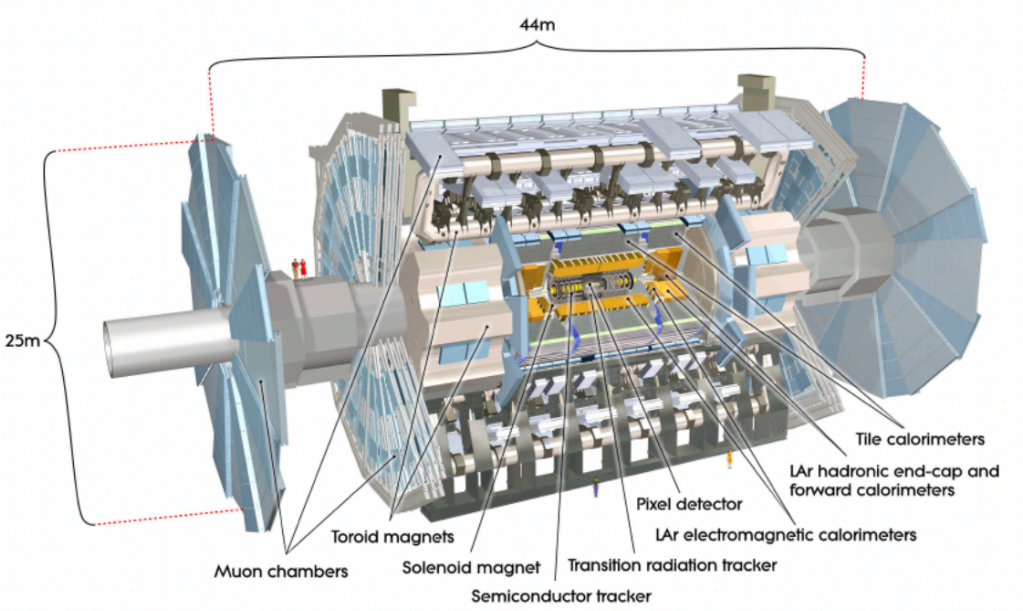


Figure 2.3 Schematic view of the ATLAS detector and its main components [24].

coordinates ϕ and r are defined as the azimuthal angle and the radial distance from the beam axis, respectively.

The pseudorapidity η for particles coming from the primary vertex [32] is defined as:

$$\eta = -\ln \tan(\theta/2),$$

where θ is the polar angle with respect to the z-axis. Therefore the η is zero for particles traveling along the x-y plane, and infinity for particles traveling along the z-axis.

The rapidity y is defined as:

$$y = \frac{1}{2} \ln \frac{E + p_z}{E - p_z},$$

where E is the energy and p_z is the momentum along the z-axis. Differences in rapidity are invariant under Lorentz boosts along the z-axis and the density of emitted particles as a function of rapidity $\frac{dN}{dy}$ is invariant under longitudinal boosts. For highly relativistic particles ($E, p \gg m$), the rapidity y is approximately equal to the pseudorapidity η.

To help the reconstruction of the collision events, quantities such as transverse momentum $p_T = p \sin\theta$ and transverse energy $E_T = E \sin\theta$ are used. Also, to measure the angular distance of two objects, a quantity ΔR is used:

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2},$$

where $\Delta\eta$ and $\Delta\phi$ are the differences in pseudorapidity and azimuthal angle, respectively.

As particles pass through, they interact with sub-systems that capture details like their

trajectory, momentum, and energy as hits and energy deposits (cells). The way particles interact with these sub-systems differs for each particle, as illustrated in Figure 2.4. This interaction influences the materials employed in the sub-detectors and guides the methods for particle reconstruction and identification in ATLAS.

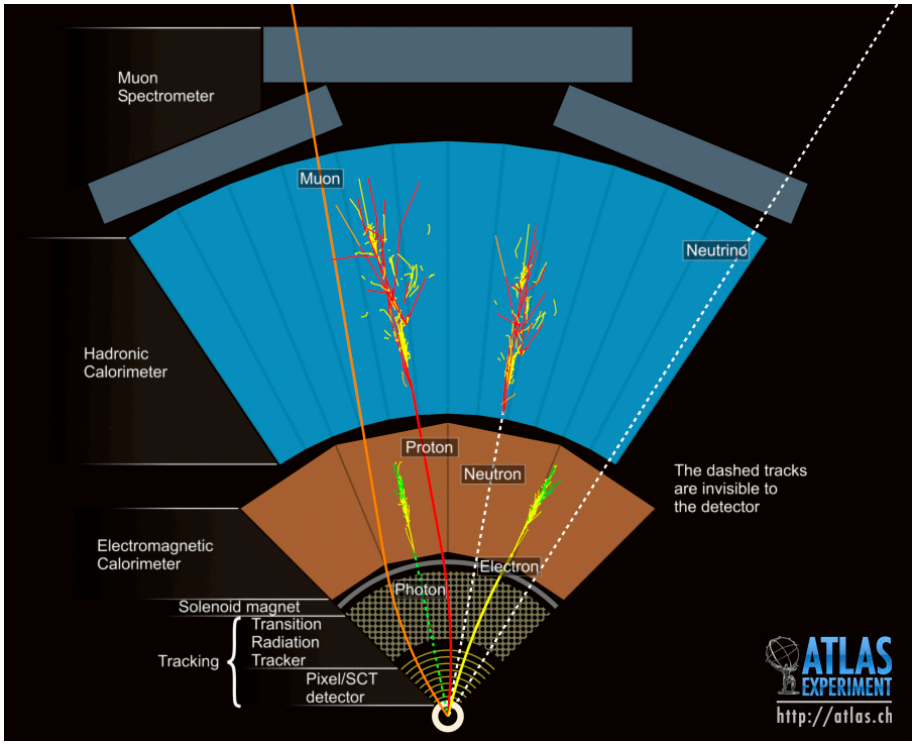


Figure 2.4 Schematic view of the interaction of different particles with the ATLAS detector [33].

In this section, the ID, calorimeters, MS and TDAQ are briefly introduced in Section 2.2.1, 2.2.2, 2.2.3 and 2.2.4, respectively.

2.2.1 Inner Detector

The inner tracking detector covers the pseudorapidity range $|\eta| < 2.5$ and provides high-precision measurements of the trajectories of charged particles in the magnetic field.

The ID is immersed in a 2 T axial magnetic field provided by a thin superconducting solenoid and contained in the cylindrical envelope of length 3.5 m and diameter 1.15 m.

It consists of three sub-detectors: the pixel detector, the semiconductor tracker (SCT), and the transition radiation tracker (TRT). The overview of the ID is shown in Figure 2.5. As shown in Figure 2.6, the innermost component is the pixel detector, followed by the SCT and the TRT. The pixel detector and the SCT cover the range $|\eta| < 2.5$, while the TRT covers the range $|\eta| < 2.0$.

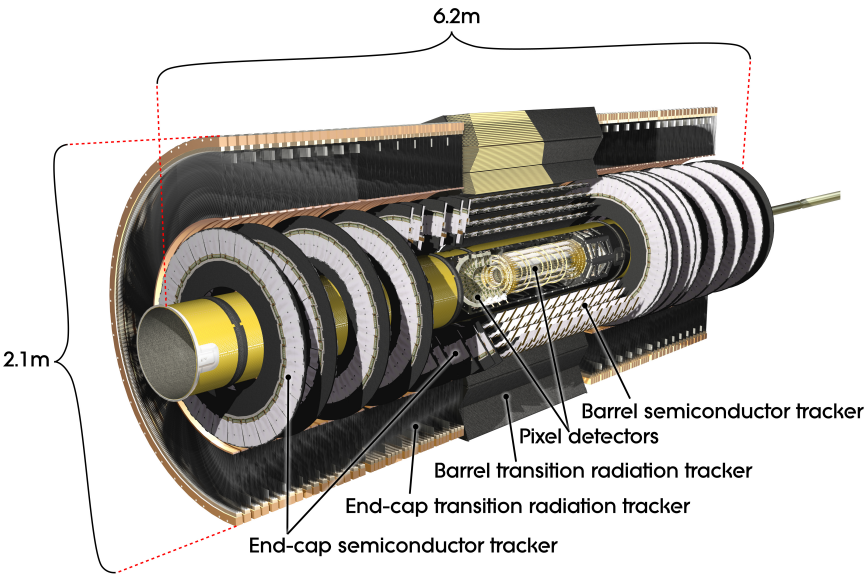


Figure 2.5 Schematic overview of the Inner Detector [34].

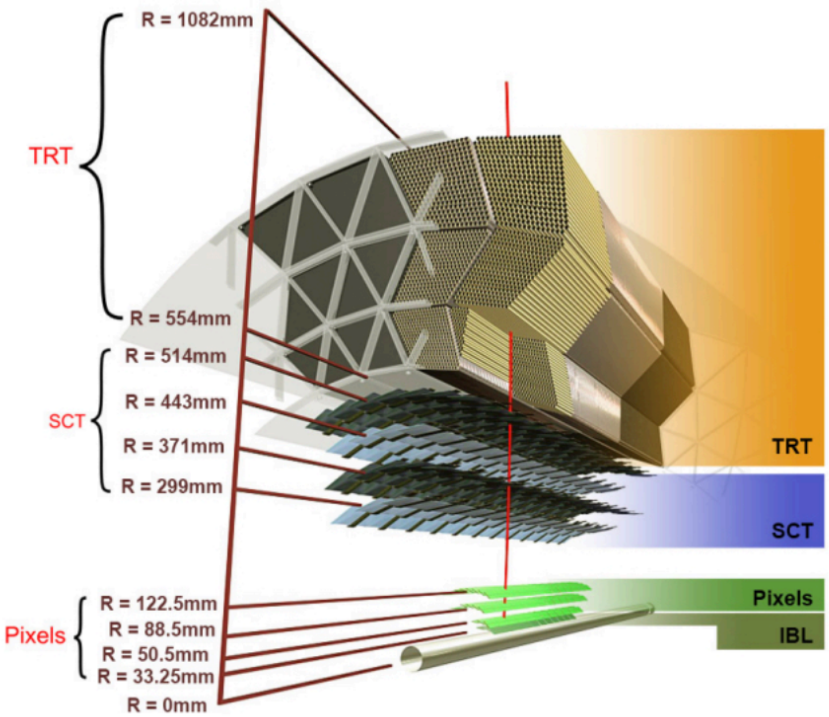


Figure 2.6 3D vision of the barrel of the Inner Detector [35].

Pixel Detector The innermost detector in ATLAS is the pixel detector, tasked with precisely measuring the trajectories of charged particles. Its role extends beyond track reconstruction, encompassing the measurement of track impact parameters relative to the primary vertex. The barrel region features 4 layers, and each end-cap region has 3 disks, resulting in a total of 80.4 million channels. The pixel's innermost layer, known as the b-layer, plays a crucial role in secondary vertex measurement.

Semiconductor Tracker The SCT is positioned following the pixel detector and consists of 4 concentric cylindrical layers of silicon microstrip detectors in the barrel region surrounding the beam axis. The end-cap region is equipped with 9 disks, each containing up to 3 rings of modules. Each SCT layer is constructed with 2 microstrips to capture hit space points, with one strip aligned parallel to the beam axis and the other rotated at a stereo angle of 40 mrad.

Transition Radiation Tracker The TRT is a tool designed for electron identification through the measurement of transition radiation. It consists of straws, each having a 4 mm diameter tube filled with a gas mixture (70% Ar, 27% CO_2, and 3% O_2 in Run 2). As charged particles pass through, they ionize the gas within the straw, causing free electrons to drift towards a centrally located gold-tungsten wire with a diameter of 31 μm. The wire is maintained at a potential of -1500 V.

2.2.2 Calorimeters

The ATLAS detector utilizes sampling calorimeters with complete ϕ coverage. The overall structure of the ATLAS calorimeter system is depicted in Figure 2.7. Electromagnetic (EM) energy measurements, characterized by high granularity, are obtained through lead/liquid-argon (LAr) sampling calorimeters. Hadronic energy measurements in the central pseudorapidity range ($|\eta| < 1.7$) are provided by a steel/scintillator-tile hadronic calorimeter known as TileCal. LAr calorimeters are employed in the endcap and forward regions to measure both electromagnetic and hadronic energies up to $|\eta| = 4.9$.

LAr Electromagnetic Calorimeter The EM calorimeter can measure electron and photon energy and a part of jet energy. It uses liquid argon (LAr) as the active material and lead as the absorber. The liquid argon is chosen because of its radiation hardness and linearity of the response. The lead absorber induces electromagnetic showers when an

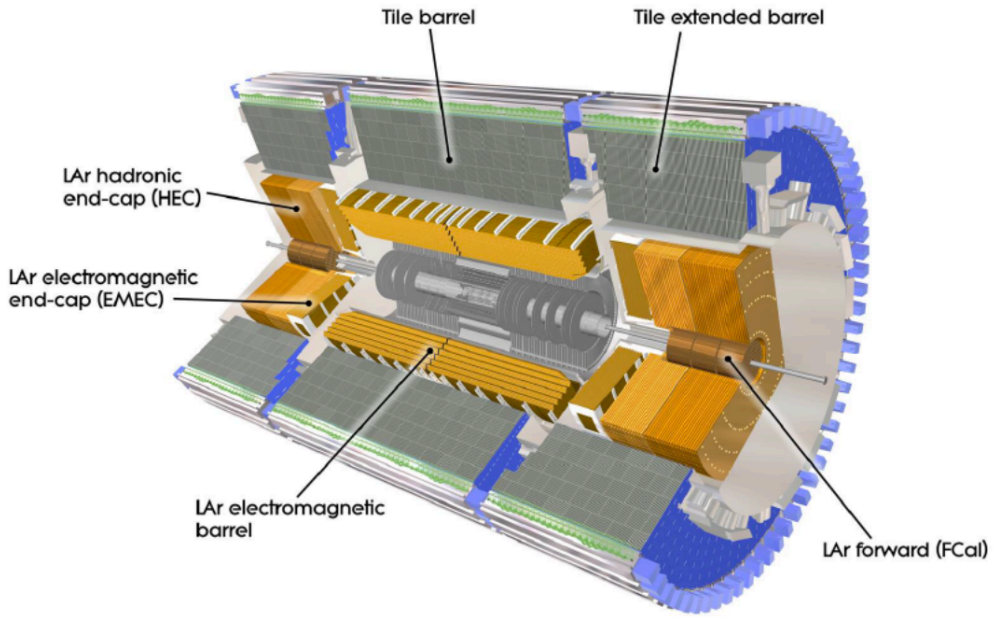


Figure 2.7 Longitudinal view of the ATLAS calorimeter system [36].

electron or a photon passes through it. The shower and the ionized particles are detected by the LAr cells and then used for the energy measurement.

The calorimeter is divided into three parts: the barrel (EMB), the end-cap (EMEC) and the forward (EMFCAL) regions. EMFCAL is described in the next paragraph of the hadronic calorimeter. The barrel region cover the pseudorapidity range $0 < |\eta| < 1.475$, while the end-cap region covers the range $1.375 < |\eta| < 3.2$. The LAr calorimeter has a four-layer structure (layers 0, 1, 2, 3), and an accordion geometry is used for the absorbers and the electrodes. Figure 2.8 shows the three-layer structure (layers 1, 2, 3) of the LAr calorimeter and the accordion geometry.

The front layer (layer 1) is highly granular of η to measure the start of the showers and to distinguish between single isolated photons and 2 photons from π^0 decays. The middle layer (layer 2) collects the majority of the shower energy. The back layer (layer 3) is used to contain the showers and measure the energy leakage beyond the middle layer. Additionally, the resampler (layer 0) is used to provide a measurement of the energy lost in front of the EM calorimeter in the range of $0 < |\eta| < 1.8$. The resolution of the EM calorimeter is parameterized as:

$$\frac{\sigma(E)}{E} = \frac{a}{\sqrt{E[\mathrm{GeV}]}} \oplus \frac{b}{E[\mathrm{GeV}]} \oplus c, \tag{2.2}$$

where a is for the stochastic behavior of the shower development, b accounts for the electronic noise and pile-up, and c quantifies the non-uniformity of the calorimeter response,

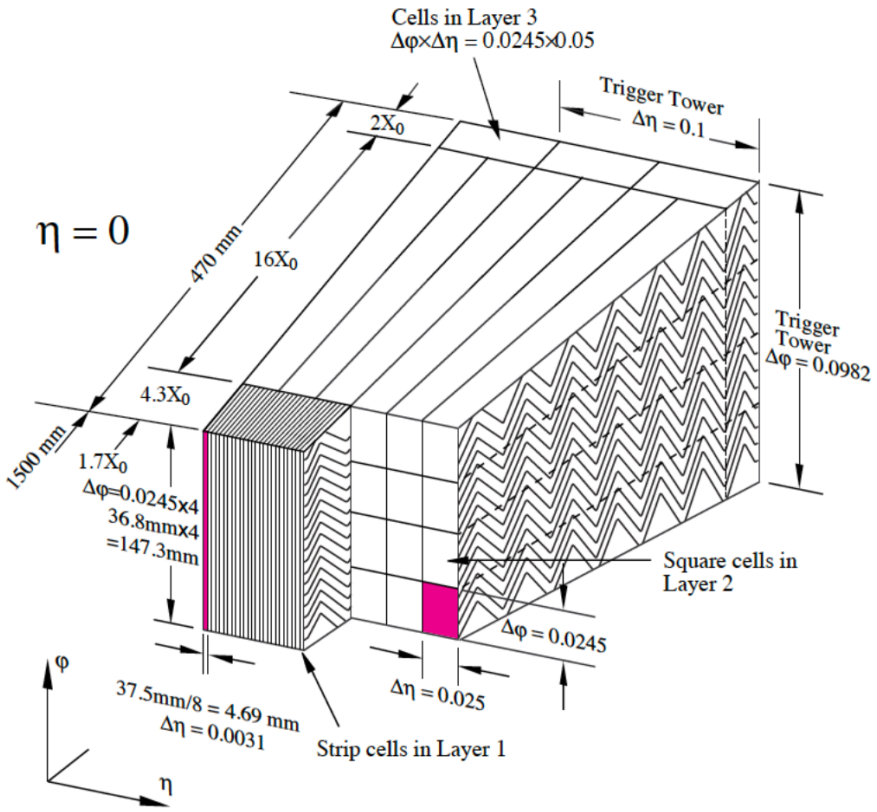


Figure 2.8 Accordion structure and the granularity of the different layers of the calorimeter [37].

aging and the radiation damage. The pile-up noise means energy deposits from other proton-proton collisions in the same or nearby bunch crossings. The intrinsic sampling term a is typically about 10% $\sqrt{\text{GeV}}$ [38], the noise term b is 10 MeV to 600 MeV without pile-up contribution [39] and is expected to be 30 MeV to 3 GeV in Run 3 pile-up conditions [40], and the constant term c is 1% to 2% [41].

Hadronic Calorimeter The hadronic calorimeter is designed to measure the energy of hadrons, which consists of the tile calorimeter (TileCal), the LAr hadronic end-cap calorimeter (HEC) and the LAr forward calorimeter (FCal).

Tile Calorimeter The tile calorimeter plays a crucial role in reconstructing the energy of hadrons and jets, as well as measuring the missing transverse energy E_T^{miss}. It spans the pseudorapidity range $|\eta| < 1.7$. Specifically, the central barrel covers $0 < |\eta| < 1.0$, while the two extended barrels span $1 < |\eta| < 1.7$. To address the transition region, special modules composed of steel scintillator sandwiches are employed. The structure of the tile calorimeter is illustrated in Figure 2.9, where each barrel is subdivided into 64 modules. The steel grid establishes a 1.5 mm modular gap at the inner radius, serving

both as the volume for readout electronics and as the flux return for the solenoid field.

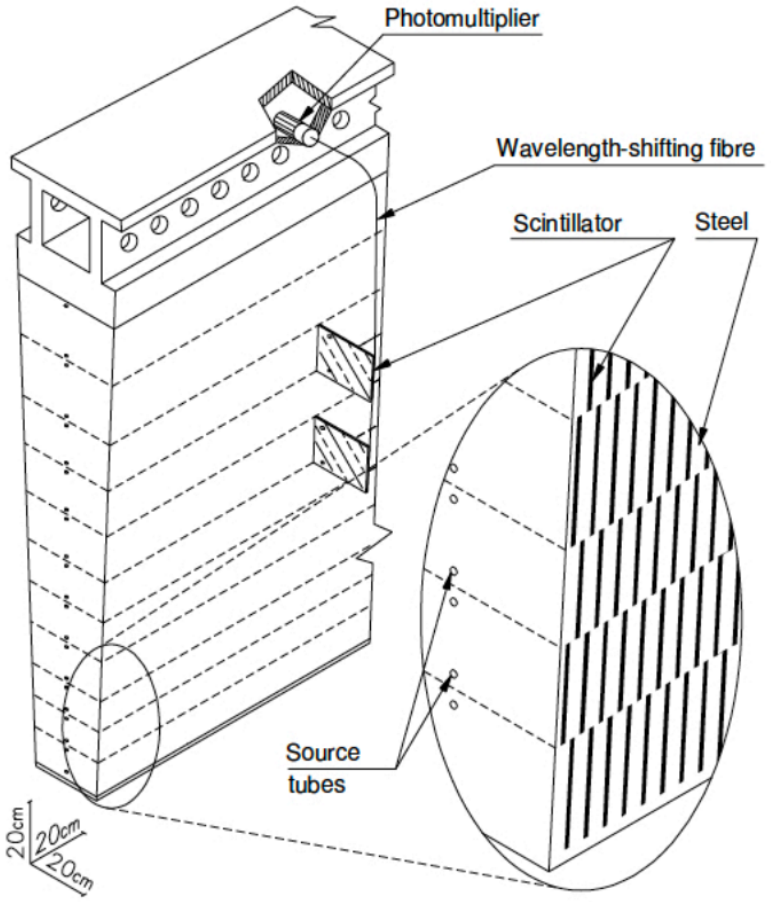


Figure 2.9 Geometry of the tile calorimeter module [24].

Hadronic End-cap Calorimeter The HEC, which is a sampling calorimeter, employs LAr as the active material and flat-plate-designed copper as the absorber integrated with the end-cap cryostat, it spans the pseudorapidity range of $1.5 < |\eta| < 3.2$. Comprising two wheels per end-cap, each wheel is made up of 32 identical wedge-shaped modules. Figure 2.10 illustrates the HEC's structure, with the front wheel containing 24 copper plates and the back wheel featuring 16 copper plates.

Forward Calorimeter The end-cap cryostat incorporates the Forward Calorimeter (FCal), covering the pseudorapidity range of $3.2 < |\eta| < 4.9$. This positioning exposes the FCal to high fluxes, necessitating a design with LAr gaps to prevent ion build-up issues and ensure maximum density. The FCal consists of three modules, illustrated in Figure 2.11. One module employs copper as the absorber to enhance resolution and heat

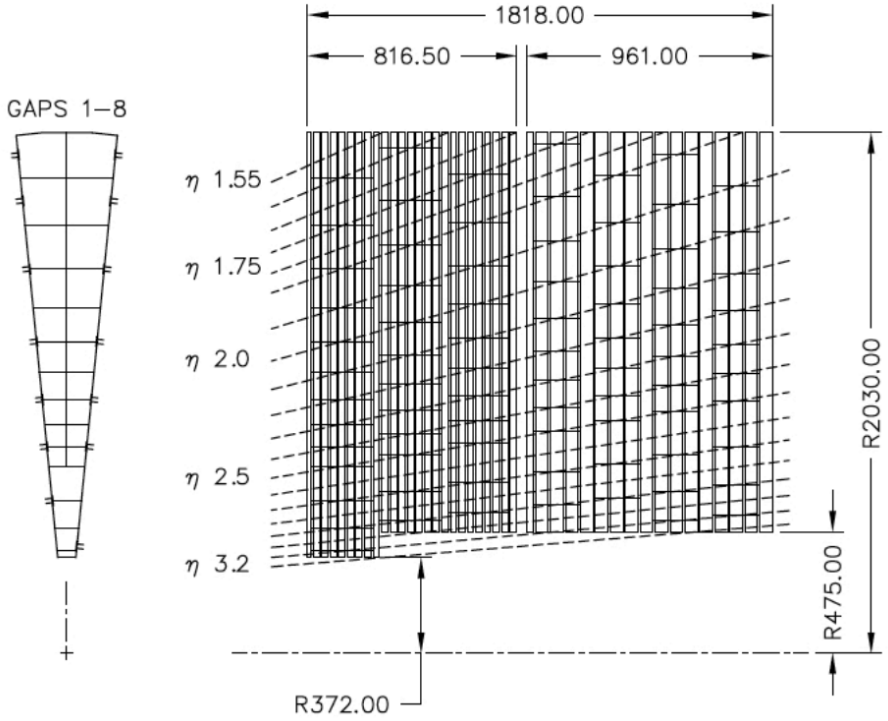


Figure 2.10 Geometry of the HEC module [24].

removal for electromagnetic measurements. The remaining two modules use tungsten as the absorber, focusing on containing hadronic showers and minimizing their spread. This configuration allows the FCal to deliver accurate measurements of forward jets while concurrently reducing background interference in the muon spectrometer.

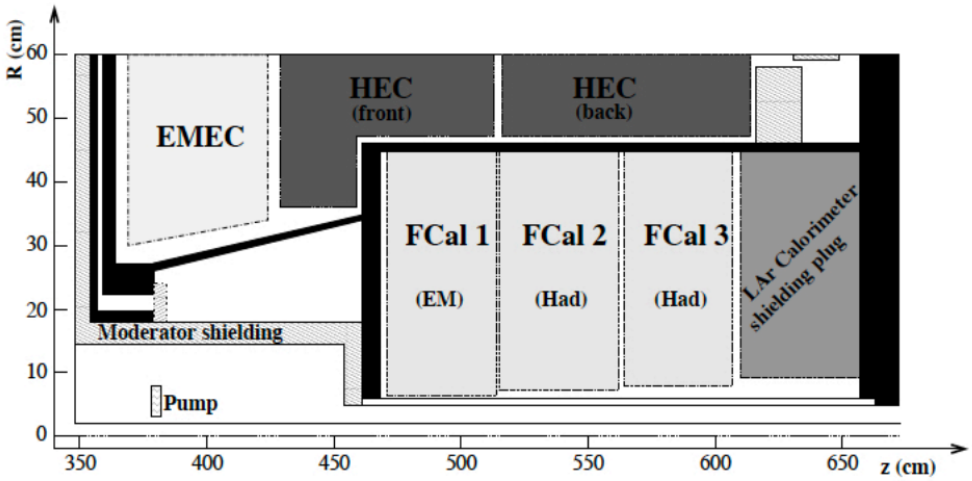


Figure 2.11 Geometry of the FCal modules located in the end-cap cryostat [24].

2.2.3 Muon Spectrometer

The muon spectrometer aims to deliver accurate standalone measurements of muons up to 3 TeV within the $|\eta| < 2.7$ range. Positioned as the outermost component of the ATLAS detector, it encases the calorimeters. The capacity to reconstruct high-momentum muons, even in the absence of ID information, facilitates rapid muon triggering in the $|\eta| < 2.4$ region.

Figure 2.12 shows the arrangement of the muon spectrometer, consisting of three large superconducting air-core toroidal magnets, each with eight coils. The field integral of the toroids ranges from 2.0 to 6.0 Tm across the muon detectors. For precise tracking, the muon spectrometer incorporates Monitored Drift Tube Chambers (MDT) and Cathode Strip Chambers (CSC). The system also includes fast detectors for triggering purposes: Resistive Plate Chambers (RPC) and Thin Gap Chambers (TGC) within the $|\eta| < 2.4$ range. This system ensures a Bunch-Crossing Identification (BCID) efficiency exceeding 99%, well-defined p_T thresholds, and the measurement of the unbending direction of muon tracks.

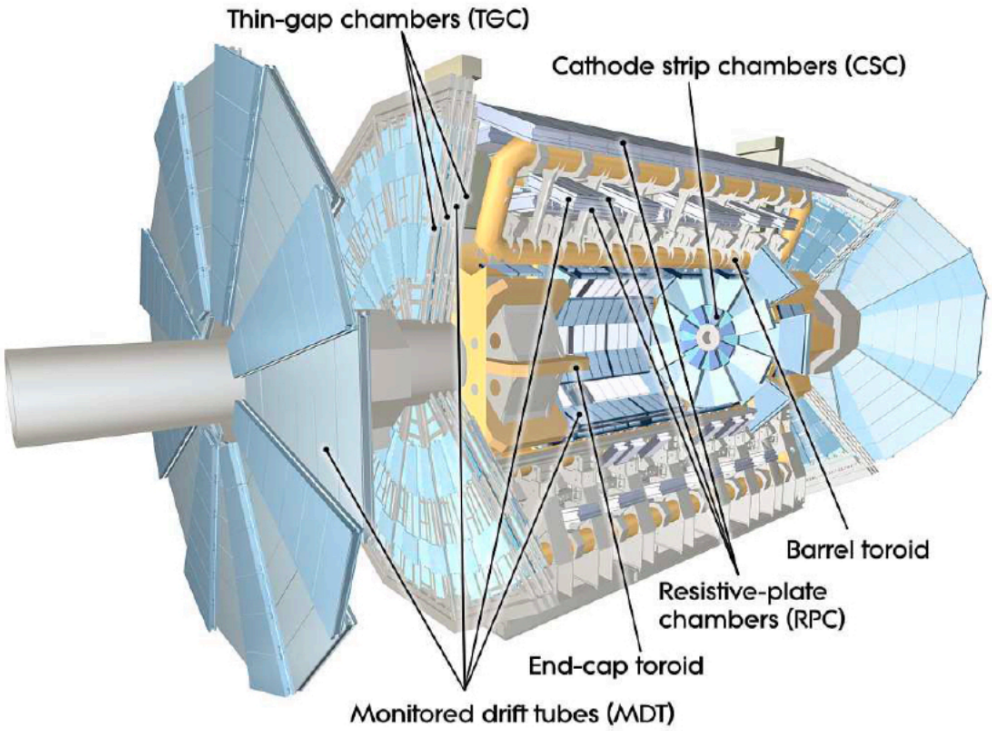


Figure 2.12 Cut-away view of the ATLAS muon spectrometer [24].

Monitored Drift Tubes The barrel MDTs cover the range $|\eta| < 1.05$, as well as the end-cap MDTs cover the range $1.05 < |\eta| < 2.7$. Their size increases with the distance from the interaction point. They are arranged in three to eight layers of drift tubes, filled with a gas mixture of 93% Ar and 7% CO_2. The average resolution for a single tube is 80 μm and for a chamber is 35 μm.

Cathode Strip Chambers The CSC is used in the forward region of the muon spectrometer, covering the range $2.0 < |\eta| < 2.7$, to improve the resolution under the high hit rate near the beam pipe. Similar to end-cap MDTs, the CSCs are arranged in 8 large sectors and 8 small sectors. Trigger time for the CSC is less than 30 ns and the average resolution is 60 μm.

Resistive Plate Chambers In the muon spectrometer's barrel region, RPCs are employed to cover the range $|\eta| < 1.05$. These RPCs consist of parallel resistive plates and a 2 μm gas-filled gap. The RPC's time resolution is approximately 1.5 nm, while the position resolution is around 1 cm.

Thin Gap Chambers The TGCs are used in the end-cap region of the muon spectrometer, covering the range $1.05 < |\eta| < 2.4$. The time resolution of the TGC is about 4 ns and the position resolution is around 2 to 6 mm.

2.2.4 Trigger and Data Acquisition System

The LHC is designed to produce a staggering 1.6 billion proton-proton collisions every second in the case of the averaged pile-up value of about 40, generating a combined data volume of approximately 60 TB/s for the ATLAS system. With protons bunching up and intersecting every 25 ns, around 40 million collisions occur per second, and each bunch carries about 10^{11} protons. Operating at the designed luminosity of 10^{34} $cm^{-2}s^{-1}$, an average of 20 interactions take place per bunch crossing. Due to this high rate, a sophisticated trigger system selectively captures only a small fraction of the collisions.

The ATLAS trigger system carries out the selection process in two stages. The first stage is the hardware-based Level-1 (L1) trigger, which reduces the event rate from the nominal 40 MHz bunch crossing rate to a maximum recording rate of 100 kHz. The second stage is the software-based High-Level Trigger (HLT), which further reduces the event rate to a few hundred Hz for permanent storage and offline analysis.

Level-1 Trigger The L1 trigger functions as a hardware-based system, utilizing information from the calorimeters and muon spectrometer at reduced granularity. Capable of processing a maximum input rate of 40 MHz, it efficiently identifies noteworthy events within a time frame of less than 2.5 μs from the collision. Subsequently, the triggered objects' geometric locations are transmitted to the next tier of the trigger system in the form of Regions of Interest (RoI) at an event rate of 100 kHz.

High-Level Trigger The HLT is a software-based system. Its operation relies primarily on commercially available computing resources and networking infrastructure. With the utilization of the complete detector granularity of the RoI supplied by the L1 trigger, the HLT is capable of managing input rates ranging from 75 to 100 kHz.

Chapter 3 Data and Monte Carlo Samples

3.1 Collision Data

The analysis targets the dataset gathered in ATLAS during Run 2 (2015-2018), which corresponds to luminosities of 36.4 fb^{-1}, 44.6 fb^{-1}, 58.8 fb^{-1} for triggers used in the 2015+2016, 2017 and 2018 data-taking periods, respectively [42]. The total luminosity is 140 fb^{-1}, which is obtained by applying detector operation quality requirements etc. to the recorded data as shown in Figure 3.1.

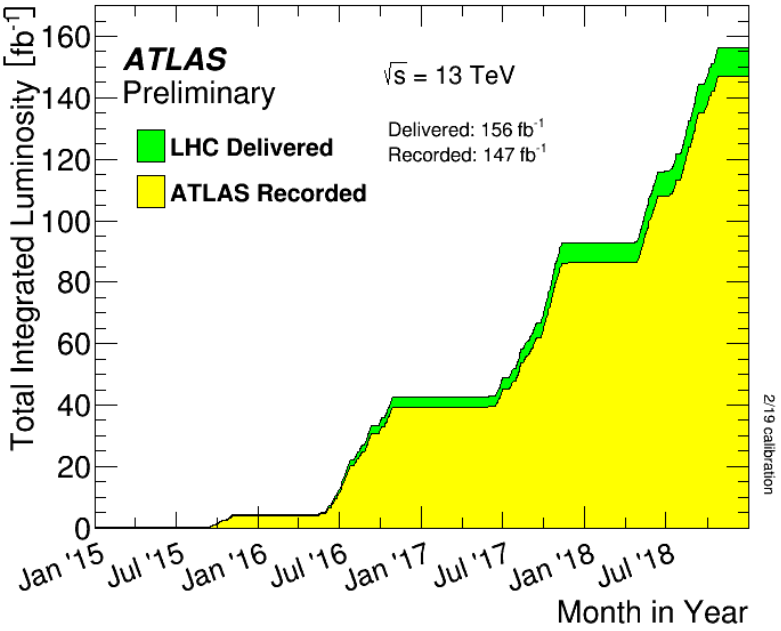


Figure 3.1 Cumulative luminosity versus time delivered to ATLAS (green) and recorded by ATLAS (yellow) during stable beams for pp collisions at 13 TeV centre-of-mass energy in LHC Run 2 [29].

3.2 Monte Carlo Samples

All simulated events were run through the full detector simulation [43] using Geant4 [44]. The effects of pile-up from multiple proton-proton interactions in the same and nearby bunch crossings are modeled by overlaying minimum bias events, simulated using the soft QCD processed of Pythia 8.186 [45] with the A3 tune [46] and the NNPDF2.3LO [47] parton distribution function (PDF) set. The average number of interactions per crossing obtained from data is used for the pile-up effect simulation. For all simulated event

samples, except for those generated using Sherpa [48], the EvtGen 1.6.0 program [49] is used to describe the decays of bottom and charm hadrons. The simulated events are processed through the same reconstruction algorithms as the data.

$Z(\rightarrow b\bar{b})$ + jets, $Z(\rightarrow q\bar{q})$ + jets ($q = u, d, s, c$) and $W(\rightarrow q\bar{q})$ + jets processes are simulated with the Sherpa 2.2.8 [50] generator using matrix elements with next-to-leading order (NLO) accuracy for one additional parton, and matrix elements accurate to leading order (LO) for up to four partons calculated with the Comix and OpenLoops libraries. They are matched with Sherpa PS [51] using the MEPS@NLO prescription and the set of tuned parameters developed by Sherpa authors. The NNPDF3.0NNLO set of PDFs [52] is used and the samples are normalized to an NNLO prediction.

The process of the multijet events, which is called multijets or dijets, is generated using Pythia 8.235 generator, using the A14 PS and hadronization tune and NNPDF2.3LO PDF set.

The process of $t\bar{t}$ is simulated with the Powheg Box v2 [53-54] generator together with the NNPDF3.0NLO [52] PDF set. The simulated events were interfaced to Pythia8.230 [46] for parton shower (PS) and hadronization using the A14 tune [55-56] together with the NNPDF2.3LO PDF set. The top-quark mass was set to 172.5 GeV and the h_{damp} parameter, which controls the transverse momentum (p_T) of the first additional emission beyond the Born configuration, was set to 1.5 times the top-quark mass. The $t\bar{t}$ production cross-section is corrected to the theory prediction calculated at next-to-next-to-leading (NNLO) order and next-to-next-to-leading log approximation. The production of $t\bar{t} + W/Z$ is simulated with the aMC@NLO 2.3.3 [57] and Pythia8.210 using NNPDF2.3NLO PDF set.

Diboson (WW, WZ and ZZ) processes were simulated with Sherpa 2.2.1, interfaced with the NNPDF3.0NNLO PDF sets for both the matrix element calculation and the PS. Sherpa provides a combination of different matrix elements with different parton multiplicities: processes with zero or one additional partons are calculated at NLO in the matrix element, while two or three additional partons are included at LO in QCD.

The $Z \rightarrow \ell\ell$ ($\ell = e, \mu$) events are generated using Sherpa 2.2.8 and with exactly the same setup as for the hadronically decaying Z events. $Z \rightarrow \ell\ell$ events have been generated with MadGraph using multileg leading order approach up to 4 additional parton in the matrix element with the ckkw merging approach. Pythia 8.230 has been used in the parton shower.

Chapter 4 Object Reconstruction and Jet Labeling

4.1 Jets

4.1.1 Large-R Jets

A boosted object originating from $H \to b\bar{b}/c\bar{c}$, $Z \to b\bar{b}/c\bar{c}$, $W \to q\bar{q}$, etc. are reconstructed as a single object, which is called large-R jet. In this analysis, the inputs to the large-R jet reconstruction algorithm are Unified Flow Objects (UFOs) [58]. UFOs are obtained from merging Particle-Flow Objects (PFOs) [59] and Track-Calo Clusters (TCCs) [60]. They have neutral and charged components. The PFO and TCC are constructed from both calorimeter clusters and tracks but their algorithms are different. The former is a typical method of so-called particle-flow, that is, cells or clusters matched to charged tracks are considered as charged components and the remaining as neutral. On the other hand, the latter is an algorithm specialized for high p_T objects, that is, the direction is obtained from a charged track and the energy from a cluster. The large-R jets are built using the anti-k_T algorithm [61] with radius parameter $R = 1.0$ implemented in FastJet [62]. Pile-up and underlying event contributions are removed via grooming with the Soft-Drop algorithm [58] along with Constituent Subtraction [63] and SoftKiller [64].

4.1.2 Track Jets

Smaller radius jets (subjets) are used to investigate the constitute of the large-R jets for the identification of heavy flavor hadrons. In this study, the variable-radius (VR) track jets are used as subjets. Charged tracks reconstructed from the inner detector are used to form subjets using the anti-k_T algorithm. For VR subjets, the radius parameter R is defined as a function of the jet p_T and a constant parameter $\rho = 30$ GeV, scaling as $R = \rho/p_T$. Minimum and maximum values for the radius parameter are kept as $R_{min} = 0.02$ and $R_{max} = 0.4$. The VR subjets are ghost-associated [65] to the large-R jet and are required to have $p_T > 7$ GeV.

4.2 Tracks

The reconstructed trajectories of charged particles, known as tracks, are reconstructed by utilizing hits that these particles create while passing through the inner detector. The

precise reconstruction of tracks is a critical task because they play a crucial role in reconstructing various other objects such as vertices and jets. The process of track reconstruction in the Inner Detector involves employing a series of algorithms, as detailed in Ref. [66-67].

Tracks are described using a perigee representation, employing five parameters and a reference point, as shown by Figure 4.1. The reference point utilized corresponds to the average position of the proton-proton interactions (beamspot position). The five parameters are outlined below:

- The transverse impact parameter d_0, representing the projection of the point of closest approach in the transverse direction.
- The longitudinal impact parameter $z_0 \sin\theta$, denoting the projection of the point of closest approach along the z-axis direction.
- The inverse transverse momentum p/p_T, providing information about the curvature of the particle track.
- The track azimuthal angle ϕ, indicating the direction of the track in the r-ϕ plane at the point of closest approach; ϕ ranges between $[-\pi, \pi]$.
- The track polar angle θ, varying within the range $[0, \pi]$.

d_0, $z_0 \sin\theta$, p/p_T are primary parameters to evaluate the tracking performance.

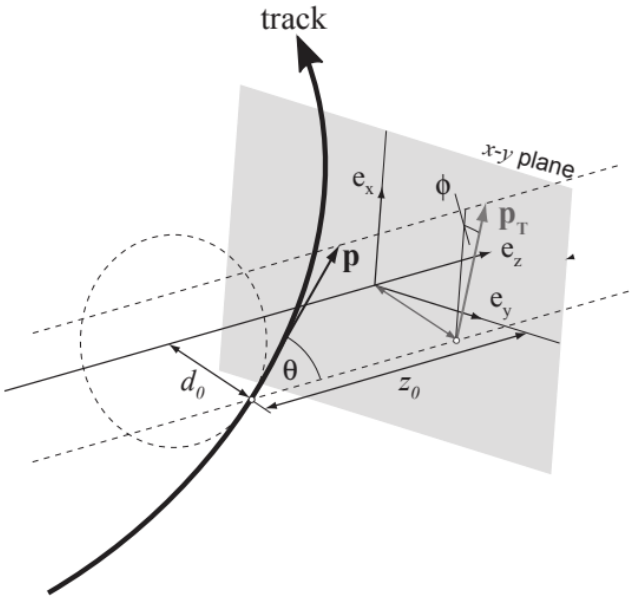


Figure 4.1 The perigee representation [68].

Tracks are ghost-associated to the large-R jet and are required to satisfy the track selection summarised in Table 4.1.

Table 4.1 Track selection requirements, where d_0 is the transverse impact parameter (IP) of the track, z_0 is the longitudinal IP with respect to the primary vertex and θ is the track polar angle. Shared hits are hits used in the reconstruction of multiple tracks. A hole is a missing hit, where one is expected, on a layer between two other hits on a track.

Parameter	Selection
p_T	> 500 MeV
$\lvert d_0 \rvert$	< 3.5 mm
$\lvert z_0 \sin\theta \rvert$	< 5.0 mm
Silicon hits	≥ 8
Shared silicon hits	< 2
Silicon holes	< 3
Pixel holes	< 2

4.3 Primary Vertex

The process of reconstructing interaction vertices involves utilizing a set of reconstructed tracks. Tracks may originate either from a primary vertex (PV), which is the interaction point between two partons, or from a secondary vertex (SV) resulting from particle decays, photon conversions, or hadronic interactions.

Generally, the reconstruction of the PV occurs through three sequential steps. Initially, a seed position is determined based on the beam spot in the transverse plane. Subsequently, the selected tracks and seeds are fitted to establish the best-estimated vertex position. Once the vertex position is defined, tracks incompatible with the vertex are eliminated. The discarded tracks are then employed to initiate a new PV, and this process is reiterated until there are no unassociated tracks remaining or no additional vertices can be identified. It is a requirement for PVs to be associated with a minimum of two tracks. Additionally, a single track may be linked to multiple vertices. The PV exhibiting the highest sum of squared transverse momenta for its associated tracks is designated as the hard-scatter PV, and the remaining PVs are categorized as interaction vertexes from the pile-up.

4.4 Muons

Muon tracks are reconstructed independently in the inner detector and the muon spectrometer. They are required to have a minimum number of hits in each system, and must be compatible in terms of geometrical and momentum matching. Finally, the information from both the inner detector and muon spectrometer systems is used in a combined fit to

refine the muon momentum measurement. In this analysis, muons are required to pass ”medium” identification and ”tight” track-based isolation criteria. Muons are required to have a transverse momentum above 20 GeV and $|\eta| < 2.5$. The muon reconstruction efficiency and the muon momentum scale and resolution corrections are measured in data and applied to the simulated events.

4.5 Electrons

Electrons are reconstructed from energy deposits in the calorimeter matched to inner detector tracks. In order to select and identify electrons, requirements are imposed on the track properties and quality, the shape of the clusters of calorimeter energy deposits and the track-to-cluster match. In this analysis, the identification is performed using a likelihood technique and is required to pass a ”medium” working point. Electrons are further required to be isolated by imposing the ”tight” isolation working point requirement. This isolation requires the absence of nearby tracks within a p_T-dependent variable-size cone around the electron. The electron energy scale is calibrated in data, and the energy resolution is calibrated in MC, using $Z \rightarrow e^+e^-$ events. Electrons are required to have a transverse momentum above 20 GeV and $|\eta| < 2.5$, excluding the crack region ($1.37 < |\eta| < 1.52$). The significance of the transverse impact parameter, $|d_0|/\sigma(d_0)$, is required to be less than 5.0. The quantity $|z_0 \sin\theta|$ is required to be less than 0.5 mm, where z_0 is the longitudinal impact parameter, to ensure that electrons are consistent with having been produced at the primary vertex.

4.6 Overlap Removal

The following overlap removal procedure is applied to resolve ambiguities in which multiple electrons, muons or jets are reconstructed from the same detector signature. First, any electrons that share a track are removed. Second, if an electron and a muon share a track, the electron is rejected if the muon is associated with a signature in the muon spectrometer, otherwise, the muon is rejected. Third, any jet within $\Delta R < 0.2$ of an electron is rejected. Fourth, any jets within $\Delta R < 0.2$ of a muon are rejected if they have fewer than three associated tracks. Fifth, any electrons or muons within $\Delta R < \min(0.4, 0.04 + 10\ \text{GeV}/p_T^{lep})$ of a jet passing the previous requirements are rejected. Finally, any large-R jet found within $\Delta R < 1.0$ of an electron is rejected.

4.7 Large-R Jets Labeling

The large-R jet truth labeling defines the acceptance of the analysis. Truth matching is done in two steps: [69]

- First, match a truth particle to truth jets such that ΔR(truth particle, truth jet) < 0.75, and get the truth label of the truth jets
- Further match a reconstructed VR jet to a truth jet such that ΔR(truth jet, reco jet) < 0.75, and copy the truth label of the matched truth jet to the reco jet

Table 4.2 shows the truth label used in the analysis and its requirements.

Table 4.2 Truth label used in the analysis and its requirements.

Num.	Label	Requirements
1	tqqb	There are a matched top quark and W boson; there is at least one B-hadron; the truth jet mass is larger than 140 GeV.
2	Wqq	There is at least one c or u-quark in the first 2 leading partons; and their invariant mass is in the 60-140 GeV range; the truth jet mass is in the 50-100 GeV range.
3	Zbb	There're at least 2 ghosted associated b-hadrons; and their invariant mass is in the 50-140 GeV range; the truth jet mass is in the 50-110 GeV range.
4	Zcc	There're at least 2 ghosted associated c-hadrons; and their invariant mass is in the 50-140 GeV range; the truth jet mass is in the 50-110 GeV range.
5	Zqq	The first 2 leading partons are with opposite charge; and their invariant mass is in the 50-140 GeV range; the truth jet mass is in the 50-110 GeV range.
6	Wqq from top	There are a matched top quark and W-boson; there is no b-hadron; the truth jet mass is in the 50-100 GeV range.
7	other from top	There is only a matched top quark.
8	other from V	There is at least one c or u-quark in the first 2 leading partons OR the first 2 leading partons have opposite charge; their invariant mass is required in the 60-140 GeV range.
9	no truth	There is no matched truth jet.
10	QCD	The truth jets are not matched to any heavy particles.

As explained in Section 3.2, There are MC samples of $Z(\rightarrow b\bar{b})+$jets, $Z(\rightarrow q\bar{q})+$ jets ($q = u, d, s, c$), $W+$ jets and $t\bar{t}$ etc. in this analysis. In the MC events, there are different types of large-R jets in one specific process, for example, even in $Z(\rightarrow b\bar{b})+$jets, there are large-R jets of Zbb, Zqq or even QCD. So, the following templates (a set of large-R jets) are prepared using the following truth-matching requirement. The details fraction of each truth label in each sample is shown in Table 4.3.

- $Z \to b\bar{b}$ template: It is created from the $Z(\to b\bar{b})$ + jets sample. The large-R jets are required to be labeled as Zbb.
- $Z(\to q\bar{q})$ + jets $(q = u, d, s, c)$ template: It is created from the $Z(\to q\bar{q})$ + jets sample. The large-R jets are required to be labeled as Zqq or Zcc.
- $W \to q\bar{q}$ template: It is created from the $W(\to q\bar{q})$ + jets sample. The large-R jets are required to be labeled as Wqq or Other from V.
- $t\bar{t}$ template: It is created from the $t\bar{t}$ sample. The large-R jets are required to be labeled as tqqb, Wqq from top or other from top.
- dijet template: It is created from the dijet sample. The large-R jets are required to be labeled as QCD.

Table 4.3 The fraction of each truth label of the leading large-R jet in $Z \to b\bar{b}$, $Z \to q\bar{q}$ $(q = u, d, s, c)$, $W \to q\bar{q}$, dijets and $t\bar{t}$ samples.

	Leading large-R jets fraction					Sub-leading large-R jets fraction				
Truth labels	$Z \to b\bar{b}$	$Z \to q\bar{q}$	$W \to q\bar{q}$	$t\bar{t}$	dijet	$Z \to b\bar{b}$	$Z \to q\bar{q}$	$W \to q\bar{q}$	$t\bar{t}$	dijet
tqqb	0	0	0	0.62	0	0	0	0	0.53	0
Wqq	0	0	0.56	<0.01	0	0	0	0.34	<0.01	0
Zbb	0.48	<0.01	0	0	0	0.39	<0.01	0	0	0
Zcc	<0.01	0.12	0	0	0	<0.01	0.08	0	0	0
Zqq	0.03	0.45	0	0	0	0.02	0.26	0	0	0
Wqq from top	0	0	0	0.07	0	0	0	0	0.08	0
other from top	0	0	0	0.11	0	0	0	0	0.15	0
other from V	<0.01	<0.01	0.01	<0.01	0	<0.01	0.02	0.03	<0.01	0
no truth	0	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
QCD	0.49	0.42	0.43	0.2	1	0.58	0.64	0.63	0.23	1

Chapter 5 Calibration Strategy

This section introduces methods to calibrate GN2X tagger signal efficiency using $Z(\rightarrow b\bar{b})$+jets events at high Z-boson transverse momenta, discusses relevant systematic uncertainties and presents the results of the calibration. This calibration method has been developed based on Ref. [22], but the analysis in this thesis is based on a new Xbb tagger and uses a new version of ATLAS software release (version 22), which will be used for Run 3 and reprocessed Run 2 data.

5.1 Methodology

The efficiency of the GN2X tagger to select $Z \rightarrow b\bar{b}$ signal events in data (ϵ^{data}) is defined as the number of data events [1] which pass the GN2X tagger requirement ($N^{\text{data}}_{\text{passed}}$) divided by the total number of signal events in data ($N^{\text{data}}_{\text{total}}$). Due to imperfect modeling of the MC simulation process, this can be different from the efficiency in simulation (ϵ^{MC}):

$$\epsilon^{\text{data}} = N^{\text{data}}_{\text{passed}}/N^{\text{data}}_{\text{total}},\ \epsilon^{\text{MC}} = N^{\text{MC}}_{\text{passed}}/N^{\text{MC}}_{\text{total}}. \tag{5.1}$$

Thus a data-to-simulation scale factor (SF) needs to be measured ($\text{SF} = \epsilon^{\text{data}}/\epsilon^{\text{MC}}$) and applied to the MC simulation to correct the difference between data and simulation. The equation can be rewritten as:

$$\text{SF} = \frac{\epsilon^{\text{data}}}{\epsilon^{\text{MC}}} = \frac{N^{\text{data}}_{\text{passed}}/N^{\text{data}}_{\text{total}}}{N^{\text{MC}}_{\text{passed}}/N^{\text{MC}}_{\text{total}}} = \frac{N^{\text{data}}_{\text{passed}}/N^{\text{MC}}_{\text{passed}}}{N^{\text{data}}_{\text{total}}/N^{\text{MC}}_{\text{total}}} = \frac{\mu_{\text{post-tag}}}{\mu_{\text{pre-tag}}}. \tag{5.2}$$

The values $\mu_{\text{post-tag}}$ and $\mu_{\text{pre-tag}}$ are the number of signal events in data divided by the number of signal events in MC simulation before and after the GN2X tagger requirement, respectively.

Huge backgrounds make the $\mu_{\text{pre-tag}}$ measurement using $Z(\rightarrow b\bar{b})$+jets events in data before the GN2X tagger impossible. Thus, the $\mu_{\text{pre-tag}}$ is measured using $Z(\rightarrow \ell^+\ell^-)$ + jets , $\ell = e, \mu$ process, which exhibits a smaller relative background contribution. The

1 Strictly speaking, the GN2X tagger is applied to each boosted object in an event. So, the number of data events means ”the number of objects”. However, in this study, there is exactly one object (”the leading large-R jet”) per event. So, the term ”event” is used.

number of signal events in data before tagging can be estimated by:

$$N^{\text{data}}_{Z\to b\bar{b}} = N^{\text{MC}}_{Z\to b\bar{b}} \cdot \frac{N^{\text{data}}_{Z\to \ell\ell}}{N^{\text{MC}}_{Z\to \ell\ell}} = N^{\text{MC}}_{Z\to b\bar{b}} \cdot \frac{N^{\text{data}}_{\ell\ell} - N^{\text{MC}}_{\text{bkg},\ell\ell}}{N^{\text{MC}}_{Z\to \ell\ell}}, \tag{5.3}$$

where $N^{\text{MC}}_{Z\to b\bar{b}}$ is the number of $Z \to b\bar{b}$ signal events in MC simulation before tagging, $N^{\text{data}}_{\ell\ell}$ is the number of selected events in data, $N^{\text{MC}}_{\text{bkg},\ell\ell}$ is the number of background events in MC simulation and $N^{\text{MC}}_{Z\to \ell^+\ell^-}$ is the number of $Z \to \ell^+\ell^-$ signal events in MC simulation. Thus, the $\mu_{\text{pre-tag}}$ can be calculated as:

$$\mu_{\text{pre-tag}} = \frac{N^{\text{data}}_{\ell\ell} - N^{\text{MC}}_{\text{bkg},\ell\ell}}{N^{\text{MC}}_{Z\to \ell\ell}}, \tag{5.4}$$

and depends only on the yields of $Z \to \ell^+\ell^-$ channel. The event selections for $Z \to b\bar{b}$ and $Z \to \ell^+\ell^-$ events are described in Section 5.2 and 5.3, respectively.

The SFs are measured as a function of the Z-boson candidate p_T. As the large-R jet $p_T > 450$ GeV, in order to have a uniform division of the p_T range with respect to available statistics, three leading large-R jet p_T bins are defined based on the number of events observed in the data: $450 < p_T < 500$ GeV, $500 < p_T < 600$ GeV and $600 < p_T < 1000$ GeV.

5.2 Event Selection for the $\mu_{\text{post-tag}}$ Measurement

5.2.1 Trigger Strategy

Unprescaled single large-R jet triggers are employed for the trigger strategy. These triggers are adapted throughout the 4 data-taking periods to accommodate machine conditions, involving to increase in the p_T threshold and to introduce a jet mass requirement. This additional requirement on the jet mass allows for a reduction in the cut for the reconstructed large-R jet p_T. The relaxation of the p_T cut significantly influences the analysis sensitivity due to the sharply declining p_T spectrum.

A summary of the triggers utilized, including 99% efficiency points determined through fitting the Fermi function, is shown in Table 5.1.

5.2.2 Pre-selection

Pre-selection requirements are applied as follows:

- Passing the detector quality criteria (for data only)
  - Passing calorimeter operation goodness criteria

Table 5.1 Overview of the triggers used for the $\mu_{\text{post-tag}}$ Measurement. They are applied as an OR and all are required to be active. The offline threshold corresponds to the offline jet cut above which the triggers are 99% efficient [70].

Year	Trigger	Offline Threshold [GeV]	Luminosity [fb^{-1}]
2015	HLT_j360_a10_lcw_sub_L1J100	$p_{T,J} > 410$ GeV	3.2
2016	HLT_j420_a10_lcw_L1J100	$p_{T,J} > 450$ GeV	33.0
2017	HLT_j440_a10t_lcw_jes_L1J100	$p_{T,J} > 450$ GeV	41.2
	HLT_j390_a10t_lcw_jes_30smcINF_L1J100	$p_{T,J} > 420$ GeV, $m_J > 50$ GeV	41.0
2018	HLT_j460_a10t_lcw_jes_L1J100	$p_{T,J} > 490$ GeV	58.5
	HLT_j420_a10t_lcw_jes_30smcINF_L1J100	$p_{T,J} > 450$ GeV, $m_J > 60$ GeV	58.5
	HLT_j420_a10t_lcw_jes_30smcINF_L1J100_a10_sub_L1SC111	$p_{T,J} > 450$ GeV, $m_J > 60$ GeV	55.4

  - Passing calorimeter noise cleaning criteria
- Passing the triggers as shown in Table 5.1
- Presence of a primary vertex
- At least two large-R calorimeter jets with $p_T > 200$ GeV to ensure the dijet topology
- Applied overlap removal on the selected events as shown in Section 4.6
- No isolated electron or muon with $p_T > 25$ GeV

5.2.3 p_T Symmetry and Rapidity Cut

In addition to pre-selections, the following event selections are applied to increase the signal significance and get rid of mismodeling phase space. Two additional cuts are applied:

- p_T symmetry cut: $\frac{p_{T,1}-p_{T,2}}{p_{T,1}+p_{T,2}} < 0.15$
- Rapidity difference cut: $\left|\Delta y_{1,2}\right| < 1.2$

where $p_{T,1}$ is the transverse momentum of the leading large-R jet (the jet with the highest p_T) and $p_{T,2}$ is the transverse momentum of the sub-leading large-R jet (the jet with 2nd highest p_T). $\Delta y_{1,2}$ is the rapidity difference between the leading and sub-leading large-R jets.

These two selections help to get rid of the mismodeling of multijet. The distribution of these two is shown in Figure 5.1 and 5.2. After the pre-selection as shown in Section 5.2.2, most of the remaining events come from dijets (multijet process). So, data is compared with the dijet MC samples. For the p_T symmetry cut, the ratio starts to decrease from 0.15. For the rapidity difference, it is tightened to 1.2 to reject more multijet events.

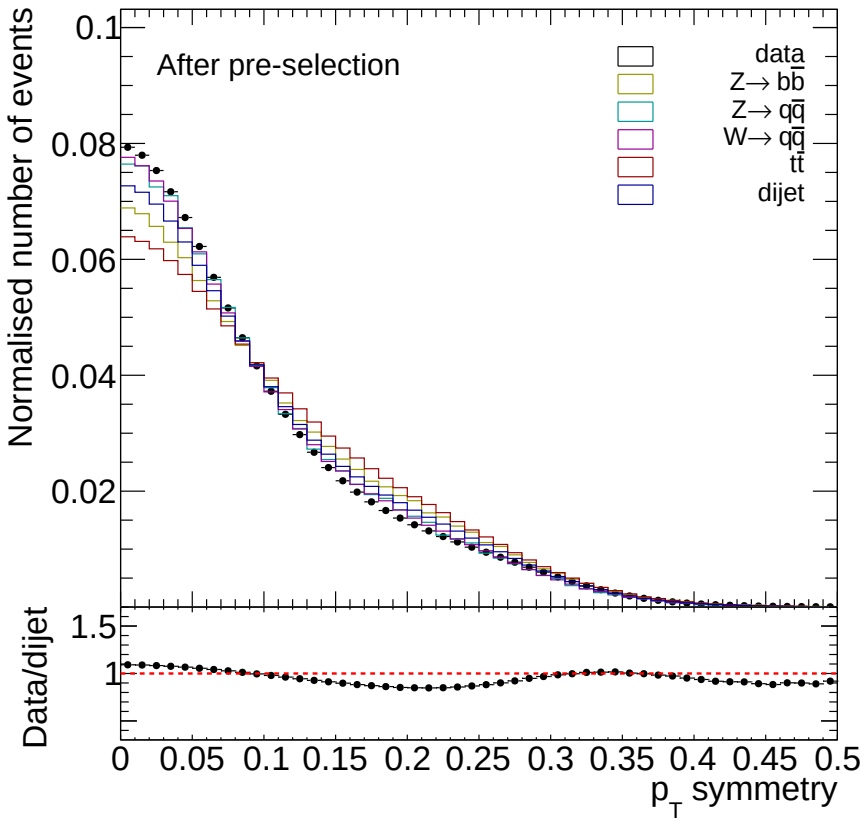


Figure 5.1 The p_T symmetry distribution of the $Z \to b\bar{b}, Z \to q\bar{q}$ $(q = u, d, s, c), W \to q\bar{q}$, dijets and $t\bar{t}$ samples.

5.2.4 Large-R Jet Candidates

The candidates for the calibration are selected from leading large-R jets with mass range $50 \leq m < 150$ GeV and transverse momentum larger than 450 GeV which can satisfy the trigger condition for all the data-taking periods. The candidates must have at least two ghost-associated VR track jets with $p_T > 7$ GeV.

The fraction of the leading large-R jets originate from different MC templates is as shown in Figure 5.3. In the $Z(\to b\bar{b})$ + jets process, the $Z \to b\bar{b}$ could be either the leading large-R jet or the sub-leading large-R jet. Since only the leading large-R jet is considered as the $Z \to b\bar{b}$ candidate, about 40% of the signal events which are the sub-leading large-R jets are lost. A similar situation happens in other processes, statistics are lost by only using leading large-R jets. However, it's not trivial to add the contribution of subleading large-R jets because of the large fraction of the wrong-labeled large-R jets. This will be one of the improvements in the future.

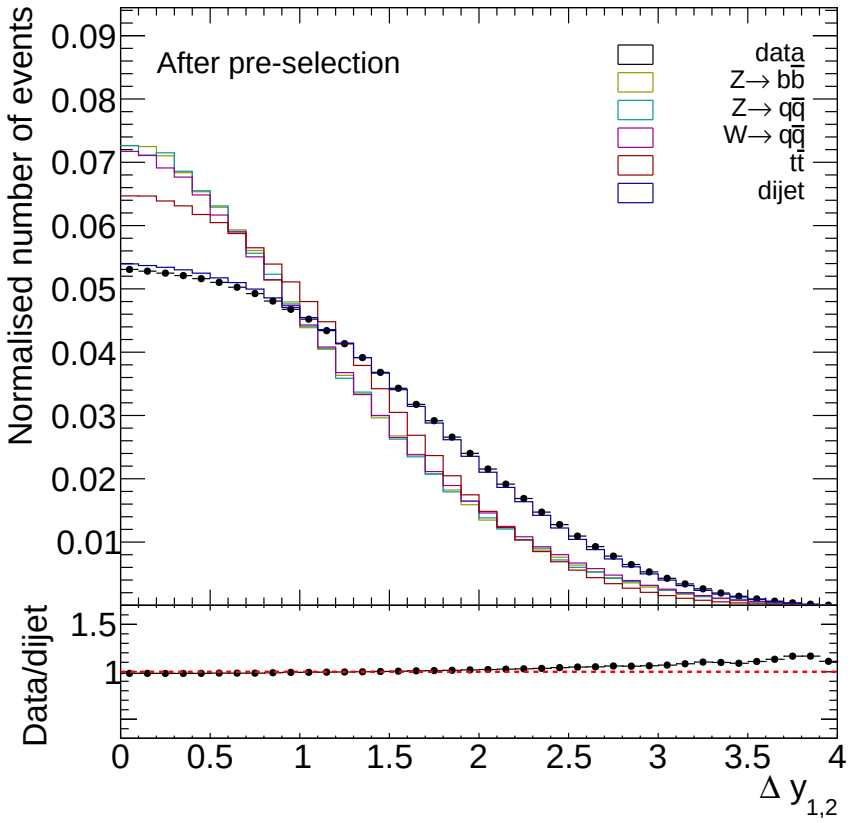


Figure 5.2 The rapidity difference distribution of the $Z \to b\bar{b}$, $Z \to q\bar{q}$ $(q = u, d, s, c)$, $W \to q\bar{q}$, dijets and $t\bar{t}$ samples.

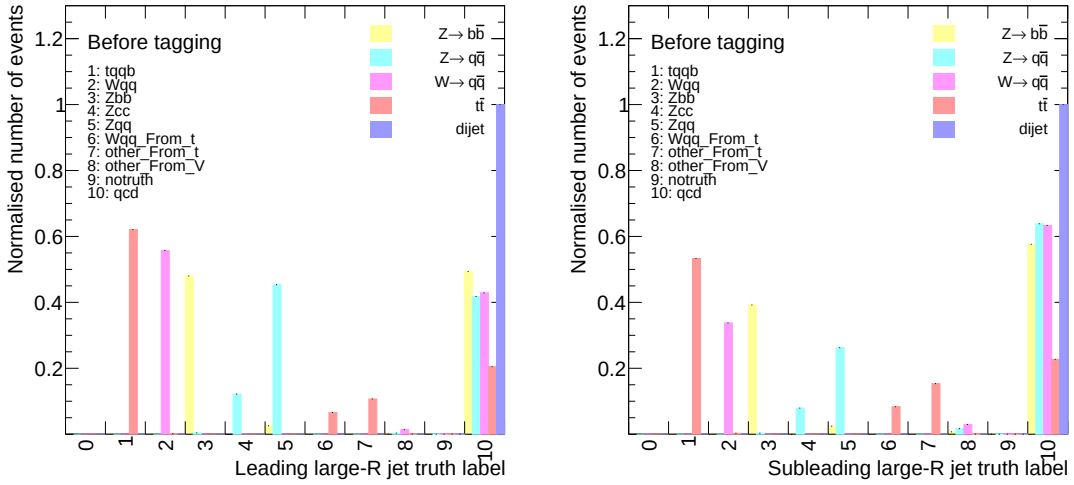


Figure 5.3 Truth labels of the leading large-R jet and the sub-leading large-R jet in $Z \to b\bar{b}$, $Z \to q\bar{q}$ $(q = u, d, s, c)$, $W \to q\bar{q}$, dijets and $t\bar{t}$ samples.

5.2.5 $X \to b\bar{b}$ Tagger GN2X

The $X \to b\bar{b}$ double b-tagger aims to identify the large-R jets that contain two b-jets from the decay of massive particles. GN2X, a new algorithm based on Graph Neural Networks (GNNs) and transformers, is trained to classify large-R jets based on their origin, discriminating jets from boosted Higgs boson decaying into pairs of bottom quarks, $H(b\bar{b})$-jets, and charm quarks, $H(c\bar{c})$-jets, from those originating from background processes.

GN2X benefits from the study of novel advances in flavor tagging of small-radius jets using GNNs and transformers. A more detailed description of the GN2X algorithm can be found in Appendix B. Also, we have implemented an alternative Xbb tagger based on GN2X, incorporating a method called subgraphs, as detailed in Appendix B.2, with the aim of enhancing the performance of the base GN2X.

GN2X provides four outputs for each large-R jet, corresponding to the probabilities of the jet being a $H(b\bar{b})$-jet (p_{Hbb}), a $H(c\bar{c})$-jet (p_{Hcc}), multijet (p_{QCD}), or a top-quark (p_{top}) jet. A discriminant, roughly corresponding to a log-likelihood ratio, is constructed from these four outputs:

$$D_{\text{Hbb}} = \ln \frac{p_{\text{Hbb}}}{f_{\text{Hcc}} \cdot p_{\text{Hcc}} + f_{\text{top}} \cdot p_{\text{top}} + (1 - f_{\text{Hcc}} - f_{\text{top}}) \cdot p_{\text{QCD}}}, \tag{5.5}$$

where f_{Hcc} and f_{top} are two free parameters that determine the relative weights of p_{Hcc} and p_{top} respectively to p_{QCD}, controlling the trade-off among $H(c\bar{c})$, top and multijet rejections. For the following studies, the values of f_{Hcc} and f_{top} are set to 0.02 and 0.25 [21], respectively, which were obtained after optimization procedure to maximize the rejection for a given signal efficiency. The tagging of a large-R jet as originating from an $H \to b\bar{b}$ decay is performed by applying a cut on the discriminant D_{Hbb}.

Figure 5.4 shows the normalized distribution of the discriminant score for $H(b\bar{b})$, $H(c\bar{c})$, top and multijet jets.

The performance of GN2X is evaluated in terms of the signal efficiency and the rejection of the background processes. The signal efficiency (ϵ) is defined as the tagging efficiency of $H(b\bar{b})$-jets. The background rejection factors are defined as the inverse of a background mis-tagging efficiency ($1/\epsilon$) for $t\bar{t}$ and multijet events.

There are different working points (WPs) for GN2X, which are shown in Table 5.2.

In summary, events used to extract the $\mu_{\text{post-tag}}$ are selected as follows:

- Pre-selection as described in Section 5.2.2

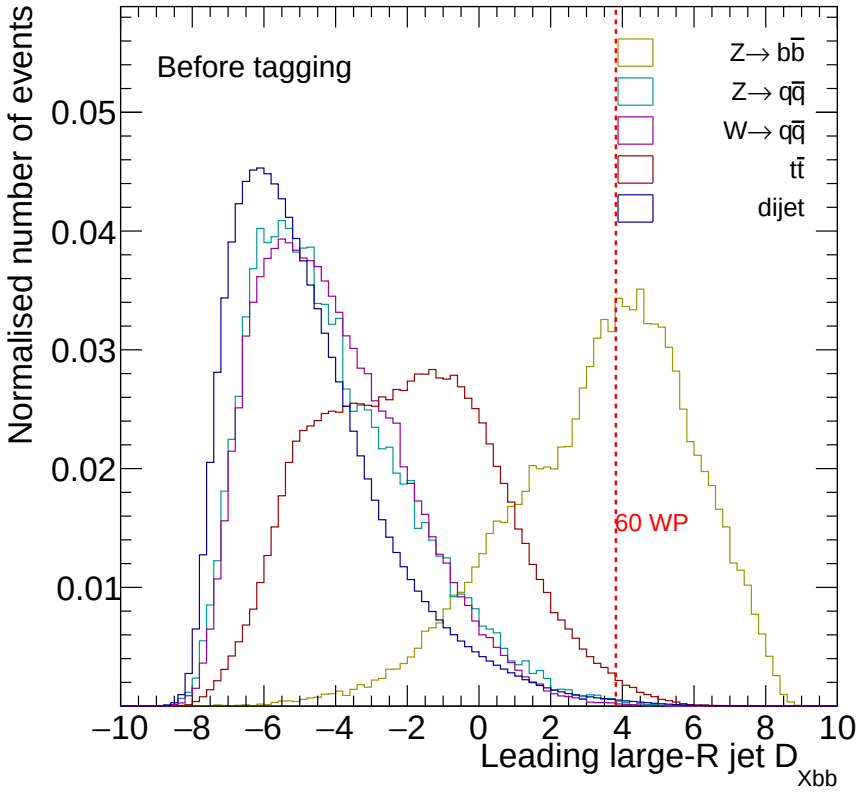


Figure 5.4 The discriminant score of $H(b\bar{b})$ jets in $Z \to b\bar{b}$, $Z \to q\bar{q}$ ($q = u, d, s, c$), $W \to q\bar{q}$, dijets and $t\bar{t}$ templates. The red line shows the cut value for 60% WP.

Table 5.2 GN2X threshold values for all WPs with $f_{\mathrm{Hcc}} = 0.02$ and $f_{\mathrm{top}} = 0.25$

Working point	50%	60%	70%	80%
Threshold value	4.335	3.818	3.166	2.211

- The leading large-R jet is required to have $p_T > 450$ GeV
- At least two ghost-associated VR track jets with $p_T > 7$ GeV
- The jet mass of leading-p_T large-R jet is required to be above 50 GeV
- Additional requirements are applied to the large-R jets to suppress the mis-modeling of multijets and to reject backgrounds as shown in Section 5.2.3
- The leading large-R jet is required to pass the $X \to b\bar{b}$ tagger

A schematic diagram of the $Z \to b\bar{b}$ event selection is also shown in Figure 5.5. After each selection and the GN2X tagging mentioned above, the number of events for $Z \to b\bar{b}$, $Z \to q\bar{q}$ ($q = u, d, s, c$), $W \to q\bar{q}$, dijets and $t\bar{t}$ templates with mass range $50 \leq m < 150$ GeV is shown in Table 5.3.

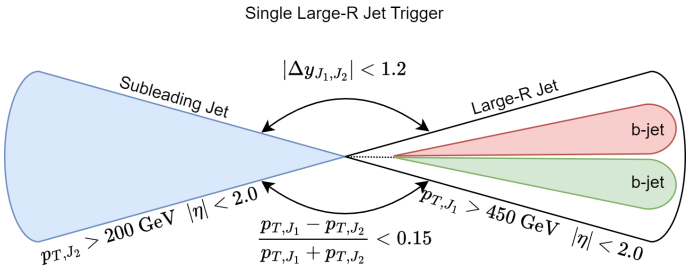


Figure 5.5 Schematic diagram of the $Z \to b\bar{b}$ event selection.

Table 5.3 The number of events for $Z \to b\bar{b}$, $Z \to q\bar{q}$ ($q = u, d, s, c$), $W \to q\bar{q}$, dijets and $t\bar{t}$ templates with mass range $50 \leq m < 150$ GeV from Run 2 MC simulation (normalized to 140 fb^{-1}) after each selection.

Selection	Number of events				
	$Z \to b\bar{b}$	$Z \to q\bar{q}$	$W \to q\bar{q}$	$t\bar{t}$	dijets
Pre-selections	52445	258324	803963	422781	254586038
p_T symmetry < 0.15	41918	205807	642392	322625	188926756
$\lvert\Delta y_{1,2}\rvert < 1.2$	31790	155751	480084	230729	119146981
Large-R jet candidates	7534	37534	115143	40312	26357350
GN2X tagger 80% WP	5957	345.8	514.3	1661	262739
GN2X tagger 70% WP	5194	141.2	76.46	715.7	145703
GN2X tagger 60% WP	4416	53.52	71.16	356.0	89019
GN2X tagger 50% WP	3712	16.40	35.37	195.1	56380

5.3 Event Selection for the $\mu_{\text{pre-tag}}$ Measurement

As discussed in Section 5.1, we use $Z(\rightarrow \ell^+\ell^-)+\text{jets}$ instead of $Z(\rightarrow b\bar{b})+\text{jets}$ to calculate $\mu_{\text{pre-tag}}$. At least one additional high p_T jet is required in order to ensure a similar event topology as the $Z(\rightarrow b\bar{b})+\text{jets}$ case. For the other selections, the p_T symmetry and rapidity difference cuts are similar to the simuation of $Z(\rightarrow b\bar{b})+\text{jets}$. The selection is summarized as follows:

- Passing the detector quality criteria (for data only)
  - Passing calorimeter operation goodness criteria
  - Passing calorimeter noise cleaning criteria
- Passing the triggers as shown in Table 5.4 and 5.5
- Presence of a primary vertex
- Applied overlap removal on the selected events as shown in Section 4.6
- At least two leptons of the same flavor
- At least one large-R jet with $p_T > 200$ GeV and $|\eta| < 2.0$
- For $Z \rightarrow \mu\mu$, at least one of the selected muons has $p_T > 27$ GeV and opposite charge
- For $Z \rightarrow ee$, at least one of the selected electrons has $p_T > 25$ GeV
- The lepton p_T balance is required: $\left(p_T^{\ell_1} - p_T^{\ell_2}\right)/p_T^{\ell\ell} < 0.8$
- In analogy to the $\mu_{\text{post-tag}}$ measurement, additional requirements are applied as follows:
  - $p_T^{\ell\ell}$ is required to be above 450 GeV and larger than the leading large-R jet p_T: $p_T^{\ell\ell} > p_T^{\text{lead.jet}}$
  - p_T symmetry requirement: $\frac{p_T^{\ell\ell} - p_T^{\text{lead.jet}}}{p_T^{\ell\ell} + p_T^{\text{lead.jet}}} < 0.15$
  - rapidity difference: $|\Delta y_{\ell\ell-\text{lead.jet}}| < 1.2$
- The signal region is defined by a mass window requirement of $66 < m_{Z\rightarrow\ell^+\ell^-} < 116$ GeV

A schematic diagram of the $Z \rightarrow b\bar{b}$ event selection is also shown in Figure 5.6.

Table 5.4 Overview of single muon triggers used for the $\mu_{\text{pre-tag}}$ Measurement. They are applied as an OR and all are required to be active.

Year	Trigger	Offline Threshold [GeV]
2015	HLT_mu20_iloose_L1MU15	$p_{T,\ell} > 20$ GeV
2015-2018	HLT_mu50	$p_{T,\ell} > 50$ GeV
2016-2018	HLT_mu26_ivarmedium	$p_{T,\ell} > 26$ GeV

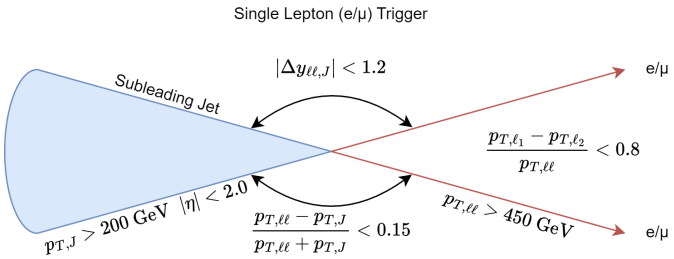


Figure 5.6 Schematic diagram of the $Z \to \ell\ell$ event selection.

Table 5.5 Overview of single electron triggers used for the $\mu_{\text{pre-tag}}$ Measurement. They are applied as an OR and all are required to be active.

Year	Trigger	Offline Threshold [GeV]
2015	HLT_e24_lhmedium_L1EM20VH	$p_{T,\ell} > 24$ GeV
	HLT_e60_lhmedium	$p_{T,\ell} > 60$ GeV
	HLT_e120_lhloose	$p_{T,\ell} > 120$ GeV
2016-2018	HLT_e26_lhtight_nod0_ivarloose	$p_{T,\ell} > 26$ GeV
	HLT_e60_lhmedium_nod0	$p_{T,\ell} > 40$ GeV
	HLT_e120_lhloose_nod0	$p_{T,\ell} > 140$ GeV

Chapter 6 Signal and Background Modeling

6.1 Modeling for the $\mu_{\text{post-tag}}$ Measurement

6.1.1 Comparison of Data and Simulation

The yields before and after the tagger are shown in Table 6.1 and 6.2. For $450 \leq p_T < 500$GeV, $500 \leq p_T < 600$GeV and $600 \leq p_T < 1000$GeV p_T bins, the comparison plots for the large-R jet kinematics before the $X \to b\bar{b}$ tagging are shown in Figure 6.1, and after the $X \to b\bar{b}$ tagging, for 60% working point, are shown in Figure 6.2. Similarly, the comparison plots for other working points are shown in Appendix A.1. Before the $X \to b\bar{b}$ tagging, for all the p_T bins, the data/MC ratio is seen to be smaller than 1 because the cross-section of dijets is overestimated. After the $X \to b\bar{b}$ tagging, the data/MC ratio is slightly larger than 1 and almost flat with a small positive slope.

Table 6.1 Event yields for different Run 2 MC samples (normalized to 140 fb^{-1}) before $X \to b\bar{b}$ tagger in different p_T bins in large-R jet mass range $50 \leq m < 150$ GeV. Statistical uncertainties of MC samples are shown.

MC Sample	$450 \leq p_T < 500$ GeV	$500 \leq p_T < 600$ GeV	$600 \leq p_T < 1000$ GeV
$Z \to b\bar{b}$	9770 ± 51	4660 ± 33	2860 ± 23
$Z \to q\bar{q}$	42260 ± 370	20030 ± 240	12430 ± 160
$W \to q\bar{q}$	129100 ± 480	61700 ± 320	38600 ± 220
$t\bar{t}$	45800 ± 140	21100 ± 100	11900 ± 75
dijet	20675500 ± 4700	9421600 ± 2800	5046300 ± 1100

Table 6.2 Event yields for different Run 2 MC samples (normalized to 140 fb^{-1}) after 60% WP of $X \to b\bar{b}$ tagger in different p_T bins in large-R jet mass range $50 \leq m < 150$ GeV. Statistical uncertainties of MC samples are shown.

MC Sample	$450 \leq p_T < 500$ GeV	$500 \leq p_T < 600$ GeV	$600 \leq p_T < 1000$ GeV
$Z \to b\bar{b}$	1880 ± 21	1620 ± 17	900 ± 11
$Z \to q\bar{q}$	30.8 ± 9.2	55.4 ± 8.6	36 ± 13
$W \to q\bar{q}$	74 ± 12	74.5 ± 9.7	75 ± 11
$t\bar{t}$	177.9 ± 9.2	160.2 ± 9.0	88.5 ± 6.8
dijet	40000 ± 210	31800 ± 140	16300 ± 59

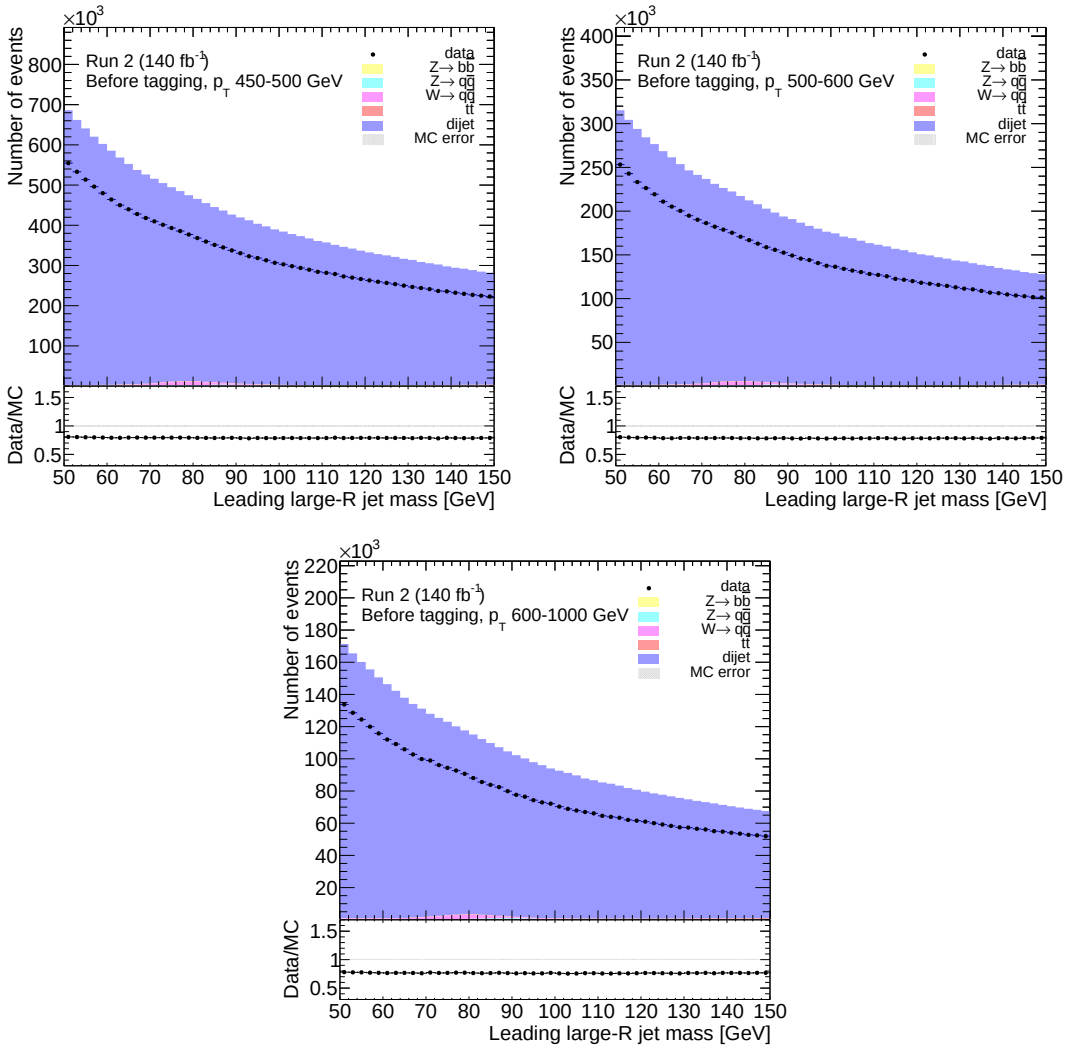


Figure 6.1 The comparison of data and MC samples prediction before $X \to b\bar{b}$ tagger for different p_T bins. Different MC samples are stacked together. The MC error is shown as the shaded band, but it's too small to be seen.

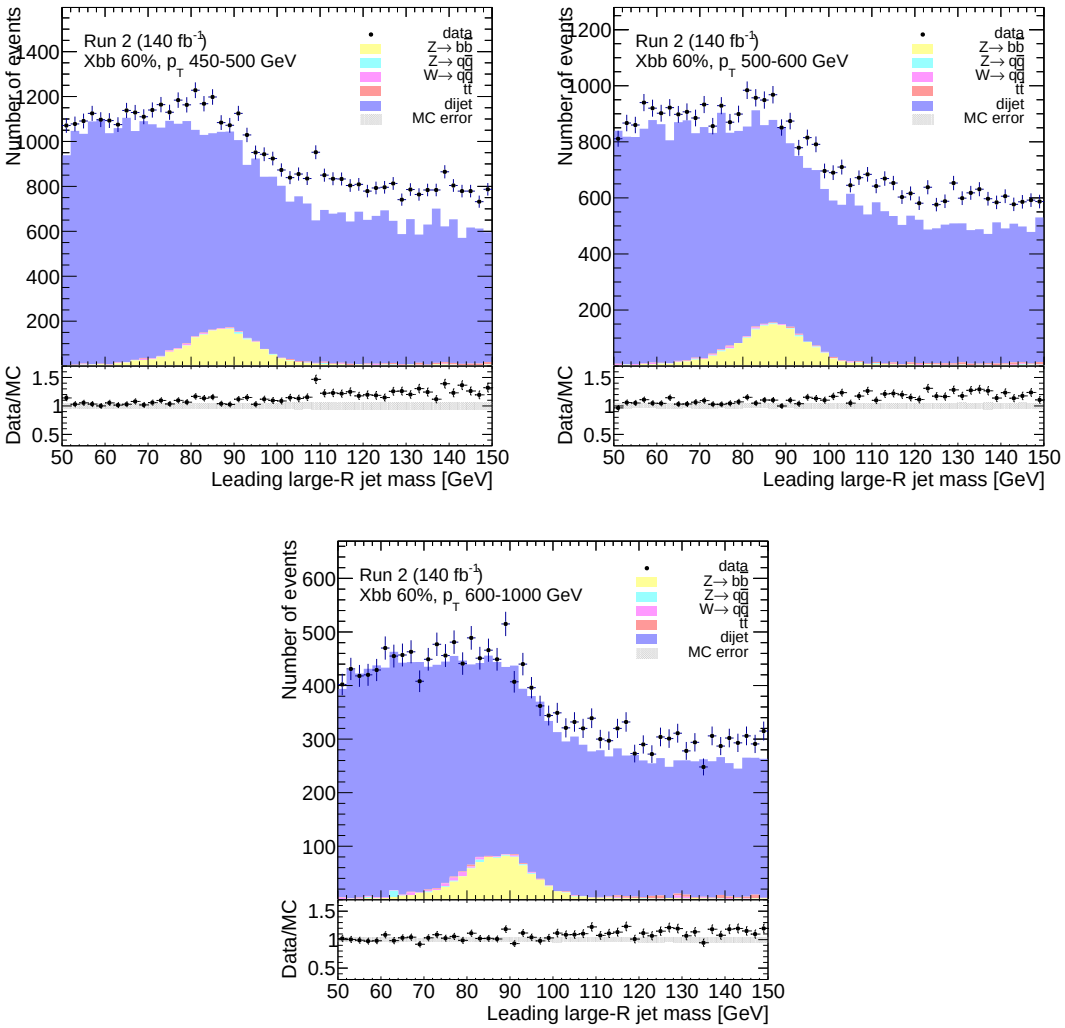


Figure 6.2 The comparison of data and MC samples prediction after 60% WP of $X \to b\bar{b}$ tagger for different p_T bins. Different MC samples are stacked together. The MC error is shown as the shaded band.

6.1.2 Signal and Background Modeling

The multijet background is determined using data-driven techniques. Other backgrounds and the signal events are modeled using MC simulations. The signal and background modeling are given in the following sections. Techniques differ for the $\mu_{\text{pre-tag}}$ and $\mu_{\text{post-tag}}$ measurements. For the $\mu_{\text{post-tag}}$ measurement a fit to the $Z \to b\bar{b}$ mass distribution is performed, while for the $\mu_{\text{pre-tag}}$ measurement only signal and background yields need to be extracted after selections described in Section 5.3. The mass distribution of $Z \to b\bar{b}$, $Z \to q\bar{q}$, $W \to q\bar{q}$, $t\bar{t}$ templates after tagging are shown in Figure 6.3.

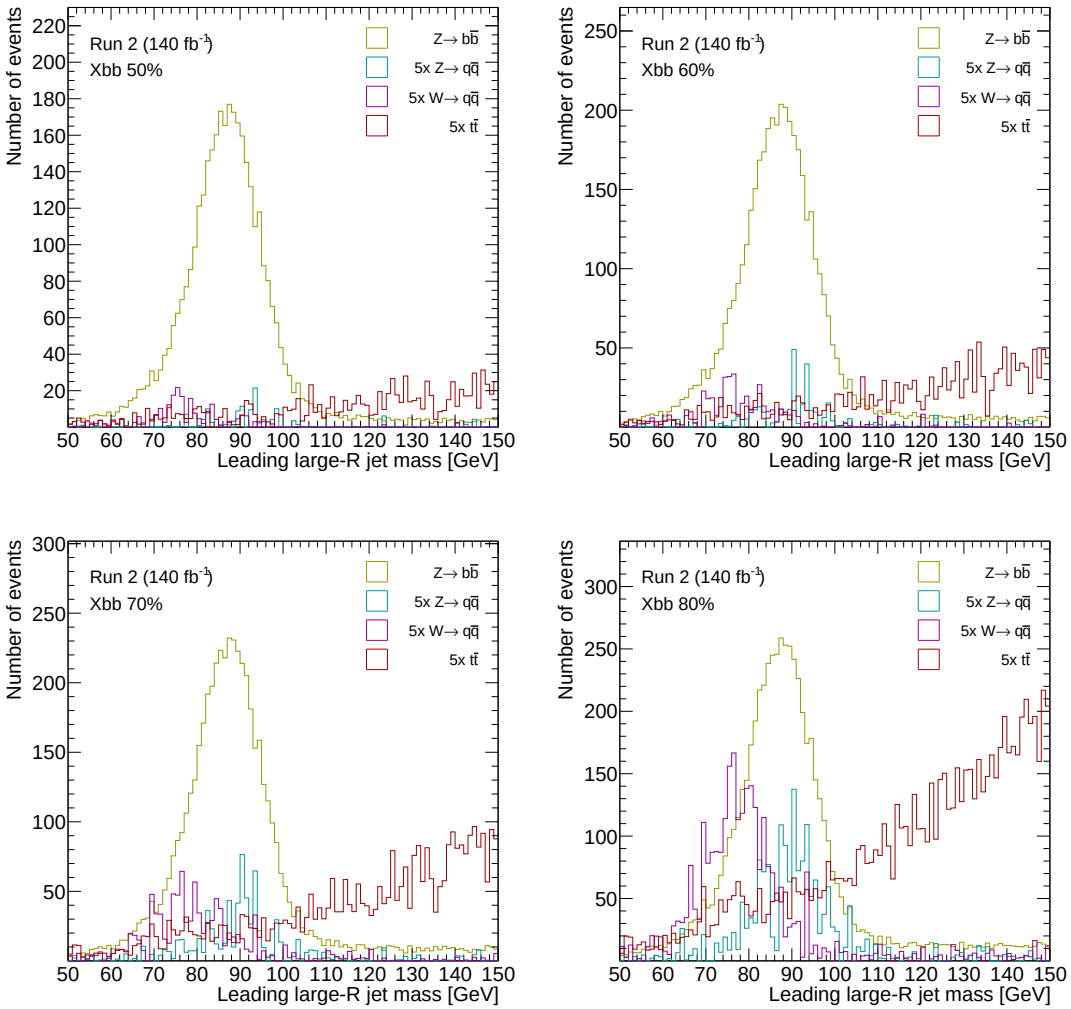


Figure 6.3 The mass distribution of $Z \to b\bar{b}$, $Z \to q\bar{q}$, $W \to q\bar{q}$, $t\bar{t}$ templates after 50%, 60%, 70% and 80% working point of GN2X tagger.

6.1.2.1 Signal Modeling

To model the $Z \to b\bar{b}$ signal invariant mass distribution for the $\mu_{\text{post-tag}}$ measurement, a double sided crystal ball (DSCB) function [71] is used:

$$f(m|m_Z, \sigma_Z, \alpha_L, \alpha_H, n_L, n_H) = N \cdot \begin{cases} \exp\left(-\frac{\alpha_L^2}{2}\right)\left[\frac{\alpha_L}{n_L}\left(\frac{n_L}{\alpha_L} - \alpha_L - t\right)\right]^{-n_L}, & t < -\alpha_L \\ \exp\left(-\frac{1}{2}t^2\right), & -\alpha_L \leq t \leq \alpha_H \\ \exp\left(-\frac{\alpha_H^2}{2}\right)\left[\frac{\alpha_H}{n_H}\left(\frac{n_H}{\alpha_H} - \alpha_H + t\right)\right]^{-n_H}, & t > \alpha_H \end{cases} \quad (6.1)$$

where $t = (m - m_Z)/\sigma_Z$, the first and the third case of the equation describes the tail and the second case describes the core of the distribution. N is a normalization parameter, m_Z and σ_Z denote the mean and standard deviation of the Gaussian core. α_L, n_L and α_H, n_H are the decay constants and normalization on the low and the high side tails.

The parameters of the DSCB function are derived fitting the $Z \to b\bar{b}$ MC templates after applying the requirements listed in Section 5.2.

Figure 6.4 and Table 6.3 shows the χ^2 fits to the signal MC samples for the $X \to b\bar{b}$ tagging requirement at 60% working point for $450 \leq p_T < 1000$ GeV. Other results for different working points are shown in Appendix A.2.

Table 6.3 The parameters of the DSCB function from a fit to the $Z \to b\bar{b}$ template for the $Z \to b\bar{b}$ MC templates after the $X \to b\bar{b}$ tagging requirement at 60% working point for $450 \leq p_T < 1000$ GeV.

N	199.1 ± 1.95
m_Z [GeV]	87.2 ± 0.08
σ_Z [GeV]	7.71 ± 0.09
α_L	1.08 ± 0.05
n_L	16.0 ± 7.95
α_H	2.11 ± 0.04
n_H	0.46 ± 0.06
χ^2	135.8
χ^2/NDF	1.46

6.1.2.2 Background Modeling

The dominant source of background in the $Z \to b\bar{b}$ $\mu_{\text{post-tag}}$ measurement is multijet events. The contribution from $W(\to q\bar{q})$+jets and $t\bar{t}$ process is found to be negligible. To confirm that ignoring them will not bias the yield estimated, a test to model the signal by comparing between using all MC samples and using only $Z \to b\bar{b}$ and dijet MC samples

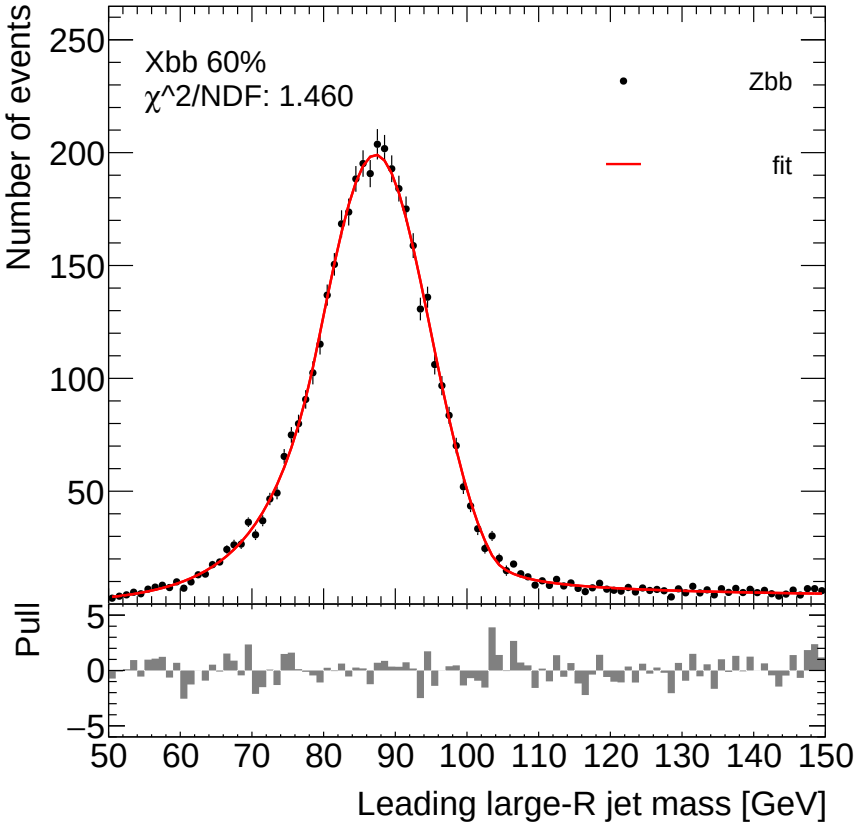


Figure 6.4 The χ^2 fits to the Z candidate mass distribution ($Z \to b\bar{b}$ template) via a DSCB function, passing the $X \to b\bar{b}$ 60% WP for $450 \leq p_T < 1000$ GeV.

is performed. The results are shown in Figure 6.5 and 6.4. After the test, the number of signal and background events by fitting is nearly the same. Thus, in this study, only dijet background is considered.

The modeling of multijet events is performed using classes of exponential and polynomial functions of 2nd to 5th order in bins of the transverse momentum of the $Z \to b\bar{b}$ candidate. The choice of the functional form is made using binned maximum likelihood fit to the sidebands of the invariant $Z \to b\bar{b}$ mass distribution in data, defined as $50 < m_{Z\to b\bar{b}} < 70$ GeV and $110 < m_{Z\to b\bar{b}} < 150$ GeV. To decide on the suitable number of free parameters (the order) of the two classes of functions: a F-test is performed [72]. The test statistic $F_{p,q}$ is calculated as:

$$F_{p,q} = \frac{\chi_p^2 - \chi_q^2}{n_q - n_p} / \frac{\chi_q^2}{n - n_q}, \tag{6.2}$$

where χ_p^2 and χ_q^2 describe quantitatively the goodness of the fit, computed in n bins of two fits with n_p and n_q degrees of freedom. The optimal functions to describe the multijet background are summarised in Table 6.5 [73]. These functions are used in the final fit

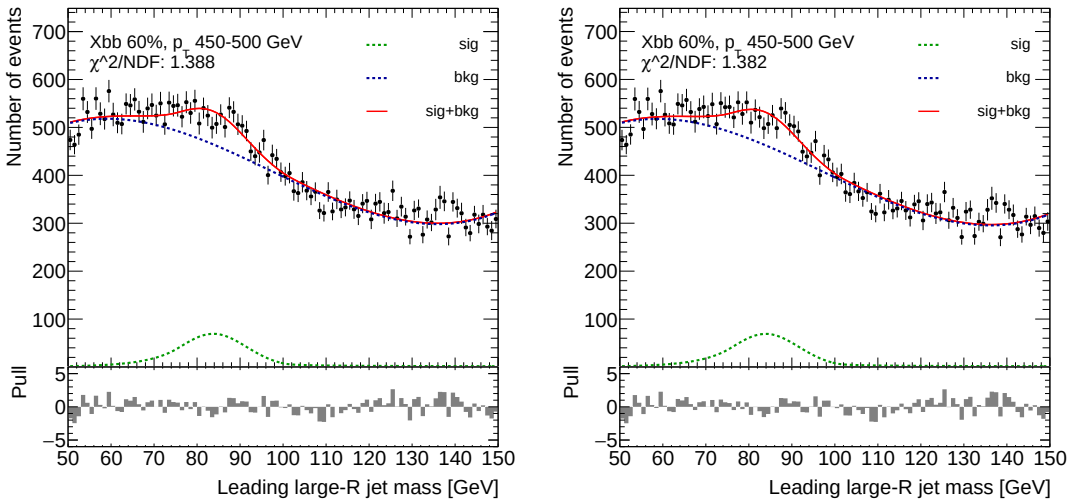


Figure 6.5 The χ^2 fits to the Z candidate mass distribution based on only $Z \to b\bar{b}$ and dijet MC templates (left) and all MC templates (right), passing the $X \to b\bar{b}$ 60% WP for $450 \leq p_T < 500$ GeV.

Table 6.4 The parameters of fitting signal and background models for different background treatments (only dijet or dijet+$W + t\bar{t}$) after the $X \to b\bar{b}$ tagging requirement at 60% working point for $450 \leq p_T < 500$ GeV.

Parameter	$Z \to b\bar{b}$ and dijet	$Z \to b\bar{b}$, dijet, $W \to q\bar{q}, t\bar{t}$
m_Z [GeV]	84 ± 1.1	84 ± 1.2
σ_Z [GeV]	7.71 (fixed)	7.71 (fixed)
α_L	1.08 (fixed)	1.08 (fixed)
n_L	16.0 (fixed)	16.0 (fixed)
α_H	2.11 (fixed)	2.11 (fixed)
n_H	0.46 (fixed)	0.46 (fixed)
a_0	-34 ± 58	-29 ± 55
a_1	1110 ± 40	1100 ± 60
a_2	-1344.1 ± 5.3	-1350 ± 65
a_3	459.7 ± 4.6	460 ± 24
χ^2	129.9	130.5
χ^2/NDF	1.38	1.39
Signal	1550 ± 254	1530 ± 244
Background	40350 ± 325	40170 ± 318

described in Section 7.2.

Table 6.5 Optimal functions to describe the multijet background in the $\mu_{\text{post-tag}}$ measurement.

$Z \to b\bar{b}$ p_T bin	Optimal function
$450 \leq p_T < 500$ GeV	$\sum_{i=0}^{3} a_i \left(\frac{m}{100[\text{GeV}]}\right)^i$
$500 \leq p_T < 500$ GeV	$\sum_{i=0}^{3} a_i \left(\frac{m}{100[\text{GeV}]}\right)^i$
$600 \leq p_T < 1000$ GeV	$a_0 \exp\left(\sum_{i=1}^{3} a_i \left(\frac{m}{100[\text{GeV}]}\right)^i\right)$

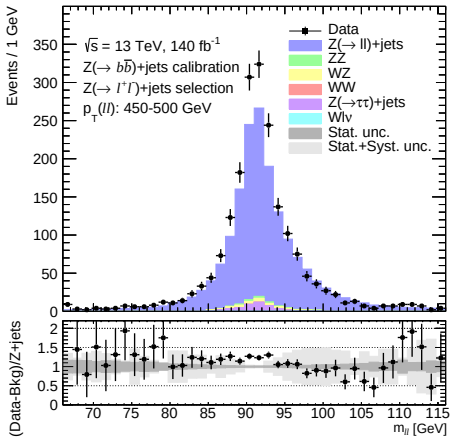
6.2 Modeling for the $\mu_{\text{pre-tag}}$ Measurement

6.2.1 Comparison of Data and Simulation

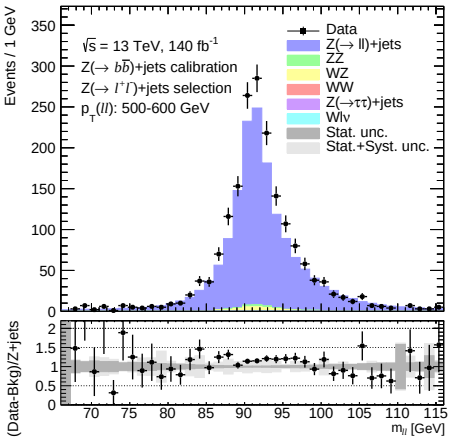
Figure 6.6 shows one of the most representative distributions for the $Z \to \ell^+\ell^-$+jets process, the Z-boson invariant mass distribution in various Z-boson candidate p_T bins. A good shape agreement between data and MC simulation is found.

6.2.2 Signal and Background Modeling

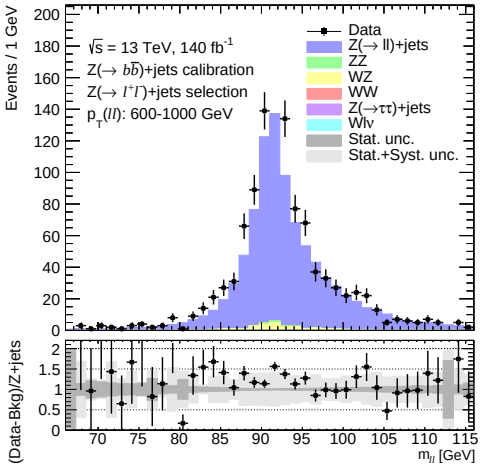
To extract signal events of $Z(\to \ell\ell)$ + jets, MC simulated events are used. the background events are subtracted from the number of observed events. The background events are obtained from the MC simulation of ZZ, WZ, WW, $Z\tau\tau$ and $W\ell\nu$. The main backgrounds are WZ and ZZ diboson processes but they are small. Only total yields are needed to calculate $\mu_{\text{pre-tag}}$.



(a) $450 \leq p_T < 500$ GeV



(b) $500 \leq p_T < 600$ GeV



(c) $600 \leq p_T < 1000$ GeV

Figure 6.6 The comparison of data and MC prediction for $Z \to \ell^+\ell^-$ in three Z-boson p_T bins.

Chapter 7 Results

In Section 7.1, the systematic uncertainties for this calibration are explained. Some values are based on previous studies. Then, the results are shown in Section 7.2.

7.1 Systematic Uncertainties

In the previous study [22], several different systematic uncertainties were studied:

- Z + jets modeling
- Fit model
- Spurious signal
- Other background modeling
- Lepton related (momentum scale and resolution, identification and trigger)
- Jet related (mass scale, mass resolution and energy scale, trigger)

The uncertainties evaluated in this thesis are as follows. But for others, the numbers are reused due to a delay in the MC preparation and so on.

Statistical Uncertainties The statistical uncertainties of the Monte Carlo simulations are considered for all analyses.

Z + jets modeling The uncertainty considered for $Z \to b\bar{b}$ shall be estimated using the alternative generator, but in this thesis, this uncertainty is not considered due to the MC sample preparation problem. The uncertainty considered for $Z \to \ell\ell$ are estimated by three aspects: generator uncertainty (Sherpa 2.2.8 VS MadGraph+Pythia8), scale uncertainties and PDF+α_s uncertainties. The details are as shown in Appendix A.5. However, in this thesis, the values of Z + jets modeling uncertainties are reused from the previous study [22], because those results cover both $Z \to b\bar{b}$ and $Z \to \ell\ell$ modeling uncertainties.

Fit Model for $\mu_{\text{post-tag}}$ An additional systematic uncertainty on the fit model due to the choice of the fit mass range is added. The signal strength $\mu_{\text{post-tag}}$ is evaluated for one alternative mass range (50-140 GeV). The uncertainty is estimated by taking the difference between the nominal and alternative fit results. The fit parameters are shown in Appendix A.3.

Spurious Signal for $\mu_{\text{post-tag}}$ The spurious signal is evaluated by fitting the mass distribution of the background-only MC samples with the signal plus background model. The spurious signal is defined as the fitted signal strength. The spurious signal test results are shown in Appendix A.4.

Other Background Modeling for $\mu_{\text{pre-tag}}$ A 20% normalization uncertainty is assigned to the VV (WW, ZZ, WZ) background. The uncertainty is estimated by recalculating the $\mu_{\text{pre-tag}}$ and extracting the SFs again.

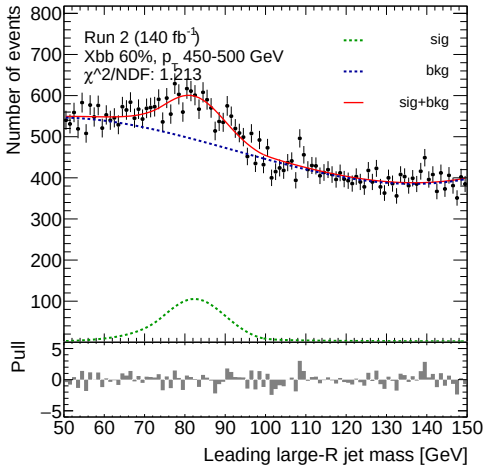
7.2 Results

The $\mu_{\text{post-tag}}$ parameter is measured using the $Z \to b\bar{b}$ invariant mass distribution for the three large-R jet p_T bins mentioned above. To extract the signal strength, a binned maximum likelihood fit is performed. The functional form that describes the smooth background is chosen using the method in Section 6.1.2.2.

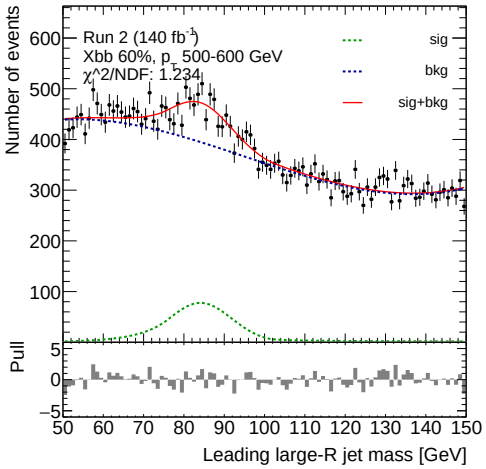
Figure 7.1 shows the $Z \to b\bar{b}$ candidate invariant mass distribution after the fit and applying the GN2X tagger at 60% working point. The pulls, defined as data minus fitted model prediction divided by the data statistical uncertainty, are mostly within 3 standard deviations. The fitting parameters are shown in Table 7.1. We perform the same fitting using MC samples, and the results are shown in Table 7.2. The $\mu_{\text{post-tag}}$ can be calculated as shown in Table 7.3. The data statistical uncertainty is calculated by setting the MC statistical uncertainty to zero, while the MC statistical uncertainty is the opposite.

The $\mu_{\text{pre-tag}}$ parameter is calculated as a ratio of observed yields in data minus the expected background yields divided by the expected $Z \to \ell^+\ell^-$ signal yields as shown in Equation 5.3. All yields are determined after the $Z \to \ell^+\ell^-$ selection requirements for each Z-boson candidate p_T bins. The $\mu_{\text{pre-tag}}$ are shown in Table 7.4 with the observed yields and expected signal and background. Similar to the $\mu_{\text{post-tag}}$, the data statistical uncertainty is calculated by setting the MC statistical uncertainty to zero, while the MC statistical uncertainty is the opposite.

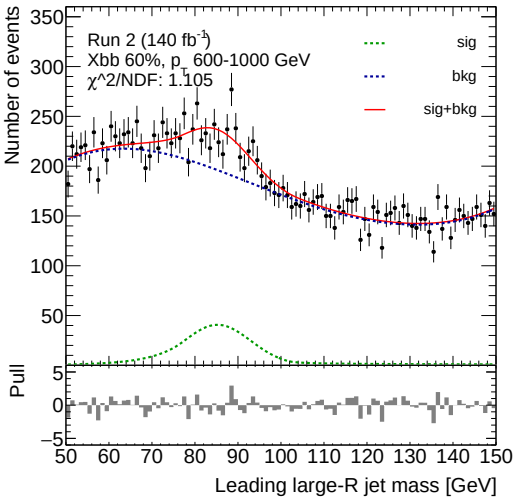
From the $\mu_{\text{post-tag}}$ and $\mu_{\text{pre-tag}}$ shown before, a SF for each p_T region can be obtained as shown in Table 7.5. Some other systematic uncertainties are not estimated in this thesis, the values from the previous study [22] are reused, including Z + jets modeling, spurious signal, lepton related, jet mass scale, jet mass resolution and jet energy scale.



(a) $450 < p_T < 500$ GeV



(b) $500 < p_T < 600$ GeV



(c) $600 < p_T < 1000$ GeV

Figure 7.1 The $Z \to b\bar{b}$ candidate invariant mass distribution and applying the $Z(\to b\bar{b})$ + jets selection and the $X \to b\bar{b}$ 60% WP for events with the large-R jet p_T in the $450 < p_T < 500$ GeV(a); $500 < p_T < 600$ GeV(b); $600 < p_T < 1000$ GeV(c) range. The fit result is shown by a red solid curve. Signal (green) and background (blue) components are shown.

Table 7.1 The parameters of fitting the real data after the $X \to b\bar{b}$ tagging requirement at 60% working point for different p_T bins.

Parameter	Large-R jet p_T		
	450-500 GeV	500-600 GeV	600-1000 GeV
m_Z	82.4 ± 0.8	84.1 ± 0.8	85.3 ± 1.1
σ_Z	7.71 (fixed)	7.71 (fixed)	7.71 (fixed)
α_L	1.08 (fixed)	1.08 (fixed)	1.08 (fixed)
n_L	16.0 (fixed)	16.0 (fixed)	16.0 (fixed)
α_H	2.11 (fixed)	2.11 (fixed)	2.11 (fixed)
n_H	0.46 (fixed)	0.46 (fixed)	0.46 (fixed)
a_0	347 ± 25	175 ± 21	35.9 ± 1.7
a_1	935 ± 59	1134 ± 47	6.8 ± 0.1
a_2	-1310 ± 54	-1470 ± 41	-7.9 ± 0.1
a_3	470 ± 21	510 ± 17	2.70 ± 0.04
χ^2	114.1	116.1	103.9
χ^2/NDF	1.21	1.23	1.11
Signal	2343 ± 252	1729 ± 230	904 ± 135
Background	45440 ± 327	35930 ± 295	17750 ± 186

Table 7.2 The parameters of fitting the MC samples after the $X \to b\bar{b}$ tagging requirement at 60% working point for different p_T bins.

Parameter	Large-R jet p_T		
	450-500 GeV	500-600 GeV	600-1000 GeV
m_Z	83.9 ± 1.1	84.9 ± 0.6	86.9 ± 0.5
σ_Z	7.71 (fixed)	7.71 (fixed)	7.71 (fixed)
α_L	1.08 (fixed)	1.08 (fixed)	1.08 (fixed)
n_L	16.0 (fixed)	16.0 (fixed)	16.0 (fixed)
α_H	2.11 (fixed)	2.11 (fixed)	2.11 (fixed)
n_H	0.46 (fixed)	0.46 (fixed)	0.46 (fixed)
a_0	-34 ± 58	88 ± 73	47 ± 84
a_1	1110 ± 40	990 ± 73	8.1 ± 0.6
a_2	-1344.1 ± 5.3	-1328 ± 81	-9.6 ± 0.6
a_3	459.7 ± 4.6	477 ± 31	3.3 ± 0.2
χ^2	130.1	89.3	94.2
χ^2/NDF	1.40	0.96	1.01
Signal	1550 ± 254	1840 ± 163	844 ± 72
Background	40400 ± 330	31680 ± 210	16460 ± 93

Table 7.3 Post-tag ($\mu_{\text{post-tag}}$) signal strength for the $X \to b\bar{b}$ tagger at 60 % efficiency WP measured using $Z(\to b\bar{b})$ + jets calibration methods.

p_T [GeV]	$450 < p_T < 500$	$500 < p_T < 600$	$600 < p_T < 1000$
$N^{\text{data}}_{Zbb,\text{GN2X}_{60\%}}$	2343 ± 252	1729 ± 230	904 ± 135
$N^{\text{MC}}_{Zbb,\text{GN2X}_{60\%}}$	1547 ± 254	1841 ± 163	844 ± 72
$\mu_{\text{post-tag}}$	1.515 ± 0.297	0.939 ± 0.150	1.071 ± 0.184
Data stat. error	0.163	0.125	0.160
MC stat. error	0.249	0.083	0.091

Table 7.4 Pre-tag ($\mu_{\text{pre-tag}}$) signal strength measured using $Z(\to \ell^+\ell^-)$+jets calibration methods.

p_T [GeV]	$450 < p_T < 500$	$500 < p_T < 600$	$600 < p_T < 1000$
$N^{\text{data}}_{\ell\ell}$	1967	1827	1137
$N^{\text{MC}}_{\text{bkg},\ell\ell}$	97.3 ± 2.4	49.7 ± 1.5	34.3 ± 1.2
$N^{\text{MC}}_{Z\to\ell\ell}$	1607.5 ± 6.9	1559.9 ± 6.1	891.2 ± 3.9
$\mu_{\text{pre-tag}}$	1.163 ± 0.028	1.139 ± 0.028	1.237 ± 0.038
Data stat. error	0.028	0.027	0.038
MC stat. error	0.005	0.005	0.006

Table 7.5 Pre-tag ($\mu_{\text{pre-tag}}$) and post-tag ($\mu_{\text{post-tag}}$) signal strength and the resulting signal efficiency scale factors (SF) for the $X \rightarrow b\bar{b}$ tagger at 60 % efficiency WP measured using $Z(\rightarrow b\bar{b}) + \text{jets}$ calibration methods. Systematic uncertainties are also shown.

p_T [GeV]		$450 < p_T < 500$	$500 < p_T < 600$	$600 < p_T < 1000$
$\mu_{\text{post-tag}}$		1.52	0.94	1.07
$\mu_{\text{pre-tag}}$		1.16	1.14	1.24
SF		1.30	0.83	0.87
Uncertainty ($\pm\sigma$) of SFs				
Data stat.		0.14	0.11	0.13
Others from previous study [22]		0.34	0.21	0.21
Syst. ($\mu_{\text{post-tag}}$)	MC stat.	0.21	0.07	0.07
	Fit model	0.16	0.10	0.02
	Spurious signal	0.21	0.03	0.08
Syst. ($\mu_{\text{pre-tag}}$)	MC stat.	<0.01	<0.01	<0.01
	Other background modeling	0.01	<0.01	<0.01
Total uncertainty		0.50	0.27	0.27

Chapter 8 Conclusion and Outlook

The Higgs boson generated with high transverse momentum presents an opportunity for measuring the charm Yukawa coupling, and also exhibits sensitivity to novel physics. To explore the Higgs boson under these conditions, advanced techniques for object reconstruction and identification, and innovative strategies for physics analysis are essential.

Throughout this thesis, the in-situ calibration of a novel $X \to b\bar{b}$ tagger using $Z(\to b\bar{b})$ + jets events is presented, and an attempt to improve the performance of the $X \to b\bar{b}$ tagger is made.

Calibrating the $X \to b\bar{b}$ tagger is a crucial endeavor to enable its utilization in physics analyses focused on boosted $b\bar{b}$ and even $c\bar{c}$ topologies. Z + jets events are employed to determine signal scale factors for $p_T > 450\,\text{GeV}$ using data collected by the ATLAS experiment in Run 2. Dijet, $\bar{t}t$, and W + jets are considered as background. The dijet background is modeled by fitting the data directly using an exponentiated polynomial or polynomial function, depending on p_T. While other backgrounds are negligible small and neglected. The scale factors measured for the $X \to b\bar{b}$ tagger at 60% working point (WP) are 1.30 ± 0.50 for $450 < p_T < 500\,\text{GeV}$, 0.83 ± 0.27 for $500 < p_T < 600\,\text{GeV}$, and 0.87 ± 0.27 for $600 < p_T < 1000\,\text{GeV}$. This calibration is achieved by the methodology documented in Ref. [22].

Concerning the calibration work, some improvements can be made in the future as follows.

Systematic Uncertainties All the systematic uncertainties described in Section 7.1 shall be evaluated. Due to the MC samples preparation problem etc., only part of the systematic uncertainties is considered in this thesis. For example, the uncertainties on the modeling of the Z + jets events can be evaluated using the alternative Herwig++ generator [74] samples. Also, a spurious signal uncertainty can be evaluated by applying signal + background fit to the mass distribution of the background-only MC samples.

Unbinned Fit The binned fit is used for $\mu_{\text{post-tag}}$ in this thesis, which differs from the unbinned fit used in the calibration work before [22]. Though the unbinned fit will take more time to perform, it is more accurate than the binned fit, and thus should reduce the systematic uncertainties of the scale factors.

Tagger Efficiency-Mass Dependence From the comparison plots of data and MC samples, the tagger efficiency is seen to be dependent on the large-R jet mass, causing the curve of the dijet background to be not a smooth falling curve but wiggled and thus it increases the difficulty of the background modeling. In this thesis, some parameters of the modeling functions are fixed, which may cause the systematic uncertainties to be underestimated.

The investigations into boosted Higgs boson physics presented in this thesis, including the $X \to b\bar{b}$ tagger introduction, calibration and improvement, has the potential for significant value in future searches. The methodology used to calibrate the $X \to b\bar{b}$ tagger using $Z \to b\bar{b}$ events, can be also applied to calibrate the $X \to c\bar{c}$ tagger, and potentially enables the observation of $H \to c\bar{c}$ (charm Yukawa), which is a goal of my further research. Also, these investigations carry relevance for the LHC Run 3 and upcoming HL-LHC plans, and we look forward to obtaining further insights in the future regarding boosted Higgs boson studies and new physics searches.

References

[1] Standard model of elementary particles. https://en.wikipedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg.

[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, 2012, 716: 1-29. DOI: 10.1016/j.physletb.2012.08.020.

[3] Chatrchyan S, et al. Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC. Phys. Lett. B, 2012, 716: 30-61. DOI: 10.1016/j.physletb.2012.08.021.

[4] ATLAS Collaboration. Combined measurement of the Higgs boson mass from the $H \to \gamma\gamma$ and $H \to ZZ^* \to 4\ell$ decay channels with the ATLAS detector using $\sqrt{s} = 7$, 8 and 13 TeV pp collision data. Phys. Rev. Lett., 2023, 131: 251802. DOI: 10.1103/PhysRevLett.131.251802.

[5] ATLAS Collaboration. Evidence of off-shell Higgs boson production from ZZ leptonic decay channels and constraints on its total width with the ATLAS detector. Phys. Lett. B, 2023, 846: 138223. DOI: 10.1016/j.physletb.2023.138223.

[6] 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, 2014, 90(11): 112015. DOI: 10.1103/PhysRevD.90.112015.

[7] Khachatryan V, et al. Observation of the Diphoton Decay of the Higgs Boson and Measurement of Its Properties. Eur. Phys. J. C, 2014, 74(10): 3076. DOI: 10.1140/epjc/s10052-014-3076-z.

[8] 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, 2015, 91(1): 012006. DOI: 10.1103/PhysRevD.91.012006.

[9] Chatrchyan S, et al. Measurement of the Properties of a Higgs Boson in the Four-Lepton Final State. Phys. Rev. D, 2014, 89(9): 092007. DOI: 10.1103/PhysRevD.89.092007.

[10] ATLAS Collaboration. Observation and measurement of Higgs boson decays to WW* with the ATLAS detector. Phys. Rev. D, 2015, 92(1): 012006. DOI: 10.1103/PhysRevD.92.012006.

[11] Chatrchyan S, et al. Measurement of Higgs Boson Production and Properties in the WW Decay Channel with Leptonic Final States. JHEP, 2014, 01: 096. DOI: 10.1007/JHEP01(2014)096.

[12] Sirunyan A M, et al. Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector. Phys. Lett. B, 2018, 779: 283-316. DOI: 10.1016/j.physletb.2018.02.004.

[13] 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, 2019, 99: 072001. DOI: 10.1103/PhysRevD.99.072001.

[14] ATLAS Collaboration. Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector. Phys. Lett. B, 2018, 784: 173-191. DOI: 10.1016/j.physletb.2018.07.035.

[15] Sirunyan A M, et al. Observation of $t\bar{t}H$ production. Phys. Rev. Lett., 2018, 120(23): 231801. DOI: 10.1103/PhysRevLett.120.231801.

[16] ATLAS Collaboration. Observation of $H \to b\bar{b}$ decays and VH production with the ATLAS detector. Phys. Lett. B, 2018, 786: 59-86. DOI: 10.1016/j.physletb.2018.09.013.

[17] Sirunyan A M, et al. Observation of Higgs boson decay to bottom quarks. Phys. Rev. Lett., 2018, 121(12): 121801. DOI: 10.1103/PhysRevLett.121.121801.

[18] ATLAS Collaboration. Combined measurements of Higgs boson production and decay using up to 80 fb^{-1} of proton-proton collision data at $\sqrt{s} = 13$ TeV collected with the ATLAS experiment. 2019. https://cds.cern.ch/record/2668375.

[19] ATLAS and CMS Collaborations. Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV. JHEP, 2016, 08: 045. DOI: 10.1007/JHEP08(2016)045.

[20] de Florian D, et al. Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector: 2/2017. 2016. arXiv: 1610.07922.

[21] ATLAS Collaboration. Transformer Neural Networks for Identifying Boosted Higgs Bosons decaying into $b\bar{b}$ and $c\bar{c}$ in ATLAS. 2023. https://cds.cern.ch/record/2866601.

[22] Calibration r21 results of atlas flavour tagging xbb. https://xbb-docs.docs.cern.ch/Calibration R21/Results/.

[23] LHC Machine. JINST, 2008, 3: S08001. DOI: 10.1088/1748-0221/3/08/S08001.

[24] ATLAS Collaboration. The ATLAS Experiment at the CERN Large Hadron Collider. JINST, 2008, 3: S08003. DOI: 10.1088/1748-0221/3/08/S08003.

[25] Mobs E. The CERN accelerator complex in 2019. Complexe des accélérateurs du CERN en 2019. 2019. https://cds.cern.ch/record/2684277.

[26] Salvachua B. Overview of Proton-Proton Physics during Run 2//9th LHC Operations Evian Workshop. 2019: 7-14.

[27] Fartoukh S, et al. LHC Configuration and Operational Scenario for Run 3. https://cds.cern.ch/record/2790409.

[28] Public atlas luminosity results for run-1 of the lhc. https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun1.

[29] Public atlas luminosity results for run-2 of the lhc. https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2.

[30] Public atlas luminosity results for run-3 of the lhc. https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun3.

[31] The hl-lhc project. https://hilumilhc.web.cern.ch/content/hl-lhc-project.

[32] ATLAS Collaboration. Vertex Reconstruction Performance of the ATLAS Detector at "\sqrts = 13~\textrmTeV". 2015. https://cds.cern.ch/record/2037717.

[33] Pequenao J. Event Cross Section in a computer generated image of the ATLAS detector. 2008. https://cds.cern.ch/record/1096081.

[34] The inner detector. https://atlas.cern/Discover/Detector/Inner-Detector.

[35] ATLAS Collaboration. Alignment of the ATLAS Inner Detector in Run-2. Eur. Phys. J. C, 2020, 80(12): 1194. DOI: 10.1140/epjc/s10052-020-08700-6.

[36] ATLAS Collaboration. ATLAS calorimeter performance Technical Design Report. 1996. https://cds.cern.ch/record/331059.

[37] ATLAS Collaboration. ATLAS liquid argon calorimeter: Technical design report. 1996. https://cds.cern.ch/record/331061.

[38] Aharrouche M, et al. Energy linearity and resolution of the ATLAS electromagnetic barrel calorimeter in an electron test-beam. Nucl. Instrum. Meth. A, 2006, 568: 601-623. DOI: 10.1016/j.nima.2006.07.053.

[39] Abreu H, et al. Performance of the electronic readout of the ATLAS liquid argon calorimeters. JINST, 2010, 5: P09003. DOI: 10.1088/1748-0221/5/09/P09003.

[40] Aleksa M, Cleland W, Enari Y, et al. ATLAS Liquid Argon Calorimeter Phase-I Upgrade Technical Design Report. 2013. https://cds.cern.ch/record/1602230.

[41] ATLAS Collaboration. Electron and photon energy calibration with the ATLAS detector using 2015–2016 LHC proton-proton collision data. JINST, 2019, 14(03): P03017. DOI: 10.1088/1748-0221/14/03/P03017.

[42] ATLAS Collaboration. Luminosity determination in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector at the LHC. Eur. Phys. J. C, 2023, 83(10): 982. DOI: 10.1140/epjc/s10052-023-11747-w.

[43] ATLAS Collaboration. The ATLAS Simulation Infrastructure. Eur. Phys. J. C, 2010, 70: 823-874. DOI: 10.1140/epjc/s10052-010-1429-9.

[44] Agostinelli S, et al. GEANT4–a simulation toolkit. Nucl. Instrum. Meth. A, 2003, 506: 250-303. DOI: 10.1016/S0168-9002(03)01368-8.

[45] Sjostrand T, Mrenna S, Skands P Z. A Brief Introduction to PYTHIA 8.1. Comput. Phys. Commun., 2008, 178: 852-867. DOI: 10.1016/j.cpc.2008.01.036.

[46] ATLAS Collaboration. The Pythia 8 A3 tune description of ATLAS minimum bias and inelastic measurements incorporating the Donnachie-Landshoff diffractive model. 2016. https://cds.cern.ch/record/2206965.

[47] Ball R D, et al. Parton distributions with LHC data. Nucl. Phys. B, 2013, 867: 244-289. DOI: 10.1016/j.nuclphysb.2012.10.003.

[48] Gleisberg T, Hoeche S, Krauss F, et al. Event generation with SHERPA 1.1. JHEP, 2009, 02: 007. DOI: 10.1088/1126-6708/2009/02/007.

[49] Lange D J. The EvtGen particle decay simulation package. Nucl. Instrum. Meth. A, 2001, 462: 152-155. DOI: 10.1016/S0168-9002(01)00089-4.

[50] Bothmann E, et al. Event Generation with Sherpa 2.2. SciPost Phys., 2019, 7(3): 034. DOI: 10.21468/SciPostPhys.7.3.034.

[51] Schumann S, Krauss F. A Parton shower algorithm based on Catani-Seymour dipole factorisation. JHEP, 2008, 03: 038. DOI: 10.1088/1126-6708/2008/03/038.

[52] Ball R D, et al. Parton distributions for the LHC Run II. JHEP, 2015, 04: 040. DOI: 10.1007/JHEP04(2015)040.

[53] Alioli S, Nason P, Oleari C, et al. NLO single-top production matched with shower in POWHEG: s- and t-channel contributions. JHEP, 2009, 09: 111. DOI: 10.1088/1126-6708/2009/09/111.

[54] Re E. Single-top Wt-channel production matched with parton showers using the POWHEG method. Eur. Phys. J. C, 2011, 71: 1547. DOI: 10.1140/epjc/s10052-011-1547-z.

[55] ATLAS Pythia 8 tunes to 7 TeV data. 2014. https://cds.cern.ch/record/1966419.

[56] ATLAS Collaboration. Summary of ATLAS Pythia 8 tunes. 2012. https://cds.cern.ch/record/1474107.

[57] Alwall J, Frederix R, Frixione S, et al. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. JHEP, 2014, 07: 079. DOI: 10.1007/JHEP07(2014)079.

[58] ATLAS Collaboration. Optimisation of large-radius jet reconstruction for the ATLAS detector in 13 TeV proton–proton collisions. Eur. Phys. J. C, 2021, 81(4): 334. DOI: 10.1140/epjc/s10052-021-09054-3.

[59] ATLAS Collaboration. Jet reconstruction and performance using particle flow with the ATLAS Detector. Eur. Phys. J. C, 2017, 77(7): 466. DOI: 10.1140/epjc/s10052-017-5031-2.

[60] ATLAS Collaboration. Improving jet substructure performance in ATLAS using Track-CaloClusters. 2017. https://cds.cern.ch/record/2275636.

[61] Cacciari M, Salam G P, Soyez G. The anti-k_t jet clustering algorithm. JHEP, 2008, 04: 063. DOI: 10.1088/1126-6708/2008/04/063.

[62] Cacciari M, Salam G P, Soyez G. FastJet User Manual. Eur. Phys. J. C, 2012, 72: 1896. DOI: 10.1140/epjc/s10052-012-1896-2.

[63] Berta P, Spousta M, Miller D W, et al. Particle-level pileup subtraction for jets and jet shapes. JHEP, 2014, 06: 092. DOI: 10.1007/JHEP06(2014)092.

[64] Cacciari M, Salam G P, Soyez G. SoftKiller, a particle-level pileup removal method. Eur. Phys. J. C, 2015, 75(2): 59. DOI: 10.1140/epjc/s10052-015-3267-2.

[65] Cacciari M, Salam G P, Soyez G. The Catchment Area of Jets. JHEP, 2008, 04: 005. DOI: 10.1088/1126-6708/2008/04/005.

[66] Cornelissen T, Elsing M, Fleischmann S, et al. Concepts, Design and Implementation of the ATLAS New Tracking (NEWT). 2007. https://cds.cern.ch/record/1020106.

[67] Cornelissen T, Elsing M, Gavrilenko I, et al. The new ATLAS track reconstruction (NEWT). J. Phys. Conf. Ser., 2008, 119: 032014. DOI: 10.1088/1742-6596/119/3/032014.

[68] Perigee representation and atlas tracking flow chart. https://atlassoftwaredocs.web.cern.ch/trackingTutorial/idoverview/.

[69] Jettruthlabelingtool.cxx. https://gitlab.cern.ch/atlas/athena/-/blob/21.2/PhysicsAnalysis/AnalysisCommon/ParticleJetTools/Root/JetTruthLabelingTool.cxx.

[70] Battaglia M, Beresford L, Celli F, et al. Measurement of Inclusive Higgs Boson Production at High p_T^H in the $H \to b\bar{b}$ Decay Mode. Geneva: CERN, 2019. https://cds.cern.ch/record/2703097.

[71] ATLAS Collaboration. Search for resonances in diphoton events at $\sqrt{s}$=13 TeV with the ATLAS detector. JHEP, 2016, 09: 001. DOI: 10.1007/JHEP09(2016)001.

[72] Lomax R. Statistical concepts: A second course. Lawrence Erlbaum Associates, 2007. https://books.google.de/books?id=p17rT373FNAC.

[73] ATLAS Collaboration. Efficiency corrections for a tagger for boosted $H \to b\bar{b}$ decays in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. 2021. https://cds.cern.ch/record/2777811.

[74] Bahr M, et al. Herwig++ Physics and Manual. Eur. Phys. J. C, 2008, 58: 639-707. DOI: 10.1140/epjc/s10052-008-0798-9.

Appendix A Additional Plots

A.1 Additional Plots for Comparison of Data and Simulation

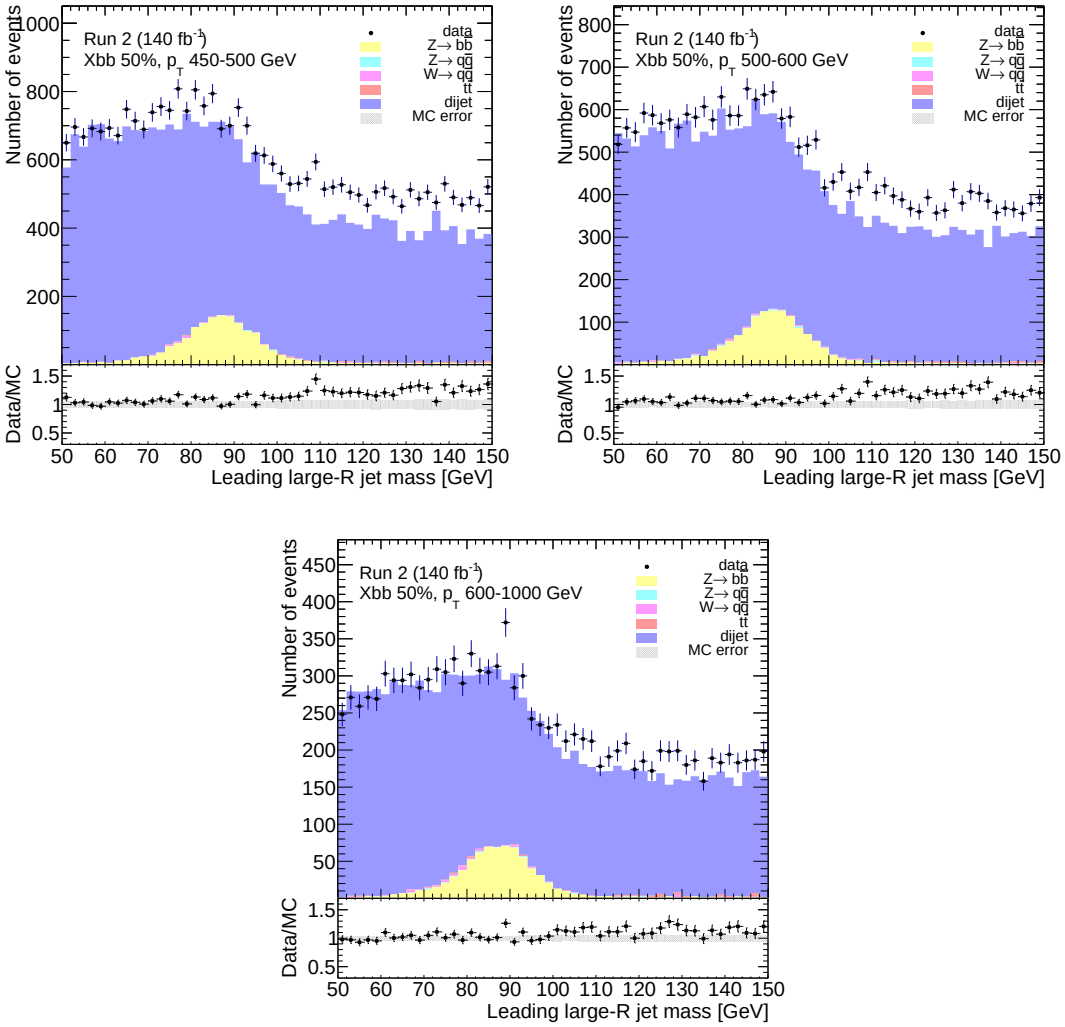


Figure A.1 The comparison of data and MC samples prediction after 50% WP of $X \to b\bar{b}$ tagger for different p_T bins. Different MC samples are stacked together. The MC error is shown as the shaded band.

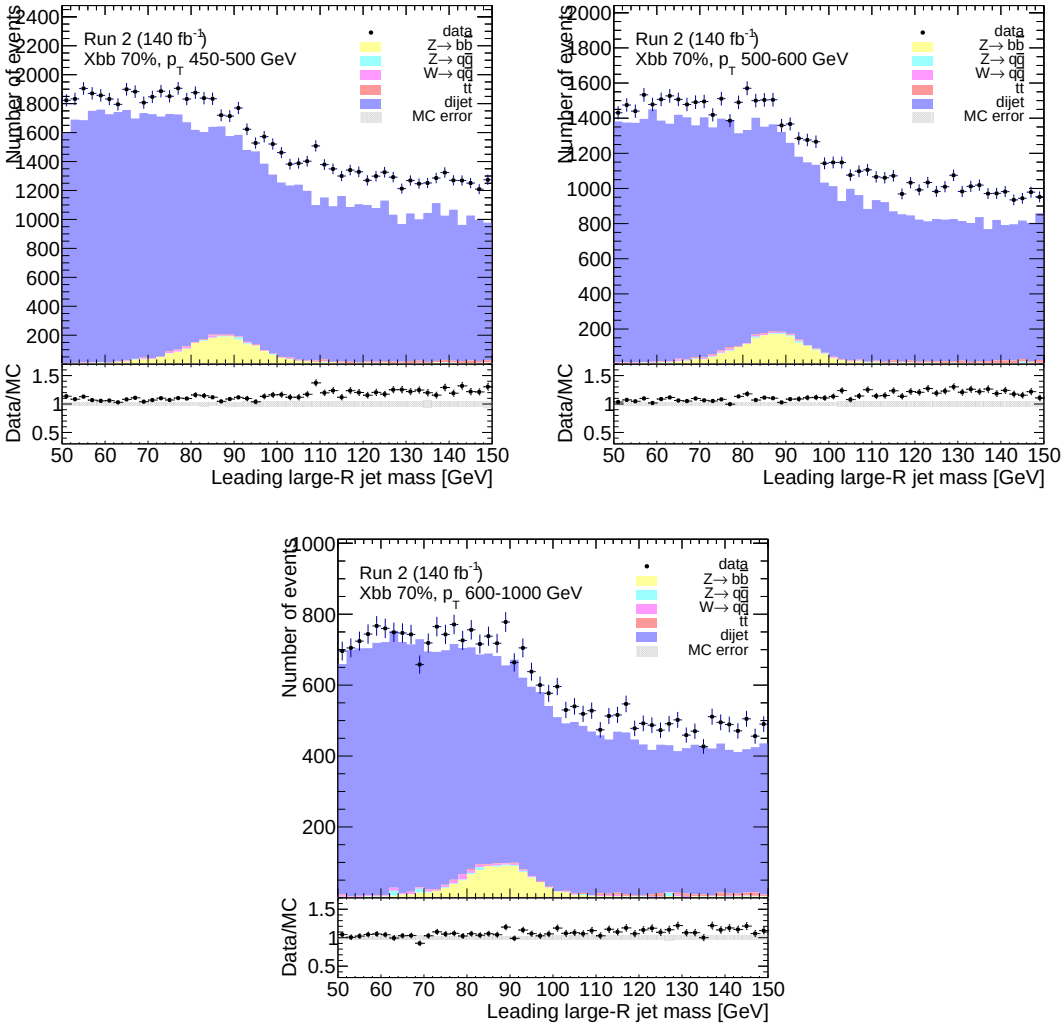


Figure A.2 The comparison of data and MC samples prediction after 70% WP of $X \to b\bar{b}$ tagger for different p_T bins. Different MC samples are stacked together. The MC error is shown as the shaded band.

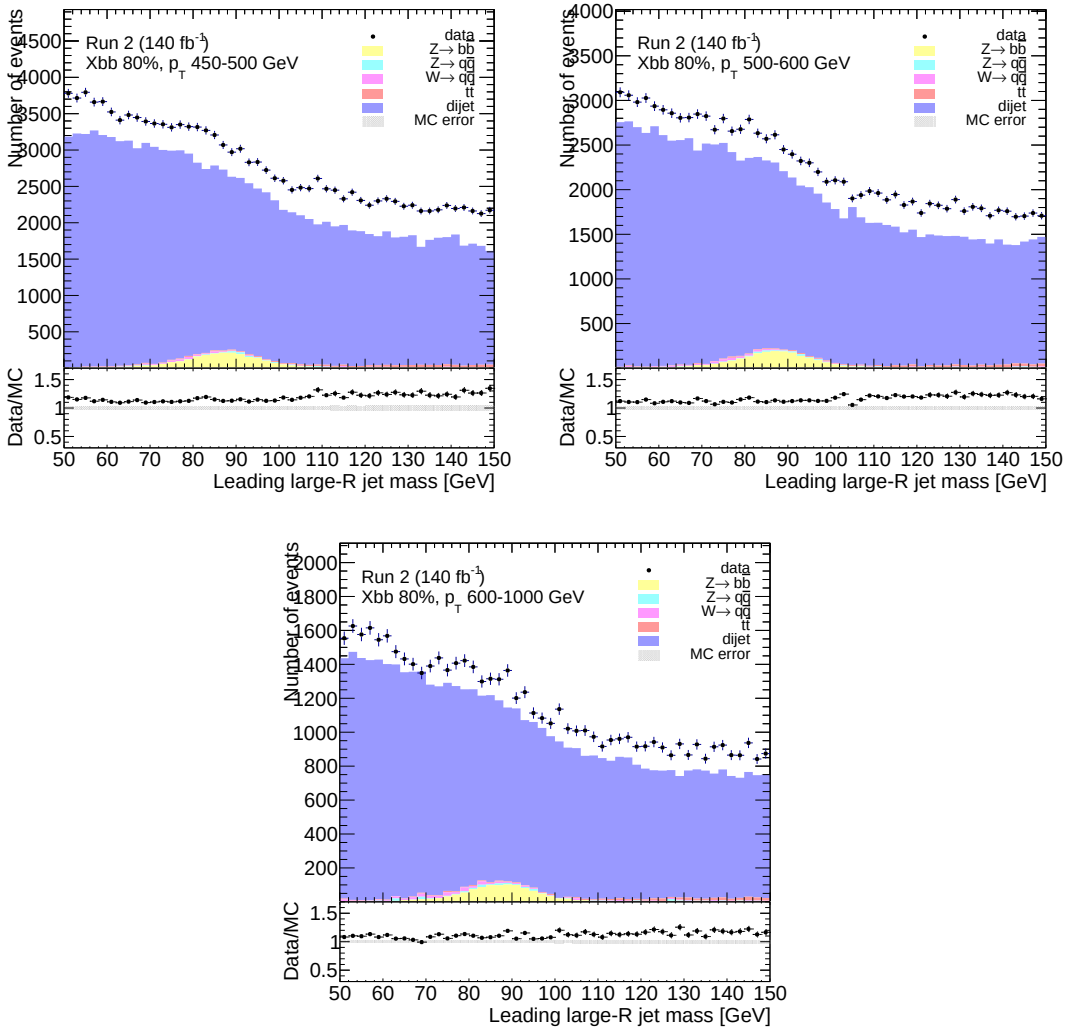


Figure A.3 The comparison of data and MC samples prediction after 80% WP of $X \to b\bar{b}$ tagger for different p_T bins. Different MC samples are stacked together. The MC error is shown as the shaded band.

A.2 Additional Plots and Tables for Signal Modeling

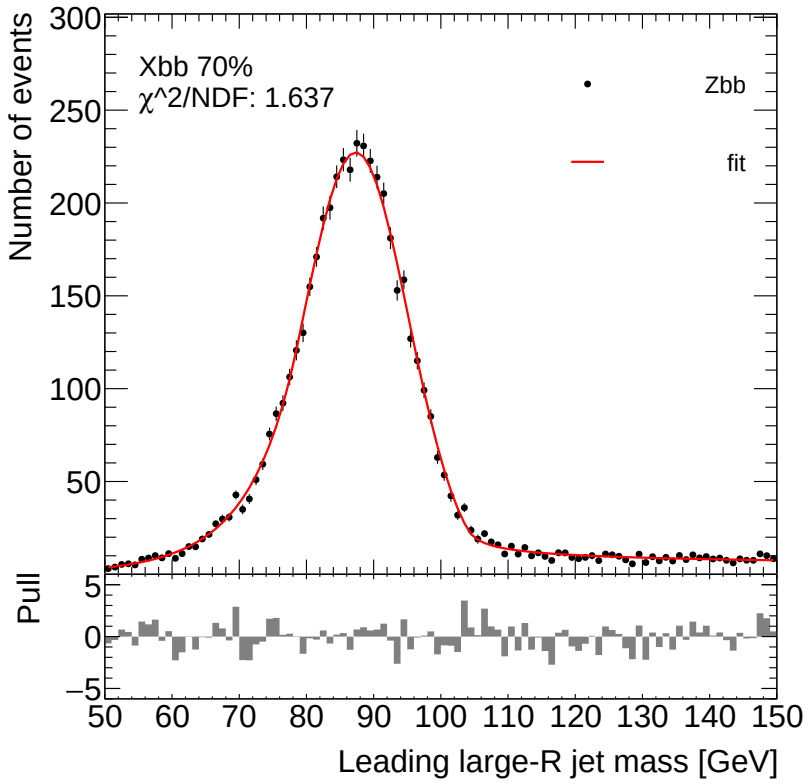


Figure A.4 The χ^2 fits to the Z candidate mass distribution via a DSCB function, passing the $X \to b\bar{b}$ 50% WP for $450 \leq p_T < 1000$ GeV.

Table A.1 The parameters of the DSCB function for the $Z \to b\bar{b}$ MC templates after the $X \to b\bar{b}$ tagging requirement at 50% working point for $450 \leq p_T < 1000$ GeV.

N	172.4 ± 4.37
m_Z	87.2 ± 0.20
σ_Z	7.56 ± 0.22
α_L	1.08 ± 0.10
n_L	13.1 ± 11.7
α_H	2.08 ± 0.13
n_H	0.62 ± 0.18
χ^2	135.4
χ^2/NDF	1.46
Yield	3709.8

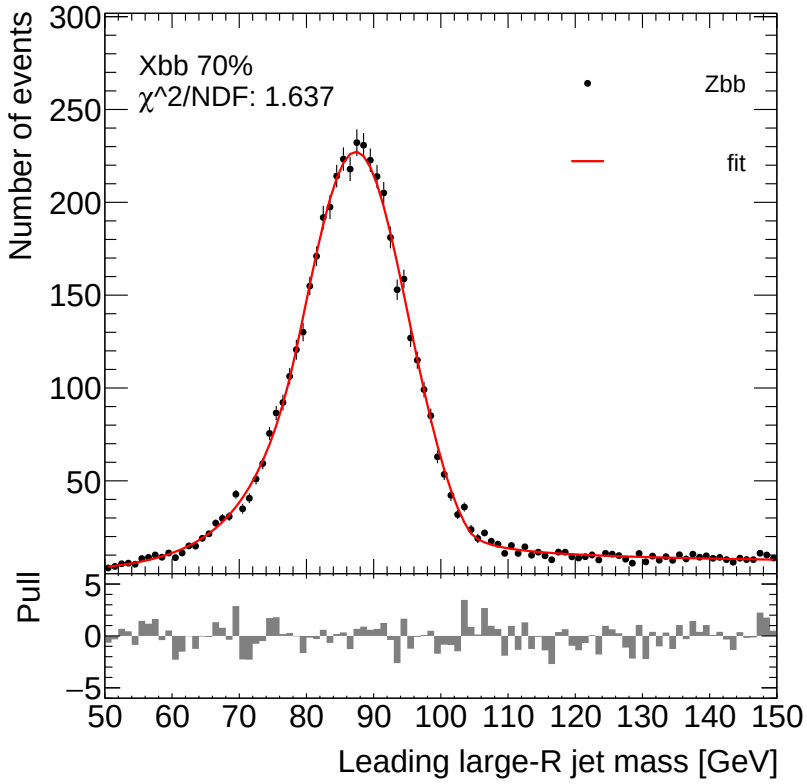


Figure A.5 The χ^2 fits to the Z candidate mass distribution via a DSCB function, passing the $X \to b\bar{b}$ 70% WP for $450 \leq p_T < 1000$ GeV.

Table A.2 The parameters of the DSCB function for the $Z \to b\bar{b}$ MC templates after the $X \to b\bar{b}$ tagging requirement at 70% working point for $450 \leq p_T < 1000$ GeV.

N	227.5 ± 4.74
m_Z	87.3 ± 0.18
σ_Z	7.86 ± 0.18
α_L	1.11 ± 0.10
n_L	13.6 ± 11.8
α_H	2.12 ± 0.08
n_H	0.31 ± 0.08
χ^2	152.1
NDF	93
χ^2/NDF	1.64
Yield	5189.3

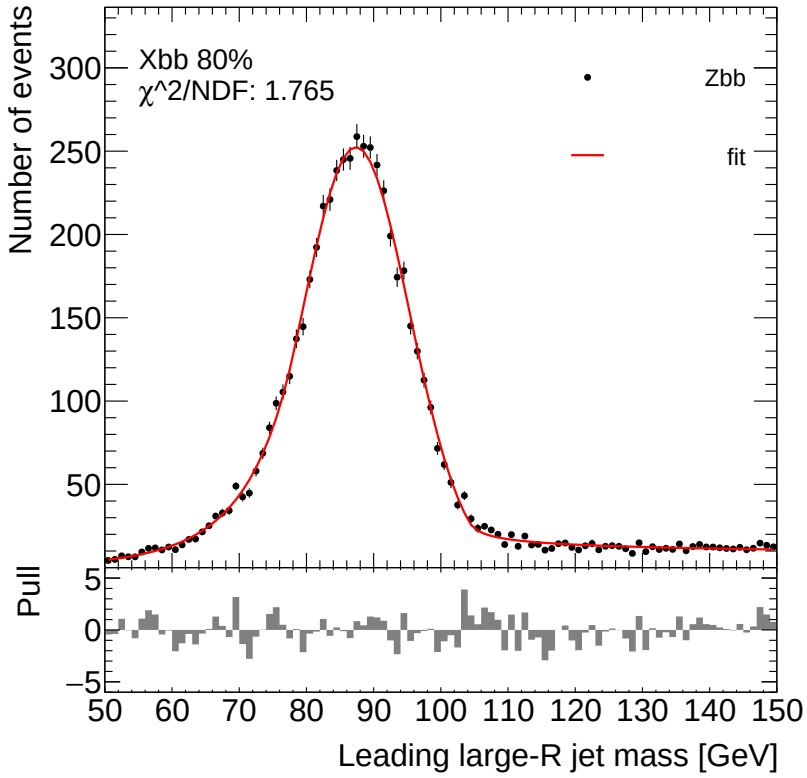


Figure A.6 The χ^2 fits to the Z candidate mass distribution via a DSCB function, passing the $X \to b\bar{b}$ 80% WP for $450 \leq p_T < 1000$ GeV.

Table A.3 The parameters of the DSCB function for the $Z \to b\bar{b}$ MC templates after the $X \to b\bar{b}$ tagging requirement at 80% working point for $450 \leq p_T < 1000$ GeV.

N	252.5 ± 5.01
m_Z	87.4 ± 0.17
σ_Z	8.00 ± 0.18
α_L	1.14 ± 0.10
n_L	9.24 ± 5.80
α_H	2.11 ± 0.07
n_H	0.23 ± 0.07
χ^2	164.2
NDF	93
χ^2/NDF	1.77
Yield	5943.0

A.3 Additional Results for Fit Model Uncertainties

Table A.4 The parameters of fitting the real data after the $X \to b\bar{b}$ tagging requirement at 60% working point for different p_T bins for alternative mass range.

Parameter	Large-R jet p_T [GeV]		
	450-500 GeV	500-600 GeV	600-1000 GeV
m_Z	82.9 ± 0.78	84.7 ± 0.89	85.4 ± 1.13
σ_Z	7.73 (fixed)	7.73 (fixed)	7.73 (fixed)
α_L	1.08 (fixed)	1.08 (fixed)	1.08 (fixed)
n_L	16.4 (fixed)	16.4 (fixed)	16.4 (fixed)
α_H	2.12 (fixed)	2.12 (fixed)	2.12 (fixed)
n_H	0.44 (fixed)	0.44 (fixed)	0.44 (fixed)
a_0	8.60 ± 9.37	-111.7 ± 7.85	30.63 ± 1.59
a_1	2139.8 ± 12.66	2153.0 ± 10.42	7.34 ± 0.10
a_2	-2645.0 ± 8.35	-2602.8 ± 11.96	-8.56 ± 0.10
a_3	942.5 ± 6.46	914.5 ± 7.21	2.93 ± 0.06
χ^2	97.1	101.0	100.6
χ^2/NDF	1.16	1.20	1.20
Signal	2055.5	1520.3	889.0
Background	46114.4	36478.4	17696.8

A.4 Additional Results for Spurious Signal Test

Table A.5 The parameters of fitting the mass distribution of the dijet MC samples using signal plus background model after the $X \to b\bar{b}$ tagging requirement at 60% working point for different p_T bins.

Parameter	Large-R jet p_T [GeV]		
	450-500 GeV	500-600 GeV	600-1000 GeV
Spurious Signal	-502.386 ± 240.502	-23.9656 ± 155.688	-111.684 ± 62.2142

A.5 Additional Results for $\mu_{\text{pre-tag}}$ Measurement

Table A.6 Systematics on $\mu_{\text{pre-tag}}$, $450 < p_T < 500$

Modeling $Z(\ell\ell)$ + jets	$^{+0.18}_{-0.14}$
Modeling ZZ	± 0.002
Modeling WZ	± 0.004
Modeling WW	± 0.000
Modeling $Z(\tau\tau)$ + jets	± 0.006
Modeling $W(\ell\nu)$ + jets	± 0.000
Luminosity	$^{+0.02}_{-0.02}$
Total	$^{+0.18}_{-0.14}$

Table A.7 Yields, $450 < p_T < 500$

$Z(\ell\ell)$ + jets	1607.5 ± 6.9
ZZ	15.3 ± 0.8
WZ	31.6 ± 1.2
WW	0.0 ± 0.0
$Z(\tau\tau)$ + jets	50.4 ± 1.9
$W(\ell\nu)$ + jets	0.08 ± 0.05
Background	97.3 ± 2.4
signal + background	1704.8 ± 7.3
data	1967

Table A.8 Systematics on $\mu_{\text{pre-tag}}$, $500 < p_T < 600$

Modeling $Z(\ell\ell)$ + jets	$^{+0.11}_{-0.09}$
Modeling ZZ	± 0.002
Modeling WZ	± 0.004
Modeling WW	± 0.000
Modeling $Z(\tau\tau)$ + jets	± 0.000
Modeling $W(\ell\nu)$ + jets	± 0.000
Luminosity	$^{+0.02}_{-0.02}$
Total	$^{+0.11}_{-0.10}$

Table A.9 Yields, $500 < p_T < 600$

$Z(\ell\ell)$ + jets	1559.9 ± 6.1
ZZ	15.4 ± 0.9
WZ	34.1 ± 1.3
WW	0.0 ± 0.0
$Z(\tau\tau)$ + jets	0.14 ± 0.07
$W(\ell\nu)$ + jets	0.10 ± 0.06
Background	49.7 ± 1.5
signal + background	1609.6 ± 6.3
data	1827

Table A.10 Systematics on $\mu_{\text{pre-tag}}$, $600 < p_T < 1000$

Modeling $Z(\ell\ell)$ + jets	$^{+0.41}_{-0.25}$
Modeling ZZ	± 0.002
Modeling WZ	± 0.006
Modeling WW	± 0.000
Modeling $Z(\tau\tau)$ + jets	± 0.000
Modeling $W(\ell\nu)$ + jets	± 0.000
Luminosity	$^{+0.02}_{-0.02}$
Total	$^{+0.42}_{-0.25}$

Table A.11 Yields, $600 < p_T < 1000$

$Z(\ell\ell)$ + jets	891.2 ± 3.9
ZZ	6.8 ± 0.6
WZ	27.4 ± 1.1
WW	0.0 ± 0.0
$Z(\tau\tau)$ + jets	-0.03 ± 0.10
$W(\ell\nu)$ + jets	0.12 ± 0.05
Background	34.3 ± 1.2
signal + background	925.5 ± 4.1
data	1137

Appendix B Identifying Boosted Higgs Bosons Decaying using Graph Neural Network

B.1 Introduction

GN2X, a new algorithm has been developed to identify the decays of high-p_T Higgs bosons to $b\bar{b}/c\bar{c}$ pairs. This algorithm is trained to classify large-R jets based on their origin, discriminating jets. The background processes are considered to be multijet processes and fully hadronic top-quark decays. GN2X utilizes the recent advances in graph neural networks (GNNs) and transformer architectures to learn the jet substructure.

In GN2X, the charged particle trajectories within the large-R jet are employed to discern the characteristic indicators of b- and c-hadron decays: displaced secondary (and perhaps tertiary, in the case of b-hadrons) vertices and tracks with substantial impact parameters. Other GN2X versions are also investigated, where trajectories are amalgamated with large-R jet calorimeter constituents and subjets.

B.2 Neural Network Architecture and Training

The base GN2X model takes three large-R jet variables and 20 variables associated with each track as input to the network. The large-R jet inputs include the jet transverse momentum, signed pseudorapidity, and mass. The complete set of inputs is presented in Table B.1. Up to 100 tracks associated with the jet are provided to the network, sorted by decreasing transverse impact parameter significance denoted as $s(d_0)$.

The GN2X architecture is an advancement of the GN1 architecture [B.1], as shown in Figure B.2. GN1 employs a Graph Neural network, while GN2X (and GN2) adopts a Transformer network architecture [B.2]. The model receives the sequence of tracks in a jet as input, with jet- and track-level inputs. Concatenating the jet and track inputs, as shown in Figure B.1, the resulting combined jet-track sequence vectors are input into a per-track initializer network. The initializer network for each input type comprises two dense layers projecting the input representations to an embedding dimension of 192.

The track representations feed into a Transformer Encoder, where the transformer architecture utilized aligns with that introduced in Ref. [B.3]. Multiple Layer Normalization layers [B.4] are incorporated to enhance stability during training, along with residual

Table B.1 Input features to the GN2X model.

Jet Input	Description
p_T	Large-R jet transverse momentum
η	Signed large-R jet pseudorapidity
mass	Large-R jet mass
Track Input	**Description**
q/p	Track charge divided by momentum (a measure of curvature)
$d\eta$	Pseudorapidity of track relative to the large-R jet η
$d\phi$	Azimuthal angle of the track, relative to the large-R jet ϕ
d_0	Closest distance from track to primary vertex (PV) in the transverse plane
$z_0 \sin\theta$	Closest distance from track to PV in the longitudinal plane
$\sigma(q/p)$	Uncertainty on q/p
$\sigma(\theta)$	Uncertainty on track polar angle θ
$\sigma(\phi)$	Uncertainty on track azimuthal angle ϕ
$s(d_0)$	Lifetime signed transverse IP significance
$s(z_0 \sin\theta)$	Lifetime signed longitudinal IP significance
nPixHits	Number of pixel hits
nSCTHits	Number of SCT hits
nIBLHits	Number of IBL hits
nBLHits	Number of B-layer hits
nIBLShared	Number of shared IBL hits
nIBLSplit	Number of split IBL hits
nPixShared	Number of shared pixel hits
nPixSplit	Number of split pixel hits
nSCTShared	Number of shared SCT hits

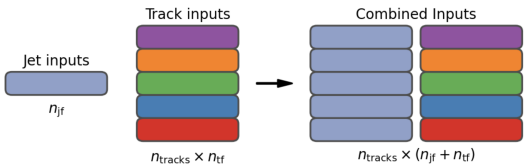


Figure B.1 The inputs to the GN2X model are the two jet features ($n_{jf} = 3$), and an array of n_t, where each track is described by 20 track features ($n_{tf} = 20$). The jet features are copied for each of the tracks, and the combined jet-track vectors of length 23 form the inputs of GN2X [B.1].

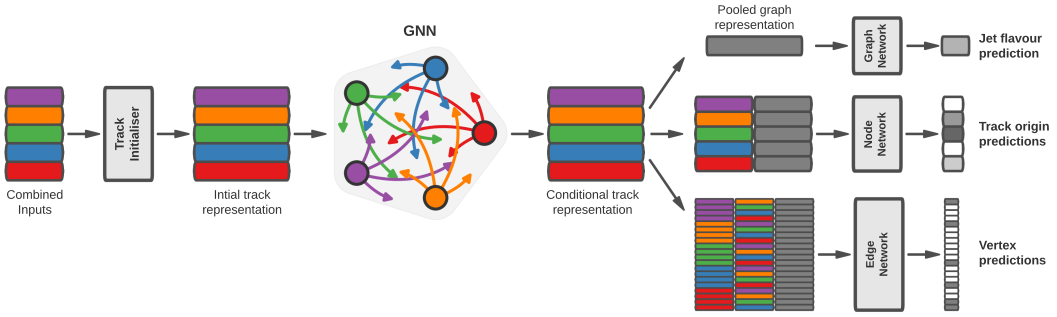


Figure B.2 The network architecture of GN1. Inputs are fed into a per-track initialization network, which outputs an initial latent representation of each track. These representations are then used to populate the node features of a fully connected graph network. After the graph network, the resulting node representations are used to predict the jet flavor, the track origins, and the track-pair vertex compatibility [B.1]

.

connections. GN2X employs 6 encoder blocks with 4 attention heads. The heterogeneous models refrain from using separate transformer encoders for each input type, given the substantial increase in parameter count.

The resulting representation of each track is then combined to create a global representation of the jet for classification. This global representation is crafted through a weighted sum over the track representations, where the attention weights for the sum are learned during training.

GN2X undergoes training employing a method akin to GN1 [B.1]. The training sessions for GN2X are conducted on a cluster equipped with 4 NVIDIA A100 GPUs, requiring approximately 1 hour to complete an epoch comprising 62 million jets. The model is fine-tuned utilizing the Adam optimizer with a batch size of 1,000 for 40 epochs. Throughout the training process, the primary and auxiliary tasks of the model are assigned weights to ensure that their losses are of comparable magnitude, mirroring the approach used in Ref. [B.1]. Additionally, the model is subjected to training with a one-cycle learning rate policy, wherein the learning rate initiates at a minimal value of 10^{-7} and steadily increases over the initial 4 epochs, reaching a maximum value of 0.005. Subsequently, the learning rate undergoes a gradual decrease, concluding at 10^{-7}.

References

[B.1] ATLAS Collaboration. Graph Neural Network Jet Flavour Tagging with the ATLAS Detector. 2022.

[B.2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in neural information processing systems, 2017, 30.

[B.3] Shleifer S, Weston J, Ott M. Normformer: Improved transformer pretraining with extra normalization. 2021.

[B.4] Ba J L, Kiros J R, Hinton G E. Layer normalization. 2016.

Appendix C Studies of training GN2X tagger using Equivariant Subgraph Aggregation Networks

This appendix presents a study of training the GN2X tagger using a new method called subgraphs. It's based on an architecture called Equivariant Subgraph Aggregation Networks (ESANs) [C.1].

C.1 Motivation

The architectures of Message-Passing Neural Networks (MPNNs) are at most as expressive as the Weisfeiler-Lehman Graph Isomorphism Test (WL test) [C.2]. However, the WL test sometimes cannot distinguish between very simple graphs, as shown in Figure C.1. To overcome this limitation, an observation is that these graphs may not be distinguishable by an MPNN, but they often contain distinguishable subgraphs. Thus, ESANs represent each graph as a set of subgraphs derived by some predefined policy.

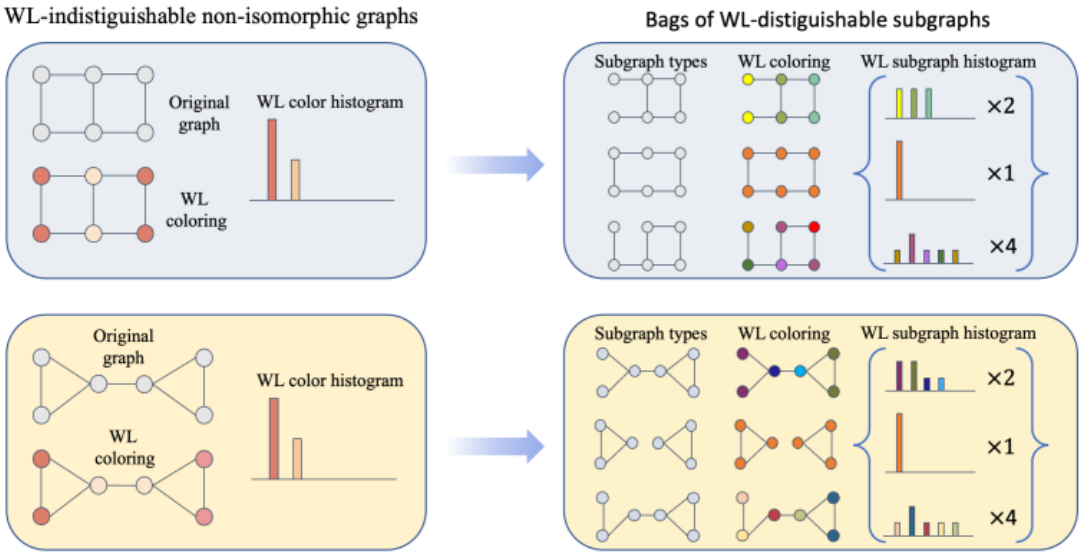


Figure C.1 Left: A pair of graphs not distinguishable by the WL test. Right: The corresponding bags (multisets) of edge-deleted subgraphs, which can be distinguished by ESAN [C.1].

Considering the subjets structure of large-R jets for the boosted Higgs boson tagging, an idea to improve the current performance of the GN2X tagger is to use the subjets as inputs of the GN2X tagger.

C.2 Architecture and Training

ESAN uses bags of subgraphs as inputs, as shown in Figure C.2. To generate subgraphs, a policy called node-deleted policy is used. In this policy, a graph is mapped to the set containing all subgraphs that can be obtained from the original graph by removing a single node. Notably, only 3 nodes to remove are considered in this study. The inputs, which are bags of subgraphs, are fed into an architecture called DSS-GNN, which is shown in Figure C.3.

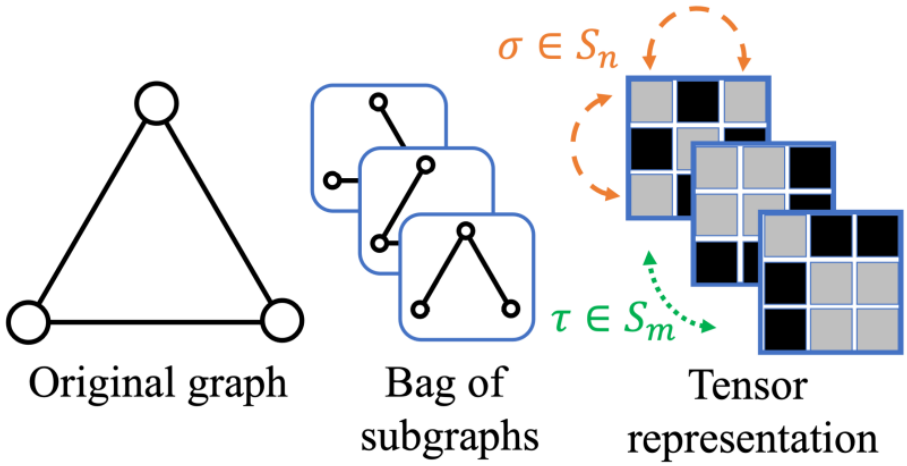


Figure C.2 The symmetry structure of a bag of subgraphs, in this case, the set of all $m = 3$ edge-deleted subgraphs. This set of subgraphs is represented as an $m \times n \times n$ tensor [C.1].

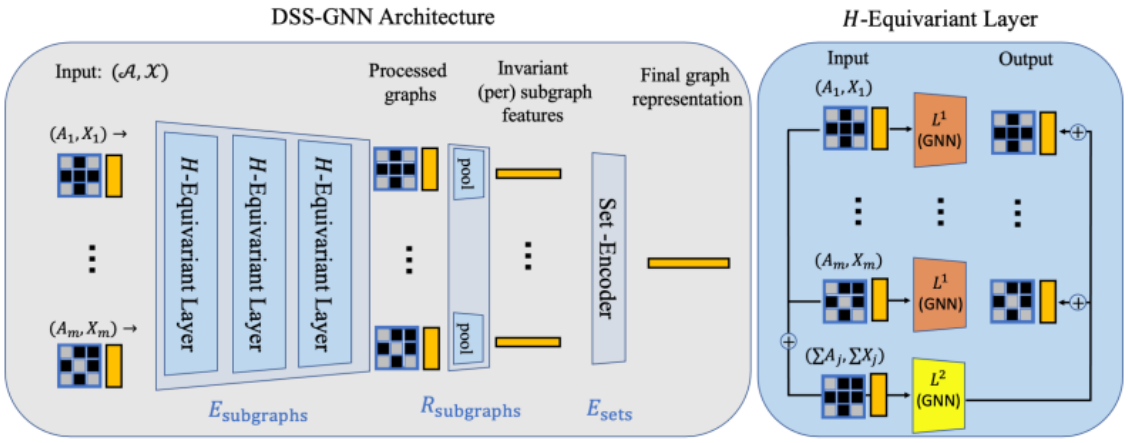


Figure C.3 DSS-GNN layers and architecture. Left panel: the DSS-GNN architecture is composed of three blocks: a Feature Encoder, a Readout Layer and a Set Encoder. Right panel: a DSS-GNN layer is constructed from a Siamese part (orange) and an information-sharing part (yellow) [C.1].

In this study, the DSS-GNN architecture will replace the GNN part of the GN2X tagger. Inside the DSS-GNN, a so-called H-equivariant Layer is used, which is a transformer encoder originally used by GN2X, for both the Siamese part and the information-sharing

part. The large-R jets will be represented as fully connected graphs in the transformer, where each node represents a track. The training process is similar to GN2X, except that the inputs are bags of subgraphs. The training sessions for this study are conducted on a cluster equipped with 2 NVIDIA A100 GPUs, requiring approximately 9 hours to complete 50 epochs comprising 1 million jets.

In addition, instead of representing the large-R jets as fully connected graphs, k-nearest neighbor graphs are also investigated. In this case, each node still represents a track, but only the k nearest neighbors of each track are connected to it, where the distance is defined as the Euclidean distance in the (η, ϕ) plane. In this study, $k = 9$ is used, which is the average number of tracks in each subjet.

C.3 Results and Conclusion

The results of this study are shown in Figure C.4 and C.5. The performance of the base GN2X tagger is also shown for comparison. The results show that the subgraph method is not so effective for fully-connected graphs. For k-nearest neighbor graphs, the subgraph methods can improve the performance, compared to the fully-connected graphs. However, graphs with attention, or so-called transformer encoder architecture, have better performance than the k-nearest neighbor graphs method. In conclusion, the transformer is a highly advanced architecture for dealing with the large-R jets, which contain lots of subjet or tracks. The transformer itself can learn the complex substructure, even better than the subgraph methods.

References

[C.1] Bevilacqua B, Frasca F, Lim D, et al. Equivariant subgraph aggregation networks. 2021.

[C.2] Leman A. The reduction of a graph to canonical form and the algebra which appears therein// 2018. https://api.semanticscholar.org/CorpusID:49579538.

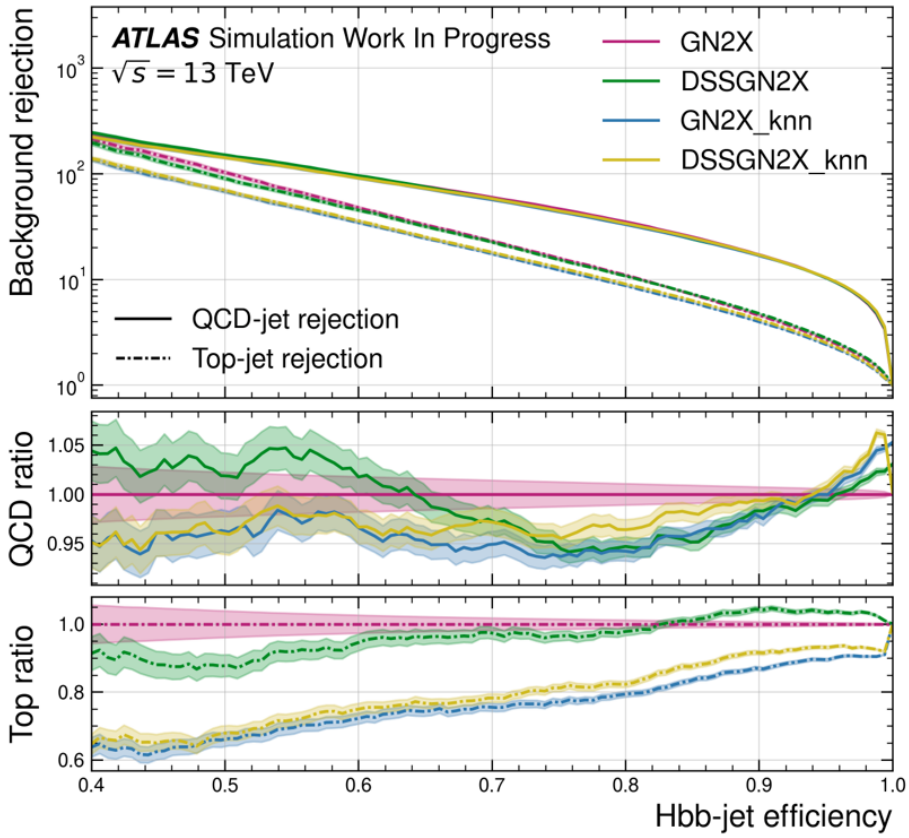


Figure C.4 The background rejection versus signal efficiency for $H \to b\bar{b}$ tagging.

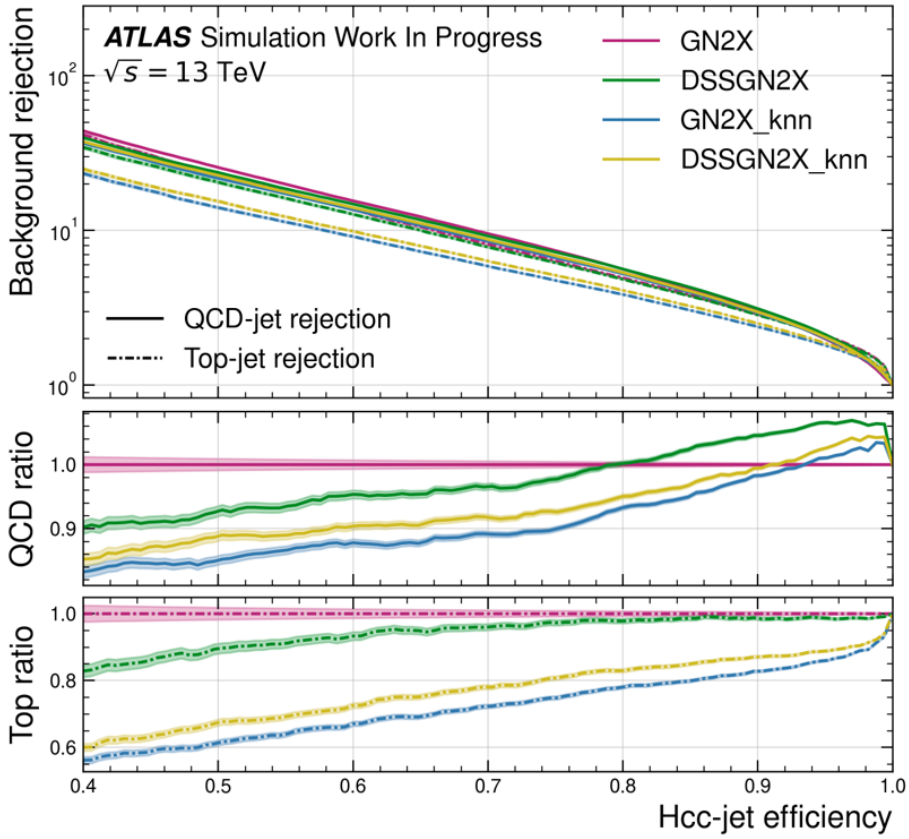


Figure C.5 The background rejection versus signal efficiency for $H \to c\bar{c}$ tagging.