Optimization of event selection for a ttH (H to bb) measurement in the dileptonic channel using data recorded at 13.6 TeV with the CMS Experiment

Bachelor Thesis

von
Kilian Krasenbrink

vorgelegt der
Fakultät für Mathematik, Informatik und Naturwissenschaften der
RWTH Aachen

im Juli 2024

angefertigt im
I. Physikalischen Institut B

bei
Prüfer: Prof. Dr. Lutz Feld
Zweitprüfer: Prof. Dr. Johannes Erdmann

Contents

1 Introduction

The Standard Model of particle physics (SM) has been highly successful at describing known interactions of elementary particles. One of its largest and latest achievements is the prediction of the existence of a new scalar elementary particle, the Higgs boson. Despite its successes, there are indications of physics beyond the SM. Among others, the SM does not include a description of gravity and does not unify all fundamental forces [1]. Furthermore, the hierarchy problem, concerned with the fine-tuning of the low Higgs mass [2] is unexplainable with the intrinsic particle content of the SM. Because high-precision measurements of SM processes allow for direct comparisons to predictions, they can reveal possible deviations and point to the origin of physics beyond the SM. In the context of Higgs physics, measurements of the coupling strengths of the elementary particles to the Higgs field are particularly important. Among all fermions the Higgs boson couples most strongly to the top quark because of its high mass. Measuring the $\mathrm{ttH}(\mathrm{H} \to \mathrm{bb})$ decay is of special interest, as it has the largest branching fraction of ttH production, which provides a direct measure of the top-Higgs coupling. The biggest challenge of this measurement at hadron colliders is imposed by separating a small signal from large dominant background processes, mainly tt production in association with two additional bottom quarks. For sensitive analyses this background has to be simulated to high precision and fine-tuned event selections and multi-variate analysis techniques have to be applied.
In preparation of future $\mathrm{ttH}(\mathrm{H} \to \mathrm{bb})$ analyses, this thesis uses data recorded with the CMS detector during the new running period at the LHC, which started in 2022 with an increased center-of-mass energy of $\sqrt{s} = 13.6\,\mathrm{TeV}$ in proton-proton collisions. This increase, along with other changes, leads to differences in data, such as increased cross sections for ttH and tt production. Consequently, a first look into the new data is necessary to check the agreement of recorded data with simulation. The changes in Run 3 also open up the possibility of new optimal event selection parameters, relevant for $\mathrm{ttH}(\mathrm{H} \to \mathrm{bb})$ analyses. Therefore, this thesis aims to provide both an initial comparison of Run 3 data with simulation and an analysis of potential event selection improvements.

2 Theoretical Foundation

This section describes the unit system used in this thesis and the needed theoretical basis of the Standard Model of particle physics. In addition, the Brout-Englert-Higgs-Mechanism and the ttH(H $\rightarrow$ bb) process are described.

2.1 Unit System

For this thesis the natural unit system is used, defined by setting the reduced Planck constant $\hbar$ and the vacuum speed of light c to unity:

$$\hbar = c = 1. \tag{2.1}$$

Thus the units of energy, momentum and mass are identical and can be given in GeV. Units of electric charge are specified in multiples of the elementary charge e. Lengths and time are given in terms of the metric system and cross sections are stated in barn ($1\text{b} = 10^{-28}\text{m}^2$).

2.2 Standard Model of Particle Physics

The SM [1, 3, 4] is the theoretical framework describing three of the four known fundamental forces through interactions between elementary particles. These forces directly arise by requiring invariance under local gauge transformations.
The full particle content of the SM is shown in Fig. 2.1. Each elementary particle has a corresponding antiparticle, distinguished by opposite signs in all additive quantum numbers, most notably their electric charge. In the following the distinction between a particle and its antiparticle will not be made, unless stated otherwise.

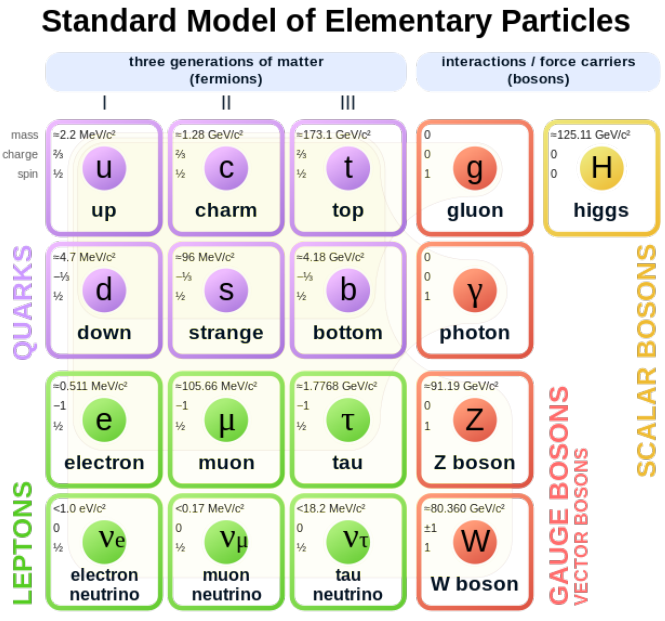


Figure 2.1: Elementary particles of the SM [5] with their mass, charge and spin. The quarks (violet) and leptons (green), which comprise all fermions of the SM, are sorted in three generations. Bosons include the gauge bosons (orange) and the Higgs boson (yellow).

All particles can be categorized as either bosons, particles with integer spin, and fermions, particles with half-integer spin. The gauge bosons mediate the three fundamental forces described in the SM. They include the strong force, the weak force, and the electromagnetic force. The Z^0 and $W^{\pm}$ bosons are the exchange particles of the weak force, while gluons g mediate strong interactions and photons γ the electromagnetic interaction.
Fermions are the constituents of matter, which can be categorized further into quarks and leptons. The distinction between leptons and quarks is due to differing interactions with the fundamental forces. Among all fermions only quarks carry the color charge of the strong force. Therefore strong force interactions of quarks are mediated through gluons, in addition to electromagnetic and weak forces. Unlike the photon, the gluon itself carries the color charge of the force it is mediating, making gluons self-interacting particles. As a result, the effective potential between two particles with colored charge rises linearly with increased distance. This is unlike the potentials of the other forces, which generally weaken with distance. Ultimately, when pulling quarks apart, it is energetically favourable to produce a quark-antiquark pair. Therefore quarks can not exist freely outside of bound states. Bound quark states are called hadrons and their formation through quarks and gluons is called hadronization. Every quark leaving an interaction point (vertex) thus undergoes hadronization, leading to a hadronic shower, called jet.
All six quarks can be arranged into three pairs sorted by mass termed as generations. Each generation consists of one up-type quark and one down-type quark. The distinction of up-type and down-type quarks is due to their electric charge, being $Q = +2/3$ and $Q = -1/3$ respectively.
Further, quarks can decay into less massive quarks through the weak interaction, with the probability of such decays determined by the CKM matrix. For instance, because top quarks decay too quickly to form bound states and the likeliest decays remain within the same generation, they almost always decay immediately into bottom quarks.
Regarding leptons, three of the six carry an electromagnetic charge and an isospin of $I_3 = -1/2$. The remaining leptons, called neutrinos, show a weak isospin of $I_3 = 1/2$ and no electric charge. Each charged lepton (electron e, moun μ, tau τ) is associated with a neutrino (electron neutrino ν_e, muon neutrino ν_μ, tau neutrino ν_τ) due to possessing the same lepton flavor. Like quarks, leptons can be subdivided into three generations as lepton pairs with the same flavour ordered according to their masses.
The Higgs boson is the singular boson of the SM with a spin of zero and is not associated with any of the fundamental forces. It was discovered in July 2012 by the ATLAS and CMS collaborations [6–8]. Its mass is measured to be approximately $m_H = 125\,\text{GeV}$. As of today, all measured Higgs properties agree with SM predictions within uncertainties [9].Given its special importance to the SM and this thesis, the Higgs mechanism is described in more detail below.

2.3 The Higgs Boson

The Z^0 and $W^{\pm}$ bosons are observed to posses relatively high masses, but the basis of the SM, the gauge invariance, requires massless gauge bosons in field free space. The Brout-Englert-Higgs-Mechanism (BEH-Mechanism [10]) solves this apparent contradiction by introducing the Higgs field Φ_H, with the Higgs boson being its respective quantum exitation. The gauge bosons interact with the Higgs field, and an additional Yukawa-type coupling is introduced for fermions. Through this the masses of all SM particles are generated by a process called spontaneous symmetry breaking. This involves the potential $V(\Phi_H)$ of the Higgs field being formulated symmetrically around $\Phi_H = 0$, which ensures local gauge invariance, but the stable ground state lying at a non-zero value Φ_0. Through the coupling of particles to this ground state they experience inertia, making them appear massive. A measure for the interaction intensity between two fields or particles is given by the coupling strength. The experienced inertia of a particle is directly

related to its coupling strength to the Higgs field, meaning that particles with higher mass experience stronger coupling. Particle masses are not predicted by the SM, but need to be measured experimentally. Given the measured masses, a measurement of the coupling strengths of elementary particles with the Higgs field can be compared to SM predictions

2.4 Measurement of ttH(H → bb)

Since the coupling strength to the Higgs is proportional to the mass of the interacting particle, the large top mass offers the strongest of the Higgs field interactions. Because the H → bb decay has the largest branching fraction of 0.58 ± 0.02 [11], probing of ttH(H → bb) decay processes is of particular interest in proton-proton collisions. This decay chain offers direct coupling strength measurements of both top and bottom quarks to the Higgs. However, measuring this Higgs decay channel is challenging due to difficulty in distinguishing it from background processes, particularly tt + jets decays. The tt production cross section at $\sqrt{s} = 13.6\,\text{TeV}$ is $\sigma_{\text{tt}} = 924.6^{+32}_{-40}\,\text{pb}$ [12], which is about three orders of magnitude larger than the ttH production cross section, calculated to be $\sigma_{\text{ttH}} = 0.57^{+0.04}_{-0.06}\,\text{pb}$ [13].
To discriminate signal from background in analyses, event selections are applied. By requiring specific signatures characteristic of the signal's decay channel, such as b-jets or leptons, the backgrounds can be reduced. In this analysis a selection is chosen requiring either two muons ($\mu\mu$ channel) or electrons (ee channel) or one electron and a muon (eμ channel) in all event final states. This selection is referred to as the dileptonic (DL) channel and it was chosen for increased trigger efficiency through the presence of leptons and lower jet counts compared to other channels for more sensitive event reconstruction. But even when applying a sharp selection on the signals final state, called a signal region, a large background contribution can not be removed effectively. The Feynman diagrams in figure 2.2 show that tt + bb processes can mimic the Higgs production channel through gluon radiation. Because the final state particles are identical in both decays, only differences in kinematic variables can be used to discriminate the ttH signal from the tt + bb background. This leads to a large background contribution even in the signal region.

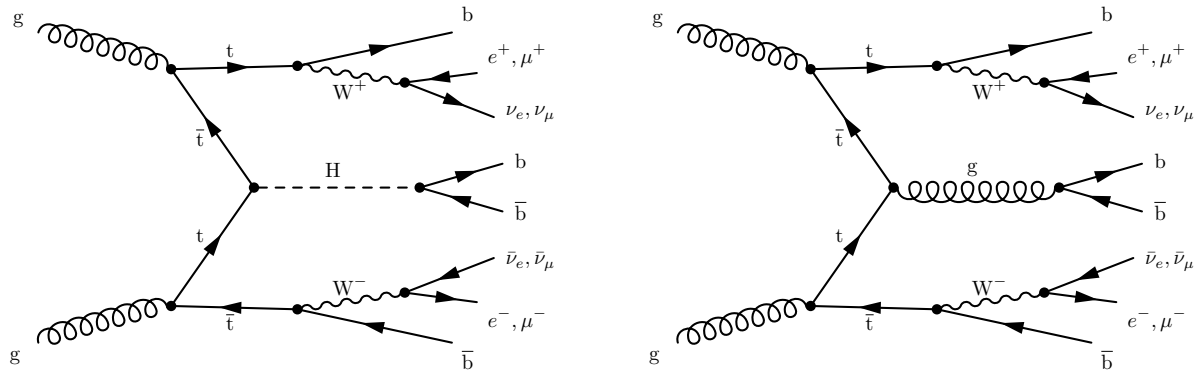


Figure 2.2: Leading-order Feynman diagram illustrating the signal process ttH, H → bb (left) and the major background process tt + bb (right), in the dileptonic channel. Both decay chains result in the same final-state particles, making the signal to background discrimination very challenging.

Sensitivity is further limited by combinatorial background due to the unambiguous reconstruction of the invariant Higgs mass, arising from the presence of more than two b-jets in the final state. Because b-jet identification only has a limited efficiency, a large background contribution also stems from jet misidentification. Therefore sophisticated analysis tools are necessary for

measurements such as ttH production rates [14]. For instance, this includes highly advanced tt + bb event simulation for background modeling and high performing b-jet identification algorithms.

3 Experimental Setup

3.1 Large Hadron Collider

The Large Hadron Collider (LHC) [15] is a circular particle accelerator situated near Geneva at CERN, the European Organization for Nuclear Research. With a circumference of 26.7 km and tunnels constructed at a mean depths of 100 m under the surface it is currently the largest and most powerful particle collider. Two oppositely accelerated beams allow for proton-proton or lead-ion collisions. The center of mass energy $\sqrt{s}$ in proton-proton collisions has increased significantly during the LHCs operating time. So far proton-proton collisions at center-of-mass energies of $\sqrt{s} = 7$, 8 and 13 TeV were conducted during two running periods from 2010-2018. Since April 2022 a center-of-mass energy of 13.6 TeV has been reached, used in the currently ongoing operation period (Run 3).
In addition to the center of mass energy a measure for a particle colliders performance is its instantaneous luminosity L. The integrated luminosity over time L_{int} defines a proportionality constant between the collision event yield N and the integrated cross section σ of the measured process [1].

$$L = \frac{n_B f_{\text{rev}} N_1 N_2}{4\pi\sigma_x\sigma_y} , \qquad L_{\text{int}} = \int L \cdot dt , \qquad N = L_{\text{int}} \cdot \sigma , \tag{3.1}$$

Here n_B corresponds to the number of bunches per beam, f_{rev} to the circulation frequency of the bunches, N_i is the number of particles in each bunch of the i-th beam, while σ_x and σ_y are the Gaussian beam profile width and height respectively. The number of proton-proton interactions per bunch crossing is referred to as pileup.
The LHC is supplied with protons by the injection chain consisting of Linac2, the Proton Synchrotron Booster (PSB), the Proton Synchrotron (PS) and the Super Proton Synchrotron (SPS) in both clockwise and counterclockwise directions. The protons are injected from the SPS into the LHC at around 450 GeV and are stored in bunches at the order of 10^{11} protons each, with a maximum capacity of 2808 bunches. These bunches are accelerated and brought into collision at the four primary beam crossings, where the detectors of the main experiments ALICE [16], ATLAS [17], LHCb [18] and CMS [19] are located. ATLAS and CMS are multi-purpose detectors, designed to study a variety of physics phenomena, wheras ALICE is a heavy-ion detector mostly studying the quark-gluon plasma. LHCb is mainly designed to measure the parameters of CP violation in the interactions of b-hadrons. The data used in this thesis was recorded with the Compact Muon Solenoid (CMS) detector, which is described in more detail in the next section.

3.2 Compact Muon Solenoid

The Compact Muon Solenoid [19–21] is a large multi-purpose particle detector. It was mainly designed for the detection of the Higgs Boson, as well as for precision measurements of the SM. The detector has the shape of a cylinder symmetrically constructed around the beam axis. It measures 22 m in length, spans a diameter of 15 m and weighs 14000 t. The core component of the detector is a superconducting solenoid with an internal diameter of 6 m and a length of 12.5 m, generating a magnetic field of 3.8 T. Inside the solenoid, there is an inner tracker consisting of a silicon pixel and strip tracker, along with a lead tungstate crystal electromagnetic calorimeter (ECAL), and a brass and scintillator hadron calorimeter (HCAL). Gas-ionization Muon Chambers are embedded into the steel flux-return yoke outside the solenoid. All systems consist of layers parallel (barrels) and orthogonal (endcaps) to the beam pipe. Forward calorimeters expand the pseudorapidity (η) coverage. A schematic cross-section of the detector is illustrated in figure 3.1

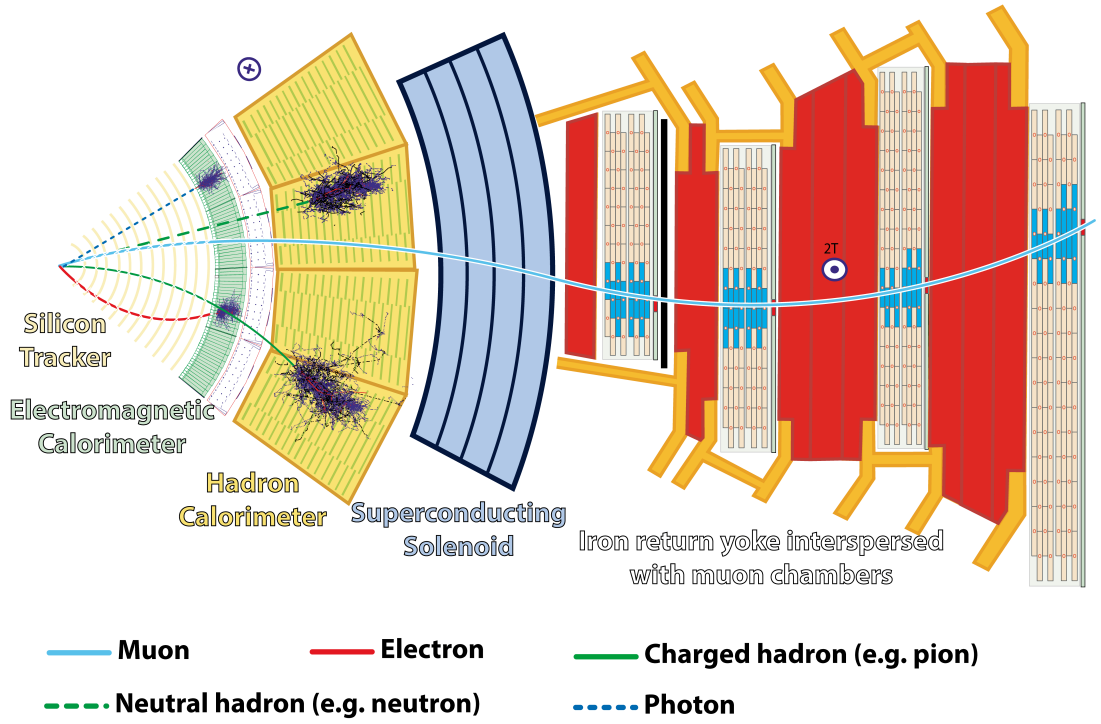


Figure 3.1: Schematic cross-section [22] of the CMS detector showing the different subdetectors and solenoid, including illustrations of different particle signatures.

Coordinate System

The cartesian coordinate system used in most CMS data analyses is centered at the nominal collision point inside the experiment. The x-axis points radially inward, toward the center of the LHC and the y-axis points vertically upward, which fixes the z-axis along the beam direction. Momenta transverse to the beam direction, denoted by p_T, are therefore measured in the x-y-plane. The azimuthal angle Φ is defined inside the x-y-plane and is measured from the x-axis. With the polar angle θ, measured from the z-axis, the pseudorapidity η is defined as:

$$\eta = -\ln\left(\tan\frac{\theta}{2}\right) \tag{3.2}$$

This is often used instead of the polar angle, because the differences of η between two particles produced in the same collision are invariant under Lorentz-boosts in the z-direction in the massless particle limit. To indicate the angular distance between two particles the combined difference ΔR is defined as:

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\Phi)^2} \tag{3.3}$$

Tracking System

The all-silicon tracking system consists of the innermost pixel detector and a surrounding strip detector, comprising an active silicone area of around 200m^2. Overall 1856 pixel modules are installed, totaling 124 million readout channels. During the second long shutdown (LS2) the detector underwent extensive repairs and the innermost layer of the pixel detector was replaced to ensure the best possible performance during Run 3. While the pixel detector covers a pseudorapidity of $|\eta| < 3.0$, the strip detector reaches $|\eta| < 2.5$. The 9.3 million silicon micro-strips

are distributed over 15148 modules and provide a resolution of 20 µm for charged particle hits at right angles. Combining information from both systems makes it possible to measure particle tracks with a momentum of around 100 GeV with a transverse momentum resolution close to 1% and an impact parameter resolution of around 10 µm. This allows for high precision vertex reconstruction.

Electromagnetic Calorimeter

The Electromagnetic Calorimeter (ECAL) outside the inner tracker uses lead tungstate crystal ($PbWO_4$) scintillators to measure the energy of photons and electrons. The scintillators are stimulated by crossing particles and emit light proportional to the energy deposit of the particle. To identify either photons or electrons and measure their momenta and energies, reconstruction algorithms based on the size and shape of the energy deposits are used. Photon and electron distinction is aided by the tracker, as only electrons are expected to have a matching track in the tracker. Further the ECAL provides preshower detection with silicon detectors preceding the crystals. The ECAL also plays an important role in the reconstruction of jets and missing transverse momentum. A total of 75848 crystals allow a pseudorapidity coverage of $|\eta| < 3$.

Hadron Calorimeter

The primary purpose of the hadron calorimeter (HCAL) is to measure the energy of neutral and charged hadrons. It further aids in the reconstruction of jets and missing transverse momentum. The HCAL consists of the hadron barrel (HB), hadron endcap (HE), hadron forward (HF) and hadron outer (HO) calorimeters. The HB and HE soround the ECAL and together span pseudorapidity ranges of up to 3.0. Both consist of alternating brass and scintillating layers. Hadrons interact strongly with the brass layers, generating hadronic showers, which are detected by the scintillators. Outside of the solenoid the HF and HO are located. While the HF consists of steel and quarz fibers, the HO is made out of plastic scintillators. The HF, lying in forward direction, extends the pseudorapidity range up to 5.2.

Muon System

The muon system is located outside of the solenoid and is used for the identification of mouns and the measurement of their momenta. It consists of drift tubes (DTs) in the barrel region ($|\eta| < 1.2$) and cathode strip chambers (CSCs) in the endcap ($0.9 < \eta < 2.4$) which offer good timing and spatial resolution. Additionally, it consists of resistive-plate chambers (RPCs) dedicated to triggering in both barrel and endcap regions and a recently installed gas electron multiplier (GEM) detector in the endcap region, which is specifically optimized for high detection rates.

Trigger System

With a proton-bunch crossing rate of 40 MHz, the amount of data produced in the detector is far too large to be stored in its entirety. Therefore the trigger system is needed to filter events of interest. This task is divided into two successive trigger systems, the programmable hardware based L1 trigger and the software-based High-Level Trigger (HLT). The L1 handles the preselection of events, receiving energy and position information from the calorimeters and muon chambers, lowering event rates to about 110 kHz. When an event is selected by the L1 trigger, it gets passed to the HLT, which uses the full extent of event information for further filtering. In total, the trigger chain reduces event rates to around 5kHz during Run 3.

4 Data Sets, Simulation and Event Reconstruction

In this section, the data sets recorded at CMS and the corresponding triggers as well as Monte Carlo (MC) based background and signal simulations used in this thesis are introduced. Then, the event reconstruction is explained and the most relevant objects for this thesis, such as jets and leptons, are defined.

4.1 Data Sets and Triggers

The CMS data set used in this thesis was recorded in 2022 at a center-of-mass energy of $\sqrt{s} = 13.6\,\text{TeV}$. The data set is further divided into primary data sets (PD), corresponding to different trigger paths. The PDs are not exclusive and might overlap in event content, leading to potential double counting of events. This is solved by orthogonalizing the primary data sets before analyses. The paths to the high-level triggers and to all used PDs are listed in Appx. 1. Because of a water leak in the ECAL in 2022, the corresponding data is split into data taken before (preEE) and after (postEE) the leak. The preEE integrated luminosity with $7.88\,\text{fb}^{-1}$ [23] includes less statistics, as the leak happened in early 2022. Therefore, the postEE data taking period was chosen, which corresponds to an integrated luminosity of $26.67\,\text{fb}^{-1}$ [23].

4.2 Simulation

The Monte-Carlo simulation consists of three major steps before event reconstruction, with the first step being event generation. Here the particle collisions are simulated. Subsequently, particle showering, including processes such as hadronizisation and final state radiations, are simulated with PYTHIA [24]. Next, the detector response is simulated with GEANT4 [25]. This ensures that the simulated samples resemble raw detector data, allowing the same reconstruction algorithm to be applied on MC and data, which is described in section 4.3.
The relevant background for this thesis consists of Drell Yan (DY) processes, diboson (ZZ, WZ, WW) and ttV (V being either W, Z) production, as well as single top production and the dominant tt background described in section 2.4. The tt, single top and diboson processes are generated at next-to-leading order (NLO) accuracy with the event generator POWHEG, while DY and ttV events are generated with MADGRAPH5_AMC@NLO at NLO accuracy. Additionally the ttH signal is generated at NLO accuracy with POWHEG. A list of all paths to the used simulated signal and background data sets can be found in Appx. 2 together with their respective process cross sections.
To compare simulation to the recorded data, the yield of simulated events has to be scaled to the recorded integrated luminosity L_{data}. Thus each event is weighted according to

$$\tau_{lumi} = \frac{\sigma \cdot \mathrm{L}_{data}}{\mathrm{N}_{sim}}, \tag{4.1}$$

where σ is the cross section of the simulated process and N_{sim} is the number of simulated events. Because pileup in MC and data may vary, additional weights are applied to match pileup distributions. Since top quark p_T distributions are observed to be softer than in simulation, events are reweighted dependent on the top p_T [26, 27]. In this thesis the $\sqrt{s} = 13\,\text{TeV}$ recommendations for top p_T reweighting are applied, because at the time of writing calculations for $\sqrt{s} = 13.6\,\text{TeV}$ have not been released yet. Further, differences in efficiencies for simulation and data have to be considered. Specifically reweighting is applied on simulation for lepton identification, as well as muon isolation, electron reconstruction and b-tagging. Such weights are evaluated on an event-by-event basis and are called scale factors (SF).

4.3 Object Definitions

Objects such as particles, jets and missing transverse momenta need to be reconstructed based on the measured detector signatures. The reconstruction is performed with the Particle Flow (PF) algorithm [28, 29], which is described below, together with the most relevant object definitions for subsequent analyses. Further, complementary identification criteria (IDs) for improved object reconstruction are introduced as well as jet-flavour identification algorithms, necessary for the identification of jets initiated by bottom quarks.

4.3.1 Particle Flow Algorithm

The PF algorithm is applied to each recorded event to identify particles and reconstruct their kinematic variables based on the full detector readout. Information from each subdetector is taken into account and the individual particle signatures are linked to form blocks, which correspond to particle candidates. Then, particle types are assigned to the identified blocks based on their composition. The particle identification is designed to assign particle identities in the expected order of decreasing discriminability of signatures. First blocks are assigned to muons, followed by electrons. Next, charged hadrons are assigned and lastly all remaining blocks are identified as either photons or neutral hadrons. After each categorization step, the identified blocks are removed for the subsequent assignments.

4.3.2 Muons

Because muons barely deposit energy inside the calorimeters, they are mainly identified through tracks reconstructed in the inner tracker and the muon chambers. The track reconstruction is done seperately in both subsystems. Muon candidates are then identified as so-called "global muons", if the reconstructed trajectories in both subsystems are compatible. For events with muons to be considered in this thesis, muon candidates have to be assigned as "global muons" and fulfill further cut-based ID requirements. Those include conditions on the goodness of the track fit based on a χ^2 test and the suppression of misidentification due to hadron punch-through into the muon chambers by requiring a minimum of stations matched with the muon candidate. Additionally there are cuts on matched tracker and muon chamber hits to reduce the number of muons from in-flight decays. To minimize miscounting from pileup, cuts are applied on the transversal (d_0) and longitudinal (d_z) distance between the reconstructed track and the primary vertex. Finally an isolation (ISO) requirement is imposed, which limits the sum of hadron and photon p_T within a cone of $\Delta R < 0.4$ around the lepton, divided by the lepton p_T, to reject muons produced in hadronic decays. Concrete values for these cuts are given in table A3.1 in Appdx. 3.

4.3.3 Electrons

Electrons leave tracks in the inner tracking system and enter the ECAL, where the resulting electron shower is absorbed completely. Electrons are therefore reconstructed by matching inner tracks with energy deposits in the ECAL. Bremsstrahlung emitted by the electrons passing the inner tracking system is also considered in the reconstruction. Because of inefficiencies in the region between the barrel and endcap section of the ECAL, electrons within $1.4442 < \eta < 1.5660$ are not taken into account. To lower misidentification rates electrons have to pass additional identification and isolation requirements. A new electron identification algorithm based on multivariate-analysis (MVA) techniques has been introduced in Run 3. In this thesis the MVA based

criteria including isolation requirements, with an efficiency of 80% are used. Concrete information on the algorithm designed with the MVA approach can be found in references [30] and [31].

4.3.4 Jets

To identify quarks and gluons, their respective hadronic showers have to be reconstructed by clustering particle candidates. In this thesis the anti-k_t algorithm [32] is employed with a distance parameter of 0.4 for jet clustering. Because contributions from pileup can bias the jet clustering, these have to be removed before clustering is applied. While charged hadron subtraction (CHS) [33] was the standard method for several years, pileup per particle identification (PUPPI) [33] is newly recommended for Run 3 analyses and is therefore employed in this thesis. While the CHS algorithm aims to remove charged hadrons unambiguously assigned to pileup interactions, PUPPI adds an estimation of neutral hadron contributions by weighting neutral hadron energies based on their probability of originating from pileup vertices. Thus the utilization of PUPPI is expected to make the reconstruction of jets and missing transverse momentum (defined in the next section) more robust against pileup.
For improved jet reconstruction, ID requirements are applied. Only jets with $|\eta| < 2.4$ are considered in this thesis and the corresponding jet IDs can be found in table A3.2 in Appdx. 3.
Certain detector regions produce anomalously high or low jet rates. To remove events with jets reconstructed in these regions, jet veto maps are applied on both simulation and data. Additionally, jets within $\Delta R < 0.4$ of the nearest lepton are vetoed to avoid lepton and jet overlap. Jet energy scale corrections are then applied on all accepted jets.
In the context of this analysis the identification of jets stemming from bottom quarks (so-called b-tagging) is of special importance. The standard in high-energy physics analyses is the utilization of deep learning techniques for jet tagging. Over the years, advancements in deep learning techniques have significantly enhanced the efficiency of b-tagging and offer improved sensitivity for Run 3 analyses. Consequently, it is interesting to directly compare the algorithms available for Run 3 in the context of ttH analyses. These algorithms include namely DeepJet [34], ParticleNet [35] and RobustParTAK4 [36, 37]. The oldest of these is the convolutional neural network DeepJet, which was the standard tool for Run 2 analyses. It is the first neural network employed in jet flavor identification, which was capable of using the entire information of all jet constituents, without prior filtering and quality checks. Through this increase in usable jet information, DeepJet achieved better sensitivities than its predecessors. The ParticleNet neural-network was the first jet-flavour identifier based on treating jets as unordered sets of particles, referred to as particle clouds [35]. Previous algorthims, like DeepJet, organized jet constituents in structures such as trees or sequences. This manually forced order on generally unordered shower particles impaired the performance of previous algorithms. The RobustParTAK4 is newly available for Run 3. The algorithm is based on a transformer network, specifically called "particle transformer" [38], trained with adversarial attacks for improved robustness against simulation mismodeling and increased tagging performance.

4.3.5 Missing Transverse Momentum

As the protons' transverse momenta are negligible before collisions, energy conservation in the transverse plane imposes that the vectorial sum of transverse momenta of all particles must be zero within detector resolution. But because neutrinos are neutral and only interact weekly, they can not be directly detected and leave an imbalance in the $\vec{p}_{\mathrm{T}}$ sum. This imbalance can be quantified through the missing transverse momentum defined as follows:

$$p_{\mathrm{T}}^{\mathrm{miss}} = \left|\vec{p}_{\mathrm{T}}^{\mathrm{miss}}\right| = \left| -\sum_{i}^{N} \vec{p}_{\mathrm{T}_i} \right| \ . \tag{4.2}$$

If all detectable particles are measured correctly, $p_{\mathrm{T}}^{\mathrm{miss}}$ corresponds to the momentum of undetectable particles. This includes potential unknown particles, or in the specific case of the SM, neutrinos. Therefore it is important to accurately reconstruct $p_{\mathrm{T}}^{\mathrm{miss}}$ for event selection. Though, the direct assignment of the missing transverse momentum to undetectable particles is limited by mismeasurements. To reduce the impact of pileup on $p_{\mathrm{T}}^{\mathrm{miss}}$, the use of PUPPI-based $p_{\mathrm{T}}^{\mathrm{miss}}$ calculations is newly recommended in Run 3 analyses and employed in this thesis. So-called event flags are applied on MC and data to remove events with anomalously high $p_{\mathrm{T}}^{\mathrm{miss}}$, caused by faulty detector parts or poor reconstruction.

5 Event Selection and Systematic Uncertainties

In this chapter the event selections for analyses in chapter 6 and 7 are introduced and the purpose of each cut is explained. Lastly, all systematic uncertainties considered in this thesis are described.

5.1 Nominal Selections

The ratio of signal events to the SM background can be reduced, by selecting events which include specific particle content and lie within a certain phase space. Specifically, to select dileptonic $\mathrm{ttH(H \to bb)}$ events, the following cuts are applied:

1. at least two jets (see Sec. 4.3.4)
2. at least one b-tagged jet
3. exactly two oppositely-charged leptons (either $\mathrm{e^{\pm}e^{\mp}}$, $\mu^{\pm}\mu^{\mp}$, or $\mathrm{e^{\pm}}\mu^{\mp}$)
4. both leptons within $|\eta| < 2.4$
5. a leading lepton with $p_{\mathrm{T}} > 25\,\mathrm{GeV}$ and a subleading lepton with $p_{\mathrm{T}} > 15\,\mathrm{GeV}$
6. invariant mass of ee and $\mu\mu$ pairs $m_{\mathrm{ee}/\mu\mu} < 76\,\mathrm{GeV}$ or $m_{\mathrm{ee}/\mu\mu} > 106\,\mathrm{GeV}$
7. $m_{\mathrm{ee}/\mu\mu} > 20\,\mathrm{GeV}$
8. $p_{\mathrm{T}}^{\mathrm{miss}} > 40\,\mathrm{GeV}$
9. minimum jet p_{T} of 30 GeV
10. maximum jet $|\eta|$ of 2.4

This selection is chosen as an equivalent to a baseline selection employed in $\mathrm{ttH(H \to bb)}$ analyses [14], which functions as a superordinate selection including all control and signal regions. It is referred to as the 2j1b selection in the following. Here, 2j1b is the abbreviation for requiring at least two jets (2) and a minimum of one b-tagged jet per event (1). The criteria 1-3 specifically aim to restrict the selection to dileptonic channel final states. To offer insight into an event selection resemblant of signal regions, an additional event selection, analogously abbreviated as the 3j3b selection, is considered. This selection requires at least three b-jets, which also directly implies at least three jets in an event, while adhering to the same cuts (3-10). The looser cut of only requiring three b-jets, while the $\mathrm{ttH(H \to bb)}$ final state includes a minimum of four b-jets, is chosen to avoid a significant decrease in event statistics. Both the 2j1b and the 3j3b selection with the cuts listed above are referred to as nominal selections. While the nominal 2j1b selection is used to provide a first look into the Run 3 data in chapter 6, chapter 7 provides data-simulation comparisons based on an optimized 3j3b selection.
Cut (2) is dependent on the employed b-tagging algorithm and the chosen working point. The working point functions as a binary classification threshold on the b-tagger output, above which a jet is regarded as b-tagged. The lepton η cut (4) is applied to restrict the range to the extend of the muon chambers and the tracker acceptance. All lepton p_{T} cuts (5) are chosen to exceed the trigger thresholds. Further, DY background events largely contribute around the Z mass ($m_{\mathrm{Z}} = 91\,\mathrm{GeV}$), and are expected to posses low missing transverse momenta ($p_{\mathrm{T}}^{\mathrm{miss}} < 40\,\mathrm{GeV}$). The corresponding cuts aiming to suppress this background are the requirements referred to as the Z window (6) and the $p_{\mathrm{T}}^{\mathrm{miss}}$ cut (8). Additionally, DY events and heavy flavour resonances in the low invariant mass range ($m_{\mathrm{ee}/\mu\mu} < 20\,\mathrm{GeV}$) are excluded (7). The jet p_{T} cut (9) is generally

applied due to reduced reconstruction efficiencies and increased energy uncertainties for low jet p_T. Lastly, the jet cut on η (10) is chosen in accordance with the jet ID requirements.

5.2 Systematic Uncertainties

In the subsequent data-simulation comparisons, several systematic uncertainties on the MC-samples are accounted for. Since this thesis offers an early look into Run 3 data, not all relevant corrections and uncertainties have been fully integrated into the analysis framework[1] at the time of writing. Jet energy scale corrections (JES), b-tagging scale factors and top p_T reweighting are applied, but their corresponding uncertainties are not included in the following. Specifically jet energy corrections are expected to be dominant sources of uncertainties in many distributions presented in chapter 6. The given uncertainties listed below therefore only serve as a first approximation.

Lepton Reconstruction

The lepton reconstruction scale factors applied on simulation mentioned in section 4.2 are affected by systematic uncertainties. To derive this uncertainty, the individual scale factors for electron identification and reconstruction as well as muon identification and isolation are individually varied by their uncertainties. The differences in simulated counts between the analysis carried out with the nominal SF and the shifted values is regarded as the uncertainty on the MC-counts.

Pileup

Pileup reweighting is based on the total inelastic proton-proton cross section of 69.2 mb [40]. The corresponding uncertainty is derived by applying alternative weights derived by shifting the proton-proton cross sections by $\pm 4.6\%$.

Theoretical Cross Sections

Because event counts are expected to scale linearly with the calculated cross sections for the individual processes, counts of each process are multiplied by the respective cross section uncertainty in percent. This amount is then regarded as an uncertainty stemming from the cross section calculation. The total cross section uncertainties for all processes are given in table 5.1

Table 5.1: Symmetric cross section uncertainties used for the modeled background processes. The total up and down uncertainties from reference [41] were used to calculate the mean symmetric uncertainties given here.

Background Process	cross section uncertainty
tt	4.8%
Drell Yan	2%
single top	3.1%
ttZ	8.9%
ttW	13.5%
WW	4%
ZZ	5.6%
WZ	6%

[1] The analysis framework PocketCoffea [39] was employed in this thesis

6 First Run 3 Data-Simulation Comparisons

This chapter presents a first look into Run 3 data-simulation comparisons. The analysis only includes events passing the nominal 2j1b event selection with the RobustParTAK4 b-tagging algorithm applied at the medium working point. The medium working point corresponds to an efficiency of around 80% on real b-jets, when applied to a tt sample [42].
The subfigures displayed in figure 6.1-6.3 consist of an upper plot, showing the background expectation from simulation and data event counts with the simulated ttH signal overlaid as a red line, and a lower plot showing the ratio of the data counts divided by the SM prediction without including the signal. In the ratio plot, simulated background counts serve as a nominal axis around which the transformed systematic uncertainties are depicted. Thus, the plot of the data-background ratio allows for direct comparisons of the extent to which the simulation reflects the shape of the data and for the quality of the simulation normalization. All simulated background processes listed in section A2.1 are considered. The top-pair production background is further categorized into events with at least one additional jet including b-hadrons (ttB) or c-hadrons (ttC), and events of all other processes (ttLF). The uncertainty bands on the MC simulation include the statistical Poisson uncertainty (stat) and systematic uncertainties as described in Section 5.2 (sys), while the theoretical cross-section uncertainty (theo) is shown separately. Error bars on data represent statistical uncertainties.

6.1 Data-Simulation Comparison Analysis

6.1.1 Number of jets/b-jets

Figure 6.1a shows the number of events which passed the nominal 2j1b selection as a function of the number of reconstructed jets (N_{jets}) in each event, while figure 6.1b shows the same for only b-tagged jets ($N_{\text{b-jets}}$). Because the signal region applies higher jet and b-jet multiplicity cuts, trends in baseline distributions directly influence signal region normalization. A downward trend in the shape of the N_{jets} distribution in figure 6.1a suggests that the simulation overestimates the number of events with higher jet multiplicity. In figure 6.1b specifically, the agreement between data and simulation seems to improve with higher b-jet multiplicity. While the the data counts are around 10% smaller than simulation in the first two bins, for higher b-jet multiplicity the data and MC agreement lies generally within the expected uncertainties. Because the first two bins include most of the selected events, this normalization error is observed in all plots of the 2j1b selection in figures 6.1-6.3. Consequently, it is expected that a selection requiring more b-jets in an event would agree better with the normalisation of the simulation. This is confirmed in chapter 7, where plots of the 3j3b selection show significant agreement improvements in all variables. Considering that major contributing uncertainties of jet energy corrections (JEC) are not included, the modeling of N_{jets} and $N_{\text{b-jets}}$ can already be considered good.

6.1.2 b-tag score

The b-tag score corresponds to the output value of the b-tagging algorithm, in this case the RobustParTAK4 algorithm, which is used for the binary jet classification into b-jets. In distribution 6.1c the count of all jets from all events passing the selection are plotted against their b-tag scores. A sudden decrease in jet counts under the working point threshold is expected, due to the cut on the number of b-jets in an event. This discontinuity can be observed at a b-tag score of around 0.45, which corresponds to the medium working point. The distribution is well modeled, but in the two lowest bins there is an excess of simulated events. It is important to note, that these lower bins include a large portion of all event statistics. This suggests that

the normalization error seen in all plots, is correlated with events including jets with low b-tag scores. Because more such events are filtered out when increasing the b-jet multiplicity requirement, this seems to be in agreement with the normalization improvements observed for the 3j3b selection in chapter 7.

6.1.3 Scalar Sum of Transverse Jet Momenta

The H_T of an event is defined as the scalar sum of all jets per event and is an important variable for monitoring jet activity in simulation. Figure 6.1*d* shows the H_T distribution for nominal 2j1b events. The distribution highlights the overall trend of simulation normalization being larger than that observed in data. Simulation is seen to be around 12% larger than the data. The shape of data has a slight downward trend in comparison to the simulation, which is more pronounced in lower bins. This might be caused by missing jet energy resolution (JER) corrections and outdated jet energy scale (JES) corrections. Changes in top p_T reweighting recommendations for $\sqrt{s} = 13.6\,\text{TeV}$ collision data might also improve the general shape. Without the normalization discrepancy, the distribution is already well modeled.

6.1.4 Leading b-jet p_T and η

The object with the largest (second largest) p_T in an event is referred to as leading (subleading). The distribution of the transverse momentum $p_T^{b_1}$ of the leading b-jet in the nominal 2j1b selection is shown in figure 6.2a, while its pseudorapidity η^{b_1} distribution is shown in figure 6.2b. Because it is crucial for $\text{ttH}(\text{H} \to \text{bb})$ analysis to reconstruct the Higgs final state, good modeling and reconstruction of kinematic variables related to b-jets is required. Specifically, b-jet kinematic variables are important inputs for signal discriminators.
As previously described for H_T, the $p_T^{b_1}$ distribution shows a slight downward trend for increasing p_T in the ratio plot, which is more pronounced for low p_T. Since H_T is the scalar sum of all jets in an event, this mismodeling of b-jet p_T directly propagates to H_T. Thus, the same arguments for possible improvements and increased uncertainties through updated JECs and top p_T reweighting apply as for the H_T distribution.
Further, there are discontinueties at around $\eta^{b_1} = -1.8$ and $\eta^{b_1} = 1.4$, which can not be explained. Decreasing the MC simulation by 12% would already show good agreement of simulation with data in both distributions. The η and p_T distributions of the subleading b-jet can be found in Appdx. 4 in figures A4.1a and A4.1b. For completeness, the Φ component distribution of the leading b-jet can be found in figure A4.1e, which is observed to be modeled well.

6.1.5 Leading Lepton p_T and η

The distribution of the leading lepton transverse momentum $p_T^{\ell_1}$ is shown in figure 6.2c, while the corresponding pseudorapidity η^{ℓ_1} distributions is plotted in figure 6.2d. A decrease of the simulation normalization would again yield good agreement in accordance with uncertainties. A slight dip in the ratio plot of $p_T^{\ell_1}$ can be observed in the region of around $180\,\text{GeV} < p_T^{\ell_1} < 280\,GeV$, though the reason for this simulation excess is at the moment, and with the currently implemented corrections, unknown. The η^{ℓ_1} ratio shows a slight parabolic shape, probably due to missing efficiency calibrations, such as trigger efficiency corrections. The sudden decrease in event counts around $|\eta^{\ell_1}| = 1.5$ are due to the cuts mentioned in section 4.3, which exclude the areas between the barrel and endcap regions of the ECAL. Additional plots of the subleading lepton p_T and η distributions can be found in Appdx. 4 in figures A4.1c and A4.1d respectively.

6.1.6 Missing Transverse Momentum and the $\Phi(p_{\mathrm{T}}^{\mathrm{miss}})$ Component

Missing transverse momentum is generally hard to model, because all kinematic variables of an event and the corresponding corrections, directly influence $p_{\mathrm{T}}^{\mathrm{miss}}$ (figure 6.3a) and $\Phi(p_{\mathrm{T}}^{\mathrm{miss}})$ (figure 6.3b). Having to account for pileup contributions and mismeasurements in data additionally complicates simulation. When the normalization is fixed and all JEC and updated top p_{T} corrections are applied, the plots shown here suggest good data-simulation agreement. Nonetheless, the $p_{\mathrm{T}}^{\mathrm{miss}}$ distribution displays an upward trend in the ratio plot for increasing p_{T}. This might be explained by rising amounts of mismeasurements for particles/jets with large momenta.
A slight substructure in the ratio plot of the $\Phi(p_{\mathrm{T}}^{\mathrm{miss}})$ distribution can be observed, but the direct causes are unknown. Considering the challenges of $p_{\mathrm{T}}^{\mathrm{miss}}$ and $\Phi(p_{\mathrm{T}}^{\mathrm{miss}})$ reconstruction, both distributions are remarkably well modelled for an early analysis.

6.1.7 Number of Vertices

Figure 6.3c depicts event counts as a function of the measured number of primary vertices in a collision (N_{PV}). The visible peak of the distribution can be observed around the value $N_{\mathrm{PV}} = 33$. In analysis of Run 2 data at $\sqrt{s} = 13\,\mathrm{TeV}$ this peak was observed at a lower value of $N_{\mathrm{PV}} = 21$. This is expected, as the amount of pileup rose with the increased center of mass energy in Run 3. Again, increasing the normalization by 10% would already yield a well described distribution, as the parabolic shape of the ratio plot would fall within the systematic uncertainty band. Since this variable is very hard to model, it is remarkable that the data-MC agreement is already comparable to reprocessed Run 2 data, not considering the normalization error.

6.1.8 Invariant Mass of Bottom Quark Pairs

Figure 6.3d shows the invariant mass distribution $\mathrm{m}_{\mathrm{bb}}^{\Delta\mathrm{R}_{\mathrm{min}}}$ of the two b-jets closest in angle, defined by possessing the minimum $\Delta\mathrm{R}$ (see equation 3.3) in an event. This is referred to as the $\Delta\mathrm{R}_{min}$ system. The distribution is chosen because the two b-jets from the Higgs boson decay are expected to be closer in angle than the b-jets from the top and anti-top decays. This grants high signal sensitivity, by directly probing the Higgs mass resonance. The visible Higgs resonance peak of the simulated $\mathrm{ttH}(\mathrm{H} \rightarrow \mathrm{bb})$ signal distribution demonstrates this. The resonance peak lies slightly below the measured Higgs mass of 125 GeV [6]. A cause for this might be neutrinos produced in the bottom quark showers, which carry momentum missing in the reconstructed jet p_{T}. In addition, the signal appears smeared around the peak, which is expected because $\Delta\mathrm{R}_{min}$ systems include false combinations not corresponding to the Higgs decay, as the simple $\Delta\mathrm{R}_{min}$ criterion has a limited accuracy.
The normalization error can also be seen here, but the ratio plot shows, that the shape of the data distribution is well modelled by the simulation. Because the invariant mass $\mathrm{m}_{\mathrm{bb}}^{\Delta\mathrm{R}_{\mathrm{min}}}$ is a high level variable, the agreement between data and simulation is strikingly good for an early analysis.
For the same reason as the $\Delta\mathrm{R}_{min}$ system, the highest p_{T} system, corresponding to the pair of b-jets with the largest vectorial sum of transverse momenta in an event, are candidates for originating from the Higgs decay. The distribution of the invariant mass for this system $\mathrm{m}_{\mathrm{bb}}^{p_{\mathrm{T}}}$ can be found in Appdx. 4 in figure A4.1f. In the nominal 2j1b selection both distributions look almost the same and no significant shape differences can be pointed out.

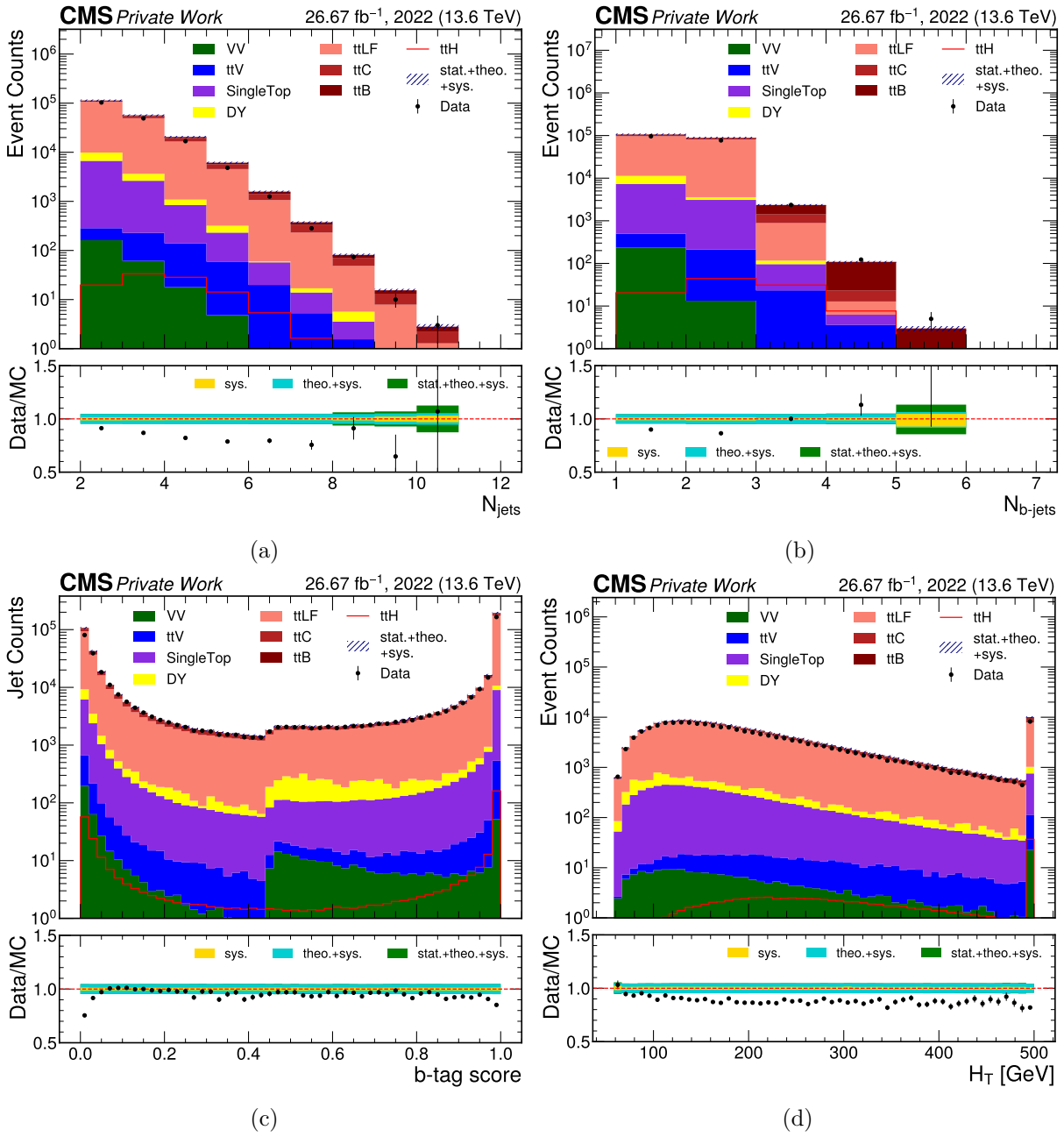
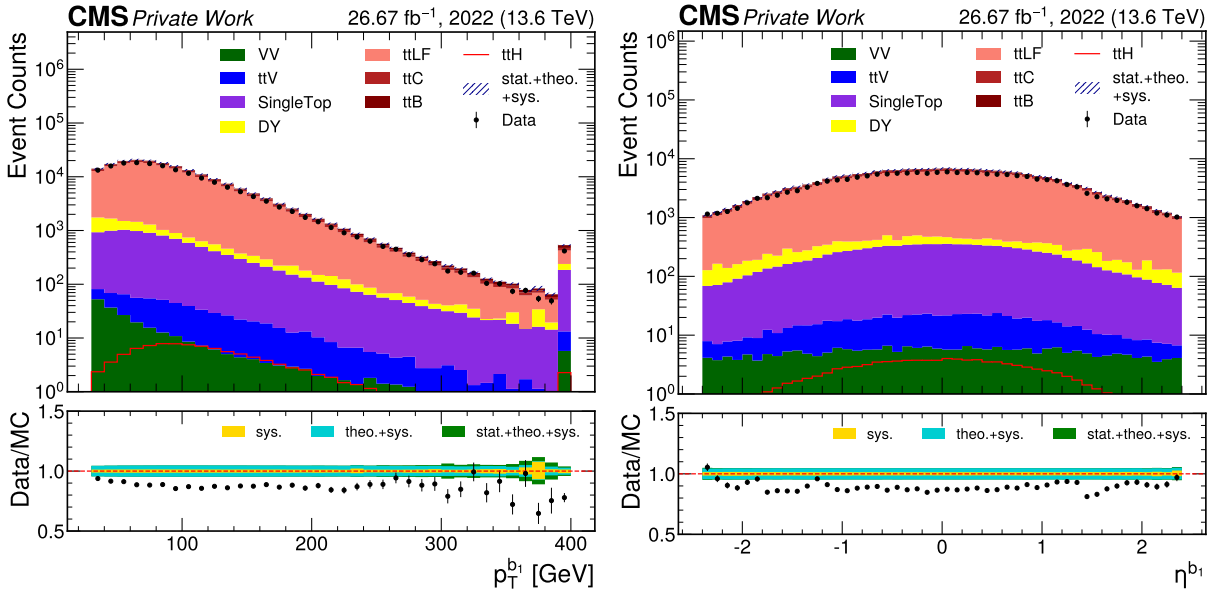
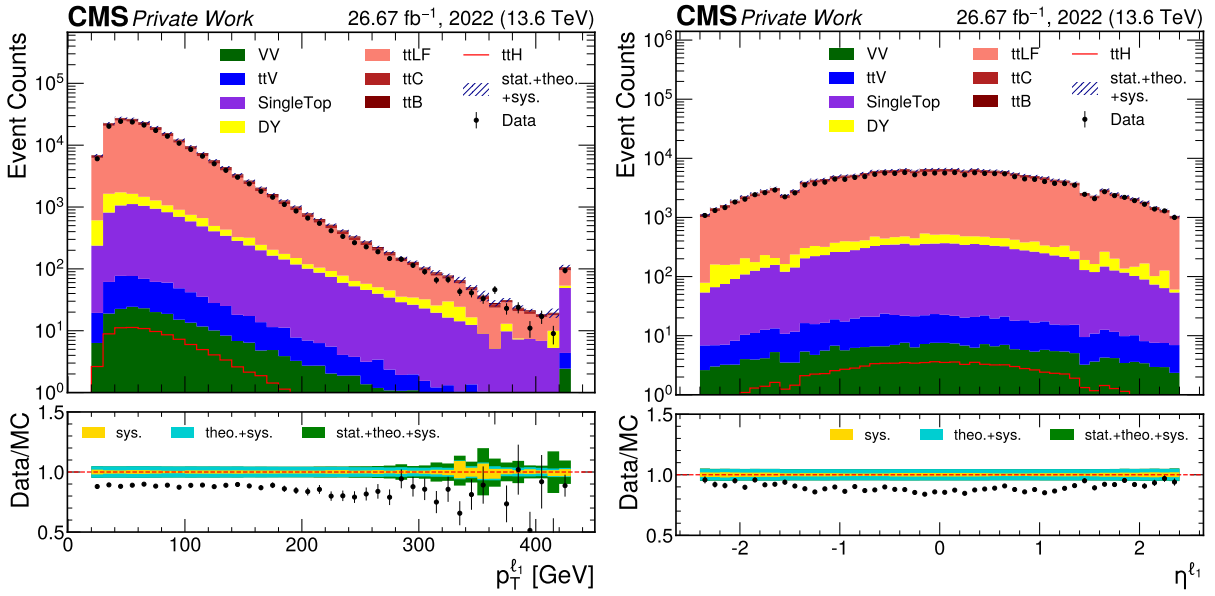


Figure 6.1: Plot (a)/(b) shows the event count as a function of the number of jets/b-jets in an event. Plot (c) shows the count of all jets in the selection as a function of their individual b-tagging scores. In (d), the event count is plotted against the sum of all jet transverse momenta H_T. All plots consider events which pass the nominal 2j1b event selection using the RobustParTAK4 algorithm for b-tagging at the medium working point. The outermost bins include overflow counts.



(a) (b)



(c) (d)

Figure 6.2: Figures (a) and (c) display transverse momentum distributions, while (b) and (d) depict the corresponding pseudorapidity of all selected events for the leading b-jet and leading lepton respectively. Events need to pass the nominal 2j1b event selection using the RobustParTAK4 algorithm for b-tagging at the medium working point. The outermost bins include overflow counts.

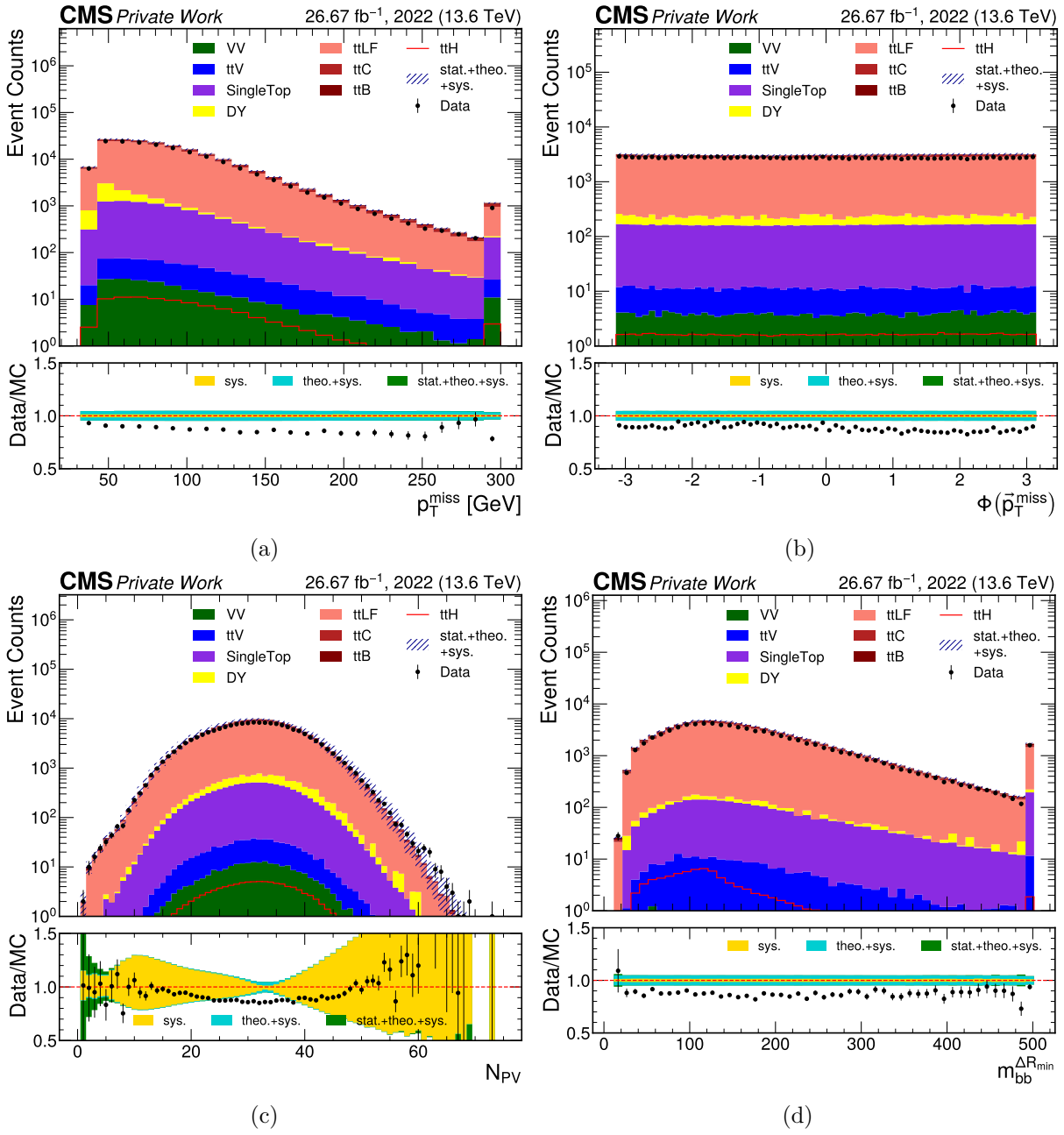


Figure 6.3: Figure (a) shows the missing transverse momentum distribution, while (b) displays the corresponding Φ component. The number of primary vertices N_{PV} is shown for all events in figure (c), while figure (d) depicts the invariant mass distribution $m_{bb}^{\Delta R_{min}}$ of the minimum ΔR system. All plots consider events which pass the nominal 2j1b event selection using the RobustParTAK4 algorithm for b-tagging at the medium working point. The outermost bins include overflow counts.

7 Event Selection Optimization

The cuts of an event selection determine signal and background efficiencies. Thus, the optimization is based on changing some of the nominal selection requirements (See Sec. 5.1), to probe for potential improvements in signal sensitivity. Specifically, the discovery significance for the ttH(H $\rightarrow$ bb) production process is used as a measure of sensitivity, which is defined in the following section. After explaining which specific event requirements are candidates for change, each one is analysed. Finally, an optimized 3j3b selection is proposed and some of the distributions discussed in chapter 7 are revisited.

7.1 Discovery Significance

A statistical measure for the signal to background discrimination and therefore the efficiency of an event selection is given by the discovery significance $\mathcal{Z}$ [43]. It is defined through the p-value, which is the probability of obtaining an observation under the background only hypothesis. The significance is then defined through

$$p = \int_{\mathcal{Z}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx. \tag{7.1}$$

Thus, $\mathcal{Z}$ corresponds to the number of standard deviations at which a one-sided tail of a normal distribution would yield an area equal to p. In the context of signal efficiency optimizations, a small p-value, or equivalently a large significance $\mathcal{Z}$, is desirable. This would indicate a low probability that the measurement only originates from background fluctuations, but rather from an additional signal. In the specific case of this thesis, the signal is the ttH(H $\rightarrow$ bb) production. In the limit of a large simulated background sample

$$\tau = \frac{\mathrm{L}_{MC}}{\mathrm{L}_{data}} \rightarrow \infty, \tag{7.2}$$

with the integrated simulated luminosity L_{MC}, an estimator from simulated data for $\mathcal{Z}$ can be derived as [43]

$$\mathcal{Z} = \sqrt{2\left((s+b)\ln\left(1+\frac{s}{b}\right)-s\right)}. \tag{7.3}$$

Here, s is the simulated signal yield, while the total background yield b is assumed to be known to high precision ($\tau \rightarrow \infty$). Further, for $s \ll b$ this estimator is approximated by

$$\mathcal{Z} \approx \frac{s}{\sqrt{b}}. \tag{7.4}$$

This approximation demonstrates that applying looser cuts can improve signal significance, as long as background and signal efficiencies are increased to the same extent. Conversely, tightening cuts may be desirable, if signal efficiencies decrease less than background.

7.2 Potential for Optimization

In this section all cuts defined in section 5.1 are examined and potential changes are highlighted.

Number of b-jets

While applying the nominal jet and b-jet multiplicity requirements, it is interesting to analyse the impact on the significance estimator of using the different b-tagging algorithms listed in

section 4.3.4.
There are three standard working points for each algorithm recommended for Run 3, namely the loose, medium and tight working points. Tighter working point requirements correspond to an increased threshold on the b-tagger output. This decreases b-tagging mistake rates but comes at the cost of lowered b-jet reconstruction efficiency. In particular, the loose (tight) working point corresponds to an efficiency of around 93% (65%) on real b-jets, when applied to a $t\bar{t}$ sample [42]. In section 7.3.1 the effects on significance for different b-tagging algorithms and working points, employed on the nominal selections, are investigated.

Range of η

The lepton η range is chosen according to the maximum coverage of the muon chambers. While muon reconstruction depends on the extent of the muon chambers, electrons can be identified across the full pseudorapidity range of the tracking system ($|\eta| < 2.5$), because the ECAL covers the $|\eta| < 3$ range. Therefore, increasing the electron η acceptance to $|\eta| < 2.5$ introduces more events into the selection, which could enhance the signal significance. This will be tested in section 7.3.2.
The jet η range can not be increased further due to the jet ID requirements.

Cuts on Transverse Momenta

The lepton p_T requirements are chosen to exceed the p_T thresholds of the applied triggers and can therefore not be lowered any further. The jet p_T cut is generally applied due to reduced reconstruction efficiencies for low jet p_T. However, since the b-tagging is calibrated down to a minimum of 20 GeV, reducing the minimum jet p_T could increase statistics and potentially improve the significance. This is investigated in section 7.3.3

DY Background Removal

The choice of the Z window is arbitrary and should be made with regard to the significance. The same argument applies for the p_T^{miss} cut. Because both requirements target DY background, they are highly correlated. Consequently, their impact on significance is analysed together in section 7.3.4.

7.3 Cut Optimization

7.3.1 b-jet Identification

The first significance optimization discussed here focuses on the comparison of different b-tagging algorithms and recommended working points. The $\mathcal{Z}$ estimator values according to equation 7.4 can be found in tables 7.1a and 7.1b, respectively for the 2j1b and 3j3b selections. To validate equation 7.4, these values were recalculated using formula 7.3, which can be found in Appdx. 6 in table A6.1.

2j1b	loose	medium	tight
DeepJet	0.224	0.233	0.237
ParticleNet	0.225	0.233	0.237
RobustParTAK4	0.225	0.233	0.237

(a)

3j3b	loose	medium	tight
DeepJet	0.504	0.815	0.892
ParticleNet	0.513	0.846	0.948
RobustParTAK4	0.517	0.853	0.959

(b)

Table 7.1: Values of calculated significance estimators $\mathcal{Z}$ for the three b-tagging algorithms and different working points, according to equation 7.4. The algorithms are employed on the 2j1b selection in table (a) and on the 3j3b selection in (b). The b-tagging SFs are not applied to obtain these values. All other mentioned corrections are considered.

As the significance estimator is observed to be the largest for all working points in both the 2j1b and 3j3b selections, the nominal choice of the RobustParTAK4 algorithm is justified. In general, a tighter working point is seen to increase the significance estimator. Specifically, for the RobustParTAK4 algorithm, the approximate $\mathcal{Z}$ estimator increases by roughly 12% in the 3j3b selection by applying a tight cut, compared to the medium working point.
To visualize the effect on background and signal event counts of b-jet multiplicity requirements and of the different working points for the RobustParTAK4 algorithm, a cut flow diagram is shown in figure 7.1. The leftmost bin starts with the loosest cuts, which get tighter with rising

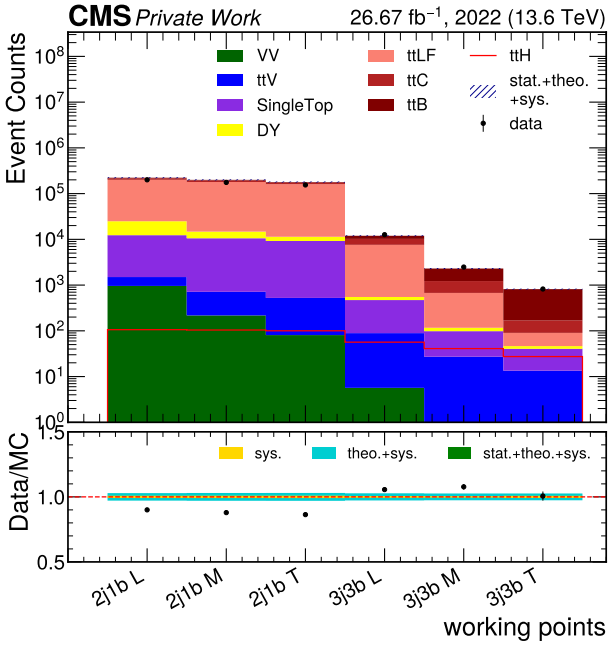


Figure 7.1: Cut flow diagram for the RobustParTAK4 algorithm at different working points and for the 2j1b and 3j3b selection. The leftmost three bins correspond to the 2j1b selection with loose, medium and tight working points, depicted in this order. The rightmost three bins show the same for the 3j3b selection.

bins, meaning the first three bins correspond to the 2j1b selection at loose, medium and tight working points and the last three depict the same for the 3j3b selection. The difference between the counts of the selections is striking, as event statistics are reduced by about two orders of magnitude, when applying 3j3b requirements. The background is decreased to around 5% of its 2j1b value, while ttH signal is only reduced to about 37%, which explains the rise in significance. Additionally, the cut flow demonstrates the challenges imposed by $\mathrm{tt} + \mathrm{bb}$ events in the signal region, explained in section 2.4. Specifically, $\mathrm{tt}+\mathrm{bb}$ events with the same final state as the signal are included in the ttB category. The most significant decrease of tt background is observed for ttLF events. The ttLF (ttC) background could be reduced to about 0.3% (3%) in the 3j3b selection, while the ttB background is only reduced to 20% of its 2j1b value. The significant reduction in background relative to the signal, due to the stricter constraint on ttLF events in the 3j3b selection, explains that the 3j3b selection in general shows a much larger significance than the 2j1b selection.
All $\mathcal{Z}$ estimator values provided in table 7.1 suggest applying the tight working point. But as analyses and fitting methods are reliant on high event statistics, it is unclear if the tight requirement limits statistical accuracy too much to be considered an improvement. This needs to

be checked separately. Nonetheless, the significance estimator presented here can guide towards potential improvements, which need to be tested more in the future.
It is important to point out, that the values in table 7.1 were derived without applying b-tagging efficiency corrections. This was necessary, because the fixed working point b-tag corrections are only implemented in the analysis framework for the RobustParTAK4 algorithm for the medium working point, at the time of writing. As all following cut optimizations are done with the nominal b-tagger and working point, b-tag corrections are applied there. The significance of the nominal 3j3b selection with b-tag corrections is notably smaller ($\mathcal{Z} = 0.785$) than without ($\mathcal{Z} = 0.853$), while in the 2j1b selection both values are the same ($\mathcal{Z} = 1.233$). The large difference in the 3j3b selection initially seems counter intuitive, as the dominant $\mathrm{tt} + \mathrm{bb}$ background and $\mathrm{ttH}(\mathrm{H} \to \mathrm{bb})$ signal yields are expected to share a similar phase space and should be effected similarly by the corrections. However, the ttLF components were specifically found to be underestimated when not applying the b-tag corrections, as this background component increases by 40% when they are applied. In comparison, the signal and ttB background yields are decreased by about 5%, while ttC yields decrease by 3% when applying b-tag corrections. Overall, the total background yield increases by approximately 7%, which almost solely originates from the ttLF components increase. This underestimation of the ttLF component causes the notable difference in the significance. This demonstrates the limitations of the estimator when not applying all efficiency corrections. Nonetheless, since tighter working points include fewer ttLF events, the significance presented for the tight working point should be closer to the value with b-tag corrections applied than for looser working points. Additionally, because the ttLF background is smaller relative to the signal when not applying the corrections, looser working points are favoured in table 7.1. Despite this, the tighter working point still possesses the largest significance estimator and the final choice of the tight working point remains valid.
All background components and the signal yield do not change notably in the 2j1b selection between the analysis with and without b-tag scale factors. The total scale factor on the background yield is dependent on the amount of correctly b-tagged and miss-tagged jets in the selection. This way miss-tag efficiencies in simulation get corrected. As the 3j3b selection requires more b-jets, the ttLF events passing the selection have a larger miss-tag ratio of jets than in the 2j1b selection. This leads to a larger impact of corrections on the ttLF component in the 3j3b selection (typically, miss-tag efficiencies receive larger Data-to-MC corrections), while its yield remains almost the same for the 2j1b selection.

7.3.2 Pseudorapidity Increase for Electrons

In this section, the maximum electron η increase from 2.4 to 2.5 is analysed. Table 7.2 displays the calculated significance values for the 2j1b and 3j3b selections. For both selections an increase can be observed for the larger electron acceptance, however the change is almost negligibly small. Nonetheless, the increase suggests at least small potential for a larger data set.

2j1b	
maximum $\lvert\eta\rvert$	
2.4	2.5
0.233	0.234

(a)

3j3b	
maximum $\lvert\eta\rvert$	
2.4	2.5
0.785	0.787

(b)

Table 7.2: Values of calculated significance estimators $\mathcal{Z}$ according to formula 7.4 for the nominal and the increased maximum electron $|\eta|$ cut in the 2j1b (a) and 3j3b (b) selections.

7.3.3 Jet p_{T} Requirement

Regarding the jet p_{T}, a decrease of the nominal 30 GeV cut is tested for 25 and 20 GeV. The corresponding results are listed in table 7.3.

2j1b		
minimum jet p_T in GeV		
20	25	30
0.225	0.233	0.237

(a)

3j3b		
minimum jet p_T in GeV		
20	25	30
0.784	0.791	0.785

(b)

Table 7.3: Values of calculated significance estimators $\mathcal{Z}$ according to formula 7.4 for the different jet p_{T} requirements on the 2j1b (a) and 3j3b (b) selections.

While for the 2j1b selection the significance decreases with a lowered jet p_{T} cut, the 3j3b selections estimator is largest for a minimum jet p_{T} of 25 GeV. Still, this increase is not substantial and only appears at the third decimal point, similar to the electron η increase.

7.3.4 Z Window and $p_{\mathrm{T}}^{\mathrm{miss}}$ Cut

Lastly, changes to the nominal Z window and $p_{\mathrm{T}}^{\mathrm{miss}}$ cut are investigated. The minimum missing transverse momentum is analyzed over a range from complete removal to 45 GeV. Similarly, the Z window cuts are examined over a range from complete removal to the standard nominal window. Because both cuts are largely correlated, as described in 5.1, every possible cut combination is tested. The calculated $\mathcal{Z}$ estimators for all combinations and both selections can be found in table 7.4.

		minimum p_T^{miss} in GeV						
	2j1b	0	10	20	30	35	40	45
Z window in GeV	no cut	0.218	0.220	0.225	0.230	0.231	0.232	0.231
	(86,96)	0.244	0.244	0.244	0.242	0.241	0.238	0.236
	(81,101)	0.245	0.245	0.244	0.241	0.239	0.236	0.233
	(79,103)	0.245	0.244	0.242	0.240	0.238	0.235	0.232
	(76,106)	0.243	0.243	0.241	0.238	0.235	0.233	0.230

(a)

		minimum p_T^{miss} in GeV						
	3j3b	0	10	20	30	35	40	45
Z window in GeV	no cut	0.869	0.865	0.858	0.841	0.832	0.820	0.807
	(86,96)	0.875	0.871	0.860	0.839	0.828	0.814	0.799
	(81,101)	0.861	0.857	0.845	0.824	0.813	0.799	0.784
	(79,103)	0.855	0.850	0.839	0.818	0.807	0.793	0.778
	(76,106)	0.847	0.842	0.830	0.809	0.799	0.785	0.770

(b)

Table 7.4: Values of calculated significance estimators $\mathcal{Z}$ according to formula 7.4 for the different $p_{\mathrm{T}}^{\mathrm{miss}}$ and Z window cuts for the invariant mass $m_{\mathrm{ee}/\mu\mu}$ on the the 2j1b (a) and 3j3b (b) selections.

Interestingly, for both the 2j1b and 3j3b selections the significance can be increased by largely loosening the Z window and $p_{\mathrm{T}}^{\mathrm{miss}}$ cuts simultaneously. Specifically, for the 3j3b selection, the significance peaks with a Z window exclusion of $m_{\mathrm{ee}/\mu\mu} \notin (86, 96)\,\mathrm{GeV}$, combined with the removal of the $p_{\mathrm{T}}^{\mathrm{miss}}$ cut, resulting in an increase of approximately 11% in the $\mathcal{Z}$ estimator. For the 2j1b selection the range of the significance peak is broader, but still suggests loosening both

the $p_{\mathrm{T}}^{\mathrm{miss}}$ and Z window requirements. While the complete removal of the Z window would notably decrease the significance in the 2j1b selection, it only slightly decreases the 3j3b estimator. This suggests that a significant portion of the DY background is already filtered by the jet and b-jet requirements in the signal region, as can be seen in the cut flow in figure 7.1. This is expected, as leading order DY processes do not include jets and processes with additional jet radiation, especially of heavy flavour jets, have significantly lower cross sections. Thus, in the 3j3b selection DY events are not dominant background processes and tighter DY cuts remove the dominant tt background and $\mathrm{ttH(H \to bb)}$ signal to the same extend, because of the similar phase space. Therefore, according to formula 7.4 the signal sensitivity decreases for stricter DY cuts. When completely removing the Z window and $p_{\mathrm{T}}^{\mathrm{miss}}$ cuts in the 3j3b selection, enough DY background gets introduced to invert this trend. Because less DY is already removed by the jet and b-jet cuts in the 2j1b selection, this turning point is reached earlier than in the 3j3b selection.

7.4 The Final Optimized Selection

Based on the changes in cuts, which yield the largest $\mathcal{Z}$ estimators presented in section 7.3, the complete optimized selection is presented here:

1. at least three jets (see Sec. 4.3.4)
2. at least three b-tagged jet
3. exactly two oppositely-charged leptons (either $\mathrm{e^{\pm}e^{\mp}}$, $\mu^{\pm}\mu^{\mp}$, or $\mathrm{e^{\pm}}\mu^{\mp}$)
4. **muons within $|\eta| < 2.4$**
5. **electrons within $|\eta| < 2.5$**
6. a leading lepton with $p_{\mathrm{T}} > 25\,\mathrm{GeV}$ and a subleading lepton with $p_{\mathrm{T}} > 15\,\mathrm{GeV}$
7. **invariant mass of ee and $\mu\mu$ pairs $m_{\mathrm{ee}/\mu\mu} < 86$ GeV or $m_{\mathrm{ee}/\mu\mu} > 96$ GeV**
8. $m_{\mathrm{ee}/\mu\mu} > 20\,\mathrm{GeV}$
9. **minimum jet p_{T} of 25 GeV**
10. maximum jet $|\eta|$ of 2.4

All requirement changes to the nominal selection are written in bold, while the $p_{\mathrm{T}}^{\mathrm{miss}}$ cut was completely removed and the RobustParTAK4 algorithm is employed at the tight working point. Consequently, b-tag scale factors are not included in analyses of the optimized selection. The significance estimator for this selection yields $\mathcal{Z} = 1.08$, which is 26.6% larger than the nominal value derived without b-tag scale factors of $\mathcal{Z} = 0.853$. This demonstrates the combined potential of all optimized cuts. Because the nominal significance is observed to be smaller when b-tag scale factors are considered ($\mathcal{Z} = 0.785$), the total increase is expected to change, but the trend should remain the same, as discussed in section 7.3.1.
Figures 7.2 and 7.3 present a selection of the variables shown for the nominal 2j1b selection in chapter 6, for the optimized 3j3b selection. As suggested by the improved modeling for higher b-jet multiplicities in the $N_{\text{b-jets}}$ distribution (Figure 6.1b), all variables show significantly improved agreement between the simulation normalization and data. As expected for the signal region, the ttH yields are now more pronounced and constitute a larger ratio of the whole simulation. Both the $N_{\text{b-jets}}$ and N_{jets} distributions (figures 7.2a and 7.2b respectively) are well modeled in shape and normalization. The plots reveal a significant reduction in statistics compared to the nominal 2j1b selection of approximately two orders of magnitude. Thus, the binning granularity had to

be decreased in all other variables to enable sensible comparisons between data and simulation. The tighter working point is expected to shift the discontinuity in the b-tag score distribution to around 0.8604, as confirmed in figure 7.2c. The corresponding shape is also well modelled. In the H_T distribution of figure 7.2d, no clear trend is visible, unlike in the baseline selection. However, a small downward trend is still visible in the $p_{\mathrm{T}}^{b_1}$ distribution of the leading b-jet (figure 7.3a), but it remains within expected fluctuations. The corresponding η^{b_1} distribution (figure 7.3b) is described well by the simulation. Further, the leading lepton $p_{\mathrm{T}}^{\ell_1}$ distribution in figure 7.3c seems well modeled in shape and normalization. Lastly, the N_{PV} distribution shows good agreement in figure 7.3d. It is important to point out, that the lowered bin resolution makes it significantly harder to distinguish trends in the plots. Additional plots of $p_{\mathrm{T}}^{\mathrm{miss}}$ and $\Phi(p_{\mathrm{T}}^{\mathrm{miss}})$, as well es leading lepton η^{ℓ_1} and leading b-jet Φ^{b_1} can be found in Appdx. 5 in figure A5.1. Despite the incomplete implementation of jet energy corrections and uncertainties, and despite applying p_{T} reweighting based on recommendations for $\sqrt{s} = 13\,\mathrm{TeV}$, the agreement between data and simulation is already very good.

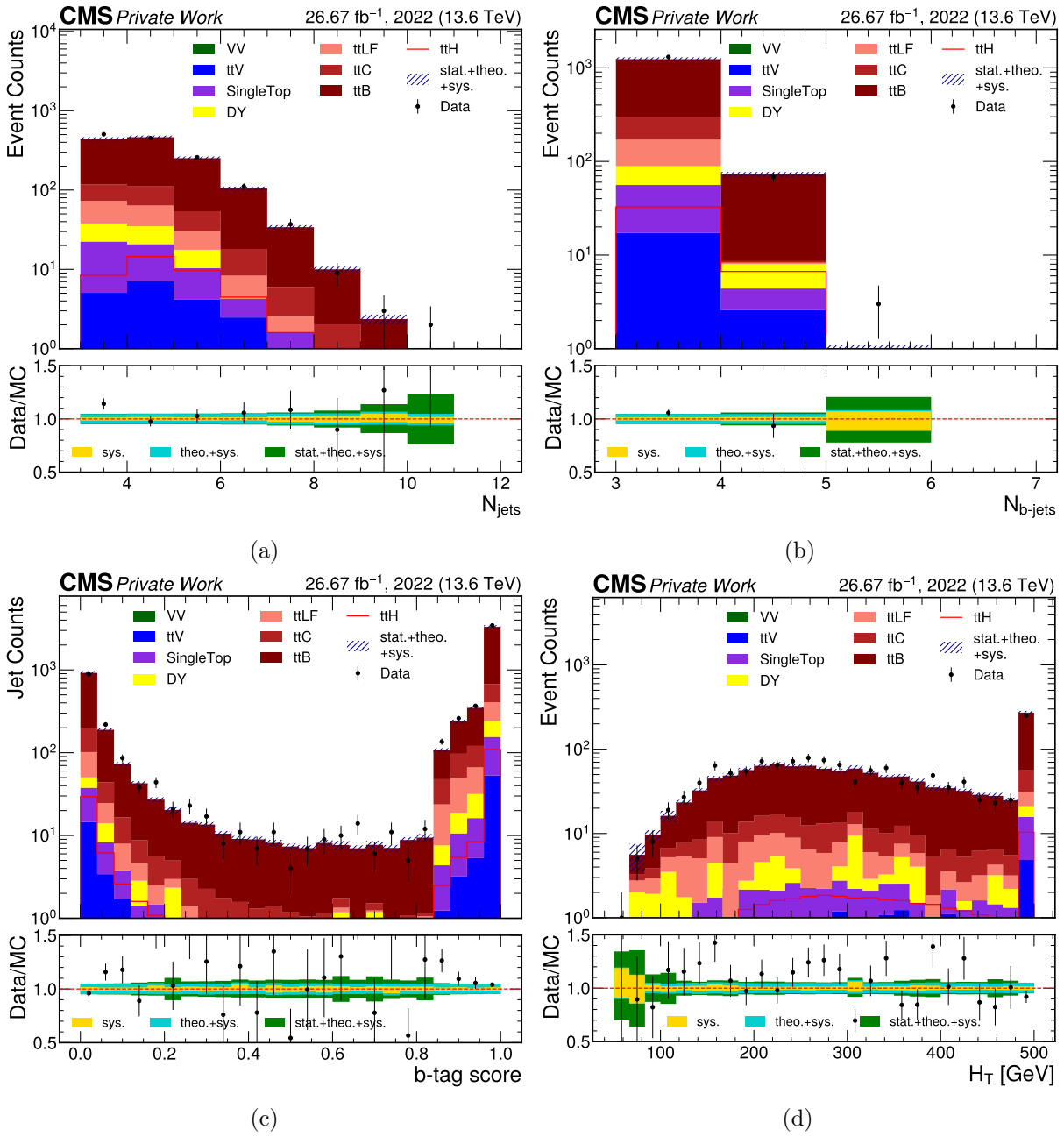
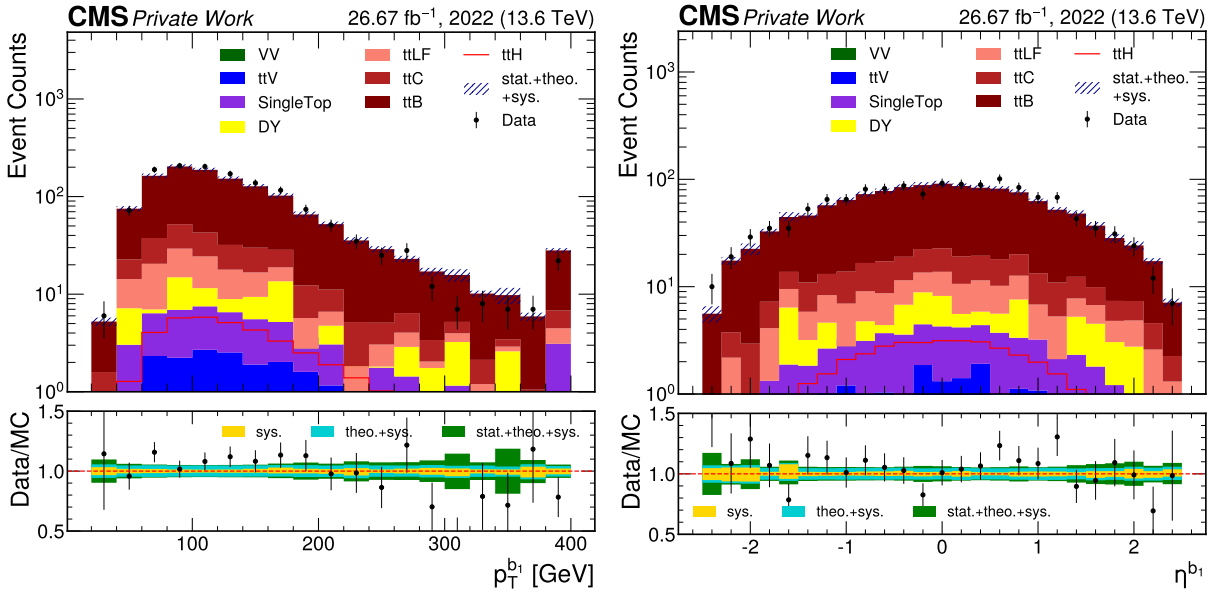
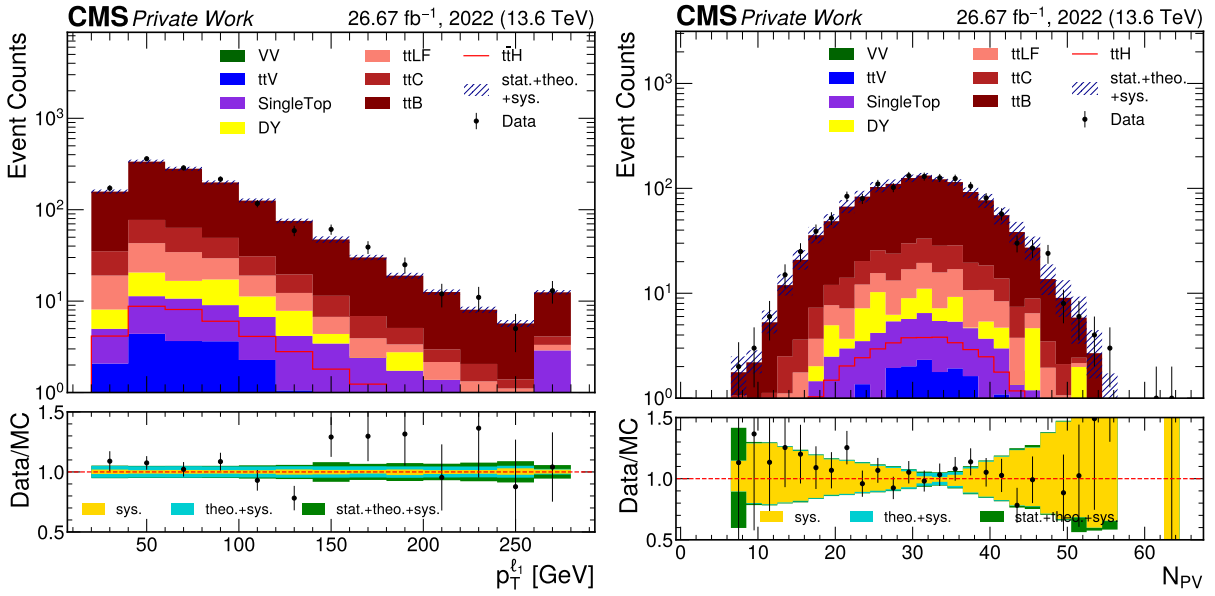


Figure 7.2: Plot (a)/(b) shows the event count as a function of the number of jets/b-jets in an event. Plot (c) shows the count of all jets in the selection as a function of their individual b-tagging scores. In (d), the event count is plotted against the sum of all jet transverse momenta H_T. All plots consider events which pass the optimized 3j3b event selection using the RobustParTAK4 algorithm for b-tagging at the tight working point without b-tag scale factors. The outermost bins include overflow counts.



(a) (b)



(c) (d)

Figure 7.3: Figures (a) and (c) display transverse momentum distributions, while (b) and (d) depict the pseudorapidity of all selected events for the leading b-jet and leading lepton respectively. Events need to pass the optimized 3j3b event selection using the RobustParTAK4 algorithm for b-tagging at the tight working point without b-tag scale factors. The outermost bins include overflow counts.

8 Summary and Outlook

This thesis provided first data-simulation comparisons of kinematic variables for data recorded with the CMS detector during Run 3 at a center of mass energy of $\sqrt{s} = 13.6$ TeV, in the context of ttH(H $\rightarrow$ bb) analysis. Moreover, an event selection optimization was performed, aiming to maximize signal sensitivity of the ttH(H $\rightarrow$ bb) process, using a simulation based significance estimator. Changed event requirements were examined, using selections employed in previous Run 2 measurements [14] as a starting point. All the studies presented used data recorded during 2022 with an integrated luminosity of $26.67\,\text{fb}^{-1}$ [23] and the corresponding MC simulations of the signal process and its backgrounds.

Two event selections were analyzed, specifically one baseline region and a selection requiring higher b-jet multiplicities, which would be a possible signal region. The nominal baseline selection was used to provide a first look into the Run 3 data quality. In the produced figures an overall overestimation of the simulation normalization was observed for all baseline selection variables. The agreement was found to improve for events with higher b-jet multiplicity, later confirmed by good normalization in the optimized signal region. But considering not all important uncertainties and corrections could be applied at the time of writing, a relatively good agreement was observed in all variables.

Next, changes on the nominal event requirements of both selections were investigated, which had potential of increasing the significance estimator. In particular, the impact of different b-tagging algorithms and working points and an increase in the electron pseudorapidity to the full detector acceptance were analyzed. Additionally, different cuts for the minimum required missing transverse momentum and the Z resonance exclusion window were tested, and a decrease in the minimum required jet p_T was examined. The optimized selection based on the cuts, which provided the largest significance, was found to increase the estimator by about 26%, compared to the nominal signal region. While a tighter working point and largely reduced DY exclusions were shown to substantially increase the significance, changes through jet p_T and electron η cuts were marginal.

Lastly, data-MC agreement was examined for the optimized signal region requirements. In all variables the initial normalization discrepancy observed in the 2j1b region was not present in the optimized region, and good shape agreement was observed.

One of the most interesting results of this thesis is the increased signal sensitivity observed by substantially loosening both the p_T^miss and Z window cuts, compared to the cuts employed in the Run 2 ttH analysis. This suggests that DY background is not substantially contributing in the signal region, and future ttH(H $\rightarrow$ bb) measurements may benefit from reduced DY requirements. Still, the estimator optimization has several shortcomings, such as only being a statistical measure not regarding systematic uncertainties and making different effects of cuts on the separate background components indistinguishable. In particular, the optimized selection was shown to improve signal sensitivity, but could not reduce the critical ttB background. Additionally, the significance estimator analysis is limited when not applying all efficiency corrections on the simulated samples. Despite this, the estimator presented here is a good first approach to identify possible improvements, which need to be investigated further with more sophisticated statistical tools once concrete analysis methods are determined for Run 3 ttH analyses.

Appendix 1: CMS Data Sets and Triggers

The used CMS data sets were recorded in the 2022 postEE period (era E-G) by the CMS detector. The paths to all data sets are summarized in table A1.1. Further, all triggers used in this thesis are listed in table A1.2.

Table A1.1: 2022 postEE data samples used for the analysis.

PD	Path
`/EGamma`	`/Run2022{E,F,G}-22Sep2023-v1/NANOAOD`
`/Moun`	`/Run2022{E,F,G}-22Sep2023-v1/NANOAOD`
`/MuonEG`	`/Run2022{E,F,G}-22Sep2023-v1/NANOAOD`

Table A1.2: List of all high-level triggers used in this thesis and their usage in corresponding final states

final state	HLT trigger paths
ee	Ele23_Ele12_CaloIdL_TrackIdL_IsoVL Ele30_WPTight_Gsf
eμ	Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL_DZ Mu12_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ Ele30_WPTight_Gsf IsoMu24
$\mu\mu$	Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass3p8 IsoMu24

Appendix 2: Simulated Samples

All paths to the used simulation datasets with their corresponding crossection are listed in table A2.1. The full paths follow the structure:

{Process Path}/Run3Summer22EENanoAODv12-130X_mcRun3_2022_realistic_postEE_{Versions}/NANOAODSIM.

For the full path of each dataset the corresponding information from table A2.1 have to be added at {Process Path} and {Versions} respectively.

Table A2.1: Simulated samples used in the analysis with their corresponding cross section.

associated background	Process Path	versions	σ [pb]
tt	/TTto2L2Nu_TuneCP5_13p6TeV_powheg-pythia8	v6-v2	98.0963
	/TTtoLNu2Q_TuneCP5_13p6TeV_powheg-pythia8	v6-v2	405.8099
Drell Yan	/DYto2L-2Jets_MLL-10to50_TuneCP5_13p6TeV_amcatnloFXFX-pythia8	v6-v2	19982.5
	/DYto2L-2Jets_MLL-50_TuneCP5_13p6TeV_amcatnloFXFX-pythia8	v6-v2	6345.99
single top	/TWminusto2L2Nu_TuneCP5_13p6TeV_powheg-pythia8	v6-v2	4.6511
	/TbarWplusto2L2Nu_TuneCP5_13p6TeV_powheg-pythia8	v6-v2	4.6511
WW, WZ, ZZ	/WWto2L2Nu_TuneCP5_13p6TeV_powheg-pythia8	v6-v2	12.98
	/WZto3LNu_TuneCP5_13p6TeV_powheg-pythia8	v6-v2	8.17
	/ZZto2L2Nu_TuneCP5_13p6TeV_powheg-pythia8	v6-v2	1.19
ttV	/TTLL_MLL-4to50_TuneCP5_13p6TeV_amcatnlo-pythia8	v6-v2	0.03949
	/TTLL_MLL-50_TuneCP5_13p6TeV_amcatnlo-pythia8	v6-v2	0.08646
	/TTNuNu_TuneCP5_13p6TeV_amcatnlo-pythia8	v6-v2	0.1638
	/TTLNu-1Jets_TuneCP5_13p6TeV_amcatnloFXFX-pythia8	v6-v4	0.25
	/TTZ-ZtoQQ-1Jets_TuneCP5_13p6TeV_amcatnloFXFX-pythia8	v6-v2	0.6209
ttH	/TTH_Hto2B_M-125_TuneCP5_13p6TeV_powheg-pythia8	v6-v3	0.331968

Appendix 3: Identification Requirements

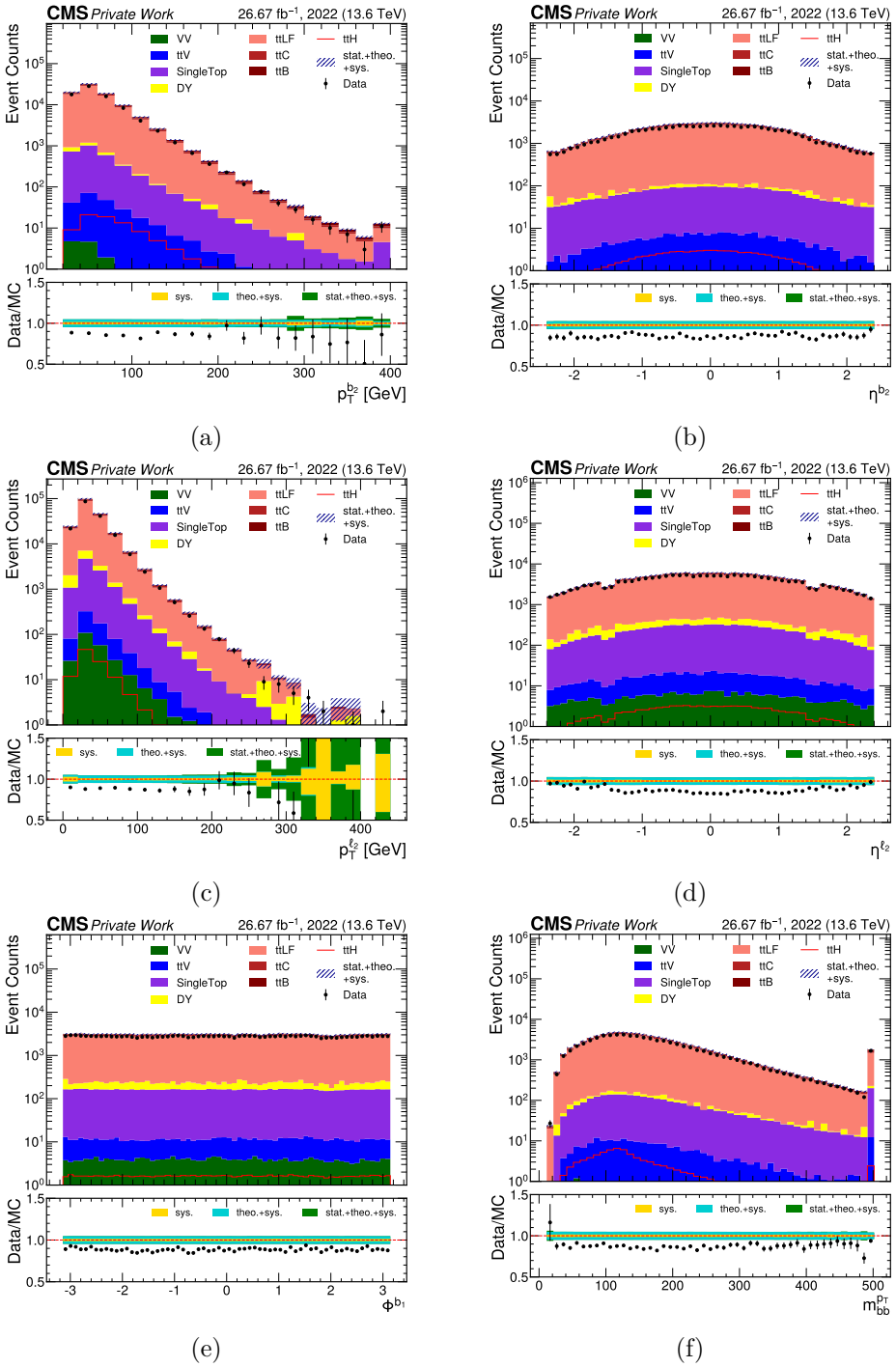
Table A3.1: List of the tight cut-based muon ID requirements [44].

Parameter	Muon ID Requirement
Global Muon	true
χ^2_{track}/n_{dof}	< 10
Muon Chamber Hits	> 0
Matched Stations	> 1
$\lvert d_0 \rvert$	< 2 mm
$\lvert d_z \rvert$	< 5 mm
Pixel Hits	> 0
Track Layer Hits	> 5
r^{μ}_{Iso}	< 0.15

Table A3.2: Jet ID [45] definitions of the TIGHTLEPVETO working point.

Parameter	Jet ID requirement
Neutral Hadron Fraction	< 0.99
Neutral EM Fraction	< 0.90
Number of Constituents	> 1
Muon Fraction	< 0.80
Charged Hadron Fraction	> 0.01
Charged Multiplicity	> 0
Charged EM Fraction	< 0.80

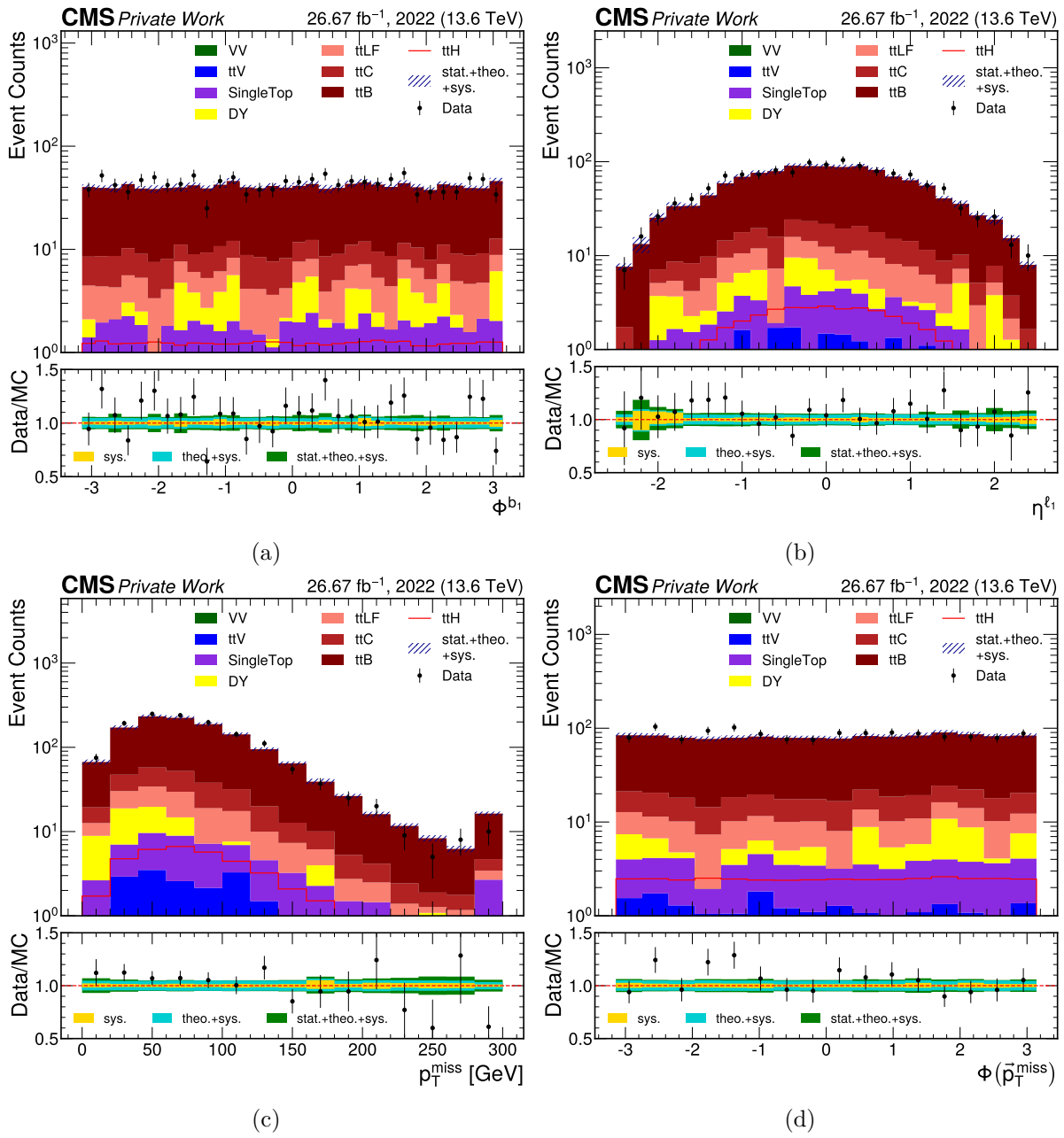
Appendix 4: Additional Baseline Plots



(a) (b) (c) (d) (e) (f)

Figure A4.1: Figure (a)/(c) shows the event counts as a function of the subleading jet/lepton p_T distribution, while figure (b)/(d) displays the subleading jet/lepton η distribution. Figure (e) depicts the Φ component of leading b-jets and (f) shows the invariant mass distribution of the highest p_T b-jet pair system. All plots consider events which pass the baseline 2j1b event selection using the RobustParTAK4 algorithm for b-tagging at the medium working point. The outermost bins include potential overflow counts.

Appendix 5: Additional 3j3b Plots



(a) (b)

(c) (d)

Figure A5.1: Figure (a) shows the distribution of the leading b-jet ϕ while figure (b) displays the leading lepton η. In figure (c) the missing transverse momentum distribution is depicted, while (d) shows the corresponding Φ component. All plots consider events which pass the optimized 3j3b event selection using the RobustParTAK4 algorithm for b-tagging at the tight working point. The outermost bins include potential overflow counts.

Appendix 6: Significance Estimator Verification

2j1b	loose	medium	tight
DeepJet	0.224	0.233	0.237
ParticleNet	0.225	0.233	0.237
RobustParTAK4	0.225	0.233	0.237

(a)

3j3b	loose	medium	tight
DeepJet	0.504	0.815	0.892
ParticleNet	0.513	0.843	0.943
RobustParTAK4	0.516	0.850	0.953

(b)

Table A6.1: Values of calculated significance estimators $\mathcal{Z}$ for the three b-tagging algorithms and different working points, according to equation 7.3. The algorithms are employed on the 2j1b selection in table (a) and on the 3j3b selection in (b). The b-tagging SFs are not applied to obtain these values. All other corrections are considered.

References

[1] C. Berger, "Elementarteilchenphysik: Von den Grundlagen zu den modernen Experimenten", Springer Spektrum, 2014. doi:10.1007/978-3-642-41753-5.

[2] S. P. Martin, "A Supersmmetry Primer", p. 1–98. World Scientific, July, 1998. doi:10.1142/9789812839657_0001.

[3] W. Demtröder, "Experimentalphysik 4: Kern-, Teilchen- und Astrophysik", Springer Spektrum, 2017. doi:10.1007/978-3-662-52884-6.

[4] W. Hollik, "Quantum field theory and the Standard Model", in *High-energy physics. Proceedings, 17th European School, ESHEP 2009, Bautzen, Germany.* 2010. arXiv:1012.3883.

[5] Wikipedia contributors, "Standard Model — Wikipedia, The Free Encyclopedia". https://en.wikipedia.org/wiki/Standard_Model. Accessed 30.05.2024.

[6] CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC", *Physics Letters B* **716** (2012), doi:10.1016/j.physletb.2012.08.021.

[7] CMS Collaboration, "Observation of a new boson with mass near 125 GeV in pp collisions at $\sqrt{s} = 7$ and 8 TeV", *Journal of High Energy Physics* (2013), doi:10.1007/jhep06(2013)081.

[8] ATLAS Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC", *Physics Letters B* **716** (September, 2012), doi:10.1016/j.physletb.2012.08.020.

[9] CMS Collaboration, "A portrait of the Higgs boson by the CMS experiment ten years after the discovery", *Nature* **607** (2022), doi:10.1038/s41586-022-04892-x.

[10] P. W. Higgs, "Broken Symmetries and the Masses of Gauge Bosons", *Phys. Rev. Lett.* **13** (Oct, 1964), doi:10.1103/PhysRevLett.13.508.

[11] LHC Higgs Cross Section Working Group, "Handbook of LHC Higgs cross sections: 4. Deciphering the nature of the Higgs sector", *CERN Yellow Reports: Monographs, Vol 2* (2017), doi:10.23731/CYRM-2017-002.

[12] M. Czakon and A. Mitov, "ATLAS-CMS recommended predictions for top-quark-pair cross sections using the Top++v2.0 program". https://twiki.cern.ch/twiki/bin/view/LHCPhysics/TtbarNNLO. Accessed 08.07.2024.

[13] A. Karlberg et al., "Ad interim recommendations for the Higgs boson production cross sections at $\sqrt{s} = 13.6$ TeV", 2024. doi:10.48550/arXiv.2402.09955.

[14] CMS Collaboration, "Measurement of the $\mathrm{t\bar{t}H}$ and tH production rates in the $\mathrm{H} \to \mathrm{b\bar{b}}$ decay channel with $138\,\mathrm{fb}^{-1}$ of proton-proton collision data at $\sqrt{s} = 13\,\mathrm{TeV}$". https://cds.cern.ch/record/2868175, 2023.

[15] L. Evans and P. Bryant, "LHC Machine", *JINST* **3** (2008), doi:10.1088/1748-0221/3/08/S08001.

[16] ALICE Collaboration, "The ALICE experiment at the CERN LHC", *JINST* **3** (2008), doi:10.1088/1748-0221/3/08/S08002.

[17] ATLAS Collaboration, "The ATLAS Experiment at the CERN Large Hadron Collider", *JINST* **3** (2008), doi:10.1088/1748-0221/3/08/S08003.

[18] LHCb Collaboration, "The LHCb Detector at the LHC", *JINST* **3** (2008), doi:10.1088/1748-0221/3/08/S08005.

[19] CMS Collaboration, "The CMS experiment at the CERN LHC", *JINST* **3** (2008), doi:10.1088/1748-0221/3/08/S08004.

[20] A. H. et al, "Development of the CMS detector for the CERN LHC Run 3", *JINST* (2024), doi:10.1088/1748-0221/19/05/P05064.

[21] R. Mommsen et al., "The CMS event-builder system for LHC Run 3 (2021-23)", 2019. doi:10.1051/epjconf/201921401006.

[22] D. Barney, "CMS Detector Slice". https://cds.cern.ch/record/2120661, 2016. CMS Collection. Accessed 30.05.2024.

[23] CMS Collaboration, "Public CMS Luminosity Information". https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults. Accessed 05.07.2024.

[24] T. Sjöstrand, S. Mrenna, and P. Skands, "A brief introduction to PYTHIA 8.1", *Comput. Phys. Commun.* **178** (2008), doi:10.1016/j.cpc.2008.01.036.

[25] GEANT4 Collaboration, "Geant4—a simulation toolkit", *Nucl. Instrum. Meth. A* **506** (2003), doi:10.1016/S0168-9002(03)01368-8.

[26] M. Czakon et al., "Top-pair production at the LHC through NNLO QCD and NLO EW", *Journal of High Energy Physics* **2017** (October, 2017), doi:10.1007/jhep10(2017)186.

[27] CMS Collaboration, "TOP PAG corrections based on theory and simulation aka NNLO-NLO weights". https://twiki.cern.ch/twiki/bin/viewauth/CMS/TopPtReweighting#TOP_PAG_corrections_based_on_the . Accessed 24.06.2024.

[28] CMS Collaboration, "Particle-flow reconstruction and global event description with the CMS detector", *Journal of Instrumentation* **12** (October, 2017), doi:10.1088/1748-0221/12/10/P10003.

[29] F. Beaudette, "The CMS Particle Flow Algorithm", 2014. doi:10.48550/arXiv.1401.8155.

[30] "MVA Based Electron ID for Run 3". https://twiki.cern.ch/twiki/bin/view/CMS/MultivariateElectronIdentificationRun3. Accessed 19.06.2024.

[31] CMS Collaboration, "Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC", *Journal of Instrumentation* **16** (May, 2021), doi:10.1088/1748-0221/16/05/p05014.

[32] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-ktjet clustering algorithm", *Journal of High Energy Physics* **2008** (April, 2008), doi:10.1088/1126-6708/2008/04/063.

[33] CMS Collaboration, "Pileup mitigation at CMS in 13 TeV data", *Journal of Instrumentation* **15** (September, 2020), doi:10.1088/1748-0221/15/09/P09018.

[34] E. Bols et al., "Jet flavour classification using DeepJet", *Journal of Instrumentation* **15** (December, 2020), doi:10.1088/1748-0221/15/12/p12012.

[35] H. Qu and L. Gouskos, "Jet tagging via particle clouds", *Physical Review D* **101** (March, 2020), doi:10.1103/physrevd.101.056019.

[36] Annika Stein, Alexandre de Moor, "Jet flavour identification for Run 3 with the RobustParTAK4 algorithm", *CMS Analysis note* **AN-2019/228**.

[37] A. Stein et al., "Improving Robustness of Jet Tagging Algorithms with Adversarial Training", *Computing and Software for Big Science* **6** (September, 2022), doi:10.1007/s41781-022-00087-1.

[38] H. Qu, C. Li, and S. Qian, "Particle Transformer for Jet Tagging", 2024. doi:10.48550/arXiv.2202.03772.

[39] Davide Valsecchi, Matteo Marchegiani, "PocketCoffea". https://pocketcoffea.readthedocs.io/en/stable/. Accessed 07.07.2024.

[40] CMS Collaboration, "Utilities for Accessing Pileup Information for Data". https://twiki.cern.ch/twiki/bin/viewauth/CMS/PileupJSONFileforData#Recommended_cross_section. Accessed 24.06.2024.

[41] CMS Collaboration, "Standard Model Cross Sections for CMS at 13 TeV". https://twiki.cern.ch/twiki/bin/viewauth/CMS/StandardModelCrossSectionsat13TeV . Accessed 26.06.2024.

[42] "2022 (Summer22EE) heavy flavour tagging scale factors". https://btv-wiki.docs.cern.ch/ScaleFactors/Run3Summer22EE/ . Accessed 08.07.2024.

[43] G. Schott, "Hypothesis Testing", ch. 3, pp. 75–105. John Wiley & Sons, Ltd, 2013. doi:https://doi.org/10.1002/9783527653416.ch3.

[44] "Muon recommendations for 2022 data and Monte Carlo". https://twiki.cern.ch/twiki/bin/view/CMS/MuonRun32022. Accessed 19.06.2024.

[45] "Jet Identification for the 13.6 TeV data". https://twiki.cern.ch/twiki/bin/view/CMS/JetID13p6TeV. Accessed 19.06.2024.