ETP-KA/2021-1

Karlsruhe Institute of Technology

Studies on $t\bar{t}+b\bar{b}$ production at the CMS experiment

Master Thesis

Emanuel Lorenz Pfeffer

At the Department of Physics
Institute of Experimental Particle Physics

Reviewer:	Prof. Dr. Ulrich Husemann
Second reviewer:	Prof. Dr. Thomas Müller
Advisor:	Dr. Matthias Schröder
Second advisor:	Jan van der Linden

Karlsruhe, 13. 01. 2021

This thesis has been accepted by the first reviewer of the master thesis.

PLACE, DATE

.....

(**Prof. Dr. Ulrich Husemann**)

I declare that I have developed and written the enclosed thesis completely by myself, and have not used sources or means without declaration in the text.

PLACE, DATE

.....

(**Emanuel Lorenz Pfeffer**)

Contents

1 Introduction

Answering the question of the most fundamental constituents of nature has always been in the nature of humankind. Particle physics is devoted to this question and tries to give an answer with the Theory of the Standard Model (SM). Within this theory, all known particles and three of the four fundamental interactions are described. One of the greatest achievements of the SM is the discovery of the predicted Higgs boson, which was discovered in 2012 at the ATLAS and CMS experiments at the CERN Large Hadron Collider (LHC) [1–5]. Even if the SM offers excellent predictions and explanations for many observed effects, some phenomena remain unexplained. Thus, the SM indubitably leaves open questions, for example the inclusion of gravity. The systematic testing, but also the attempted refutation of the theoretical model therefore constitutes an important task in particle physics.

In this thesis, the interplay of two elementary particles of the SM, the two heaviest quarks top (t) and bottom (b), is studied in detail. Processes in which a top quark-antiquark pair occurs in associated production with two bottom quarks ($t\bar{t}+b\bar{b}$) play an important role in particle physics. The two quarks show a high mass difference and their associated production is therefore particularly difficult to model, accompanied by large uncertainties [6]. In addition, events involving this process constitute a large irreducible background in measurements of $t\bar{t}+H$ production in $H \to b\bar{b}$ decays. These measurements are an essential test of the SM and an important constraint of physics beyond the SM. Consequently, the $t\bar{t}+b\bar{b}$ process needs to be thoroughly understood.

Events comprising $t\bar{t}+b\bar{b}$ processes are analyzed at two different Monte Carlo event simulation levels in this thesis.
The first part compares different existing event simulations used at the CMS experiment for the $t\bar{t}+b\bar{b}$ process in the single-lepton channel and investigates possible differences in the modeling. The analysis is technically realized in a way that allows a direct comparison of the simulations with those of the ATLAS experiment. The comparison of simulations of both experiments is performed building on this analysis. This enables the design of a common strategy between ATLAS and CMS.

The second part focuses on strategies for assigning b jets to their origin. Since top quarks decay into b quarks almost exclusively, the final state of an event involving the $t\bar{t}+b\bar{b}$ process usually consists of four or more b jets. However, it is unknown in the event reconstruction which b jets result from top quark decays and which originate from

additional gluon radiation and subsequent splittings into pairs of b quarks in the event. Two main methods for possible assignments are presented. Applying a predefined metric to evaluate the performance of the assignment, the first method follows a straightforward approach and examines kinematic observables for different characteristics depending on the origin of the b jet.
The second method pursues a more sophisticated approach using deep neural networks. The deep neural networks are trained to learn more intricate characteristics of the objects in order to apply various origin assignment strategies based on the networks' output. All methods will be evaluated and compared according to the metric defined at the beginning of the study.

This thesis is structured in eight chapters. This introduction forms the starting point. In Chapter 2, theoretical foundations of the SM are introduced and specified for physics at hadron colliders. The experimental view and environment, meaning the setup of the Large Hadron Collider and the CMS experiment, is addressed in Chapter 3. In Chapter 4, the event reconstruction methods for data analysis at the CMS experiment are explained. Chapter 5 lays the groundwork for the two studies building on it. In this chapter, the $t\bar{t}+b\bar{b}$ process is defined and delimited. Additionally, different levels of simulations are delineated to differentiate the two studies following in Chapter 6 and 7. In Chapter 6, the comparison of the simulations for events involving $t\bar{t}+b\bar{b}$ processes for ATLAS and CMS are performed. The different methods for assigning the b jets to their origin are evaluated in Chapter 7. A conclusion of the key findings rounds off the thesis in Chapter 8.

2 Theoretical foundations

In this chapter a theoretical basis for the studies in this thesis is given. Initially, in section 2.1 the Standard Model as the fundamental theory of particle physics is introduced and the elementary particles and interactions are described. This theoretically very vast subject is specified in section 2.2 for physics at hadron colliders, which also forms the transition to the experimental discussion and environment in Chapter 3.

2.1 The Standard Model

The Standard Model (SM) of elementary particle physics is a fundamental, relativistic and renormalizable quantum field theory [1–3, 7–19]. All known elementary particles and their interactions via three of the four fundamental forces are described in this theory. This includes the strong interaction, the weak interaction and the electromagnetic interaction, while gravity is not covered by the SM. The SM is a theory that is widely tested and in many facets validated [4, 5, 20–30]. The theoretical foundations of the SM are based on symmetry considerations of gauge theories, whose principles are summarized below. The elementary particles and their interactions arise from these gauge theories from a theoretical perspective. The following summary is based on [31, 32].

2.1.1 Gauge symmetries and interactions

In quantum field theories, particles are the interpretation of excitations of the fields. The state of a field is generally represented by a Lagrangian density, which is given by $\mathcal{L}\left(\phi(x), \partial_\mu \phi(x)\right)$ and thus the field-theoretical analog to the Lagrangian formalism. The Lagrangian density $\mathcal{L}$ is a function of the field $\phi(x)$ and the covariant derivative $\partial_\mu \phi(x)$ in dependence of the space-time coordinates x.

In the following, the conceptual approach from the starting point of a gauge theory to obtaining the Lagrangian density is demonstrated for the quantum field theory of electrodynamics, referred to as Quantum Electrodynamics (QED). Free fermions, i.e. particles with spin-1/2, obey the Dirac equation. The corresponding Lagrangian density of a free fermion can be written as

$$\mathcal{L}_{\text{fermion}} = \bar{\psi}(x)\left(i\gamma^\mu \partial_\mu - m\right)\psi(x) \quad , \tag{2.1}$$

where $\psi(x)$ denotes the Dirac spinor, $\bar{\psi}(x)$ the Dirac adjoint defined as $\bar{\psi}(x) = \psi^\dagger(x)\gamma^0$, γ^μ the corresponding Dirac matrices and m the fermion mass. This Lagrangian density is

already invariant under a global unitary transformation $U(1)$, as can be proven easily [31]. However, demanding the invariance of the system under a local $U(1)$ transformation, some substitutions are required. Carrying out the calculations, the substitution

$$\partial_\mu \rightarrow D_\mu = \partial_\mu + iqA_\mu \tag{2.2}$$

is obtained, where A_μ denotes a gauge field. The arising constant q is a consequence of the continuous symmetry, which leads to the conservation of a quantity according to Noether's theorem [33]. For this field the transformation

$$A_\mu(x) \rightarrow A_\mu(x) - \frac{1}{q}\partial_\mu \alpha(x) \tag{2.3}$$

with the space-time dependent phase $\alpha(x)$ is required, in order to satisfy the invariance of the system under the local gauge symmetry. In this case A_μ is the vector field of the gauge boson and understood as the field of a photon. The Lagrangian density for the dynamics of A_μ can be expressed via $\mathcal{L} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu}$ where $F_{\mu\nu}$ denotes the electromagnetic field tensor. The field tensor $F_{\mu\nu}$ is associated with the photon field A_μ via $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. Thus, the Lagrangian density of QED becomes

$$\mathcal{L}_{\mathrm{QED}} = (i\bar{\psi}\gamma^\mu\partial_\mu\psi - m\bar{\psi}\psi) - q\bar{\psi}\gamma^\mu A_\mu\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} \quad . \tag{2.4}$$

The four terms can be distinguished as follows. The first two terms are already known from equation 2.1, which are the Lagrangian density of the free fermion including a mass term. For instance, the fermion could be an electron with its mass $m = m_e$. The third term describes the interaction of the fermion ψ with the photon field A_μ via a coupling strength proportional to q, which can be identified as the electric charge. The last term denotes the propagation of photons as described. As a matter of fact, massless photons arise from the Lagrangian density of the QED since no mass term occurs for A_μ. Indeed, a mass term would break the demanded symmetry.

The concept shown for QED can be extended to more sophisticated theories like quantum chromodynamics (QCD) [11]. However, in contrast to the procedure shown for QED, non-Abelian groups are applied. For QCD, the special unitary group $SU(3)$ is used. This group consists of eight generators, which finally lead to eight different gauge bosons of QCD called gluons. Furthermore, three charges exist, which are called color charges and are name-giving (chromo) for QCD. The three color charges are called red, green and blue. In contrast to QED, gluons carry the charge associated with their quantum field theory, which is not the case for photons. In quantum field theories, quantum fluctuations arise, leading to the energy dependence of the coupling constants, which is referred to as running coupling. Together with the self-interaction of the gluons, this results in the color confinement and asymptotic freedom properties of QCD. The Lagrangian density of the QCD is given by

$$\mathcal{L}_{\mathrm{QCD}} = \sum_q \bar{\psi}_{q,i}(i\gamma^\mu\partial_\mu - m_q)\psi_{q,i} - g_\mathrm{s}\left(\bar{\psi}_{q,i}\gamma^\mu T^a_{ij}\psi_{q,j}\right)A^a_\mu - \frac{1}{4}F^a_{\mu\nu}F^{\mu\nu}_a \quad , \tag{2.5}$$

where first two terms denote the Lagrangian density of a free quark q with mass m_q. The third term describes the quark-gluon coupling proportional to the coupling g_s. The eight generators of the group $SU(3)$ are T^a_{ij}, which are represented by the Gell-Mann matrices. The fourth term denotes the gluon propagation and self-interaction with $F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + g_\mathrm{s} f^{abc} A^b_\mu A^c_\nu$ and the structure constants f^{abc}. Analogous to QED, there are no mass terms for the gauge bosons in this equation, therefore the gluons are massless as are the photons.

Establishing a quantum field theory for weak interactions leads to several challenges. Since only left-handed particles and right-handed anti-particles interact weakly, which was experimentally proven, the theory must be a chiral theory [20, 21]. Moreover, it was found that the gauge bosons of the weak interaction are massive [25–27]. The solution lies in the unification of the weak interaction with the electromagnetic interaction and the principle of spontaneous symmetry breaking as described in the Higgs mechanism [1–3].

The symmetry group of the electroweak interaction is $SU(2)_L \times U(1)_Y$. The weak isospin is introduced as a chiral characteristic in the group $SU(2)_L$, which applies only to left-handed particles. For the $U(1)_Y$ gauge group, the hypercharge Y is introduced. In addition, three gauge bosons are obtained from the $SU(2)_L$, which are massless due to symmetry conservation. Similar to the $U(1)$ group of QED, a massless gauge boson arises from analogous considerations for the $U(1)_Y$ group.

In order to obtain the experimentally manifested masses for the three gauge bosons, the Higgs mechanism is applied [1–3]. This mechanism introduces a complex scalar field ϕ and the Lagrangian density

$$\mathcal{L}_{\text{Higgs}} = (D_\mu \phi)^\dagger (D^\mu \phi) - \mu^2 \phi^\dagger \phi - \lambda (\phi^\dagger \phi)^2 \quad , \tag{2.6}$$

where the parameters are specified as $\lambda > 0$ and $\mu^2 < 0$. The relation between the two parameters is given by the vacuum expectation value $v = \sqrt{-\mu^2 / (2\lambda)}$ of the field ϕ after the choice of the ground state. This causes a spontaneous symmetry breaking of the electroweak gauge group and eventually leads to a scalar boson with mass $m_H = \sqrt{2\lambda v^2}$, which is called the Higgs boson. Via this mechanism, the previously massless gauge bosons mix to form the observable bosons $W^\pm$ and Z^0 with masses $m_W = \frac{1}{2} g v$ and $m_Z = \frac{1}{2}\sqrt{g^2 + g'^2}\, v$ after the symmetry breaking, whereas the photon remains massless. Hence, the mass terms depend on the vacuum expectation value and the coupling constants g and g' of the gauge group.

The masses of the fermions are given by the Yukawa interaction of the form

$$\mathcal{L}_{\text{Yukawa}} = -y_{\text{f}}\, \bar{\psi}_{\text{L}}(x) \phi(x) \psi_{\text{R}} \quad , \tag{2.7}$$

where y_{f} is the Yukawa coupling of the fermion f, L/R are the chiral components of the spinor and ϕ is the scalar field [34]. The mass relation to the vacuum expectation value and the Yukawa-like coupling between a fermion and the Higgs boson reads $m_{\text{f}} = \frac{1}{\sqrt{2}} y_{\text{f}} v$.

2.1.2 Elementary particles

Some elementary particles resulting from the theoretical considerations of gauge symmetries have already been discussed in the previous section. In this section, the particles of the SM will be considered from a more experimental perspective. Although the theoretical descriptions in the previous section define masses with respect to other quantities, e.g. the fermion masses or the Higgs boson mass, their exact value has to be determined experimentally since they remain as free parameters in the SM. The determination of these parameters, for example the measurement of the Yukawa coupling strength of different particles, is an important test of the SM.
The bosons and their key characteristics are already theoretically introduced and are summarized in Table 2.1. Altogether there are four spin-1 bosons of the electroweak theory, which are called γ, $W^\pm$ and Z^0. Only the two $W^\pm$ bosons are electrically charged. The $W^\pm$ bosons have a mass of about 80 GeV, the Z^0 boson is more massive with about 91 GeV [35]. The eight gluons of the QCD interact via the strong force and carry color charge. Gluons are electrically neutral, massless and spin-1 bosons similar to the intermediate particles of

Table 2.1: Standard Model gauge bosons and the Higgs boson. Gauge bosons are vector bosons with spin-1, the Higgs boson is a scalar boson with spin-0. The data is taken from [35]

Boson	Spin	Interaction	Charges	Mass
Gluons (g)	1	strong	color	massless
Photon (γ)	1	electromagnetic		massless
$W^{\pm}$ boson	1	electroweak	electric and weak	(80.379 ± 0.012) GeV
Z^0 boson	1	electroweak	weak	(91.188 ± 0.002) GeV
Higgs boson (H)	0		weak	(125.10 ± 0.14) GeV

the electroweak sector. The Higgs boson is a spin-0 particle, does not carry any charge and is the heaviest boson with about 125 GeV.

Fermions, particles with spin-1/2, are divided into two groups called leptons and quarks. An overview of the fermions is given in Table 2.2. Leptons include the electron (e), muon (μ) and tauon (τ), all three being electrically charged and differ in their masses. Associated to these leptons are neutrinos, which are correspondingly called electron neutrino, muon neutrino and tau neutrino. The three neutrinos are electrically neutral and have a small mass that has not been accurately measured so far.

The group of quarks also consists of six different elementary particles, which are called up (u), down (d), charm (c), strange (s), top (t) and bottom (b). They are electrically charged in thirds and carry color charge. The mass spectrum is spread over many orders of magnitude, from about 2 MeV for the up quark to about 173 GeV for the top quark. The properties of the top quark are discussed more detailed in the following section.

2.1.3 Top quark physics

The top quark is the heaviest particle of the SM and its lifetime is $5 \cdot 10^{-25}$ seconds [35]. Consequently, the lifetime is so short that the top quark does not hadronize but decays instead. More than 99 % of the top quark decays result in a b quark and a W boson.
The transition of quarks into other quarks is described by the CKM formalism. In this formalism, the transition into a b quark and a W boson is strongly favored compared to a transition into an s quark and a W boson or a c quark and a W boson for the top quark. Also the mass difference to other quarks is very large, which therefore results in a short lifetime.

The top quark transition is characterized by the W boson decay due to the approximately uniform decay of the top quark. The W boson can decay into one of the three leptons and a neutrino ($W \to \ell\nu_\ell$ with $\ell = e, \mu, \tau$) or into a pair of one of the five remaining quarks and an antiquark of different flavor ($W \to q\bar{q}'$) [35].
In this thesis, events with pairs of two top quarks ($\mathrm{t\bar{t}}$) are considered, resulting in different final states. In the dilepton channel, both W bosons decay into charged leptons and neutrinos. However, charged lepton in this case refers only to electron and muon, since the tauon can further decay hadronically and thus forms a special category which is not directly taken into account in the analyses presented in this thesis. If one of the two W bosons decays into a quark-antiquark pair while the other decays leptonically, it is referred to as the single-lepton channel. In the third case both W bosons decay into quark-antiquark pairs, which is called all-hadronic channel. The last category includes

Table 2.2: Standard Model fermions. Fermions are spin-1/2 particles and categorized in two groups. The first group is represented by the 6 quarks, the second group by the 6 leptons. Antiparticles are not demonstrated. The masses range over many orders of magnitude, from less than 1.1 eV to more than 10^{11} eV. No uncertainty is given for the electron and muon since it is many orders of magnitude smaller than the given value. All data is extracted from [35], except for the neutrinos [36].

Fermion	Interaction	Electric charge	Mass
up quark (u)	strong, electroweak	$+2/3$	$2.2^{+0.5}_{-0.3}$ MeV
down quark (d)	strong, electroweak	$-1/3$	$4.7^{+0.5}_{-0.2}$ MeV
charm quark (c)	strong, electroweak	$+2/3$	(1.27 ± 0.02) GeV
strange quark (s)	strong, electroweak	$-1/3$	93^{+11}_{-5} MeV
top quark (t)	strong, electroweak	$+2/3$	(172.76 ± 0.30) GeV
bottom quark (b)	strong, electroweak	$-1/3$	$4.18^{+0.03}_{-0.02}$ GeV
electron neutrino (ν_e)	weak	0	< 1.10 eV
electron (e)	electroweak	-1	0.51 MeV
muon neutrino (ν_μ)	weak	0	< 1.10 eV
muon (μ)	electroweak	-1	105.60 MeV
tau neutrino (ν_τ)	weak	0	< 1.10 eV
tauon (τ)	electroweak	-1	(1776.9 ± 0.1) MeV

all decays involving tauons, which can be further distinguished depending on the decay in dedicated analyses. This results in branching fractions of about 5 % for the dilepton channel, 30 % for the single-lepton channel, 44 % for the all-hadronic channel, and 21 % for the tauon category [35].

2.2 Hadron collider physics

To study elementary particles, interactions and structures as well as to test the SM at high energies, hadron colliders are a suitable choice. The following two sections deepen the theoretical foundations specific to research at hadron colliders.

2.2.1 Cross sections

In a scattering process, the interaction rate is defined by

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \sigma \cdot L_{\text{inst.}} \quad , \tag{2.8}$$

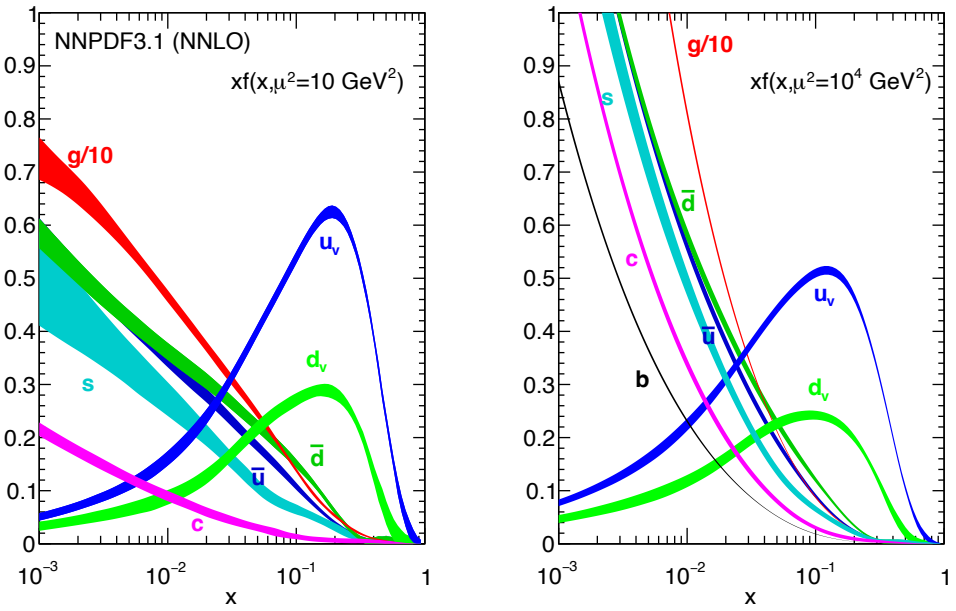
where σ is called cross section and $L_{\text{inst.}}$ is the instantaneous luminosity. The cross section is a quantity of the probability for the occurrence of a process, the instantaneous luminosity is a measure of the performance of a hadron collider and essentially describes how many collisions there are per time and area. Hence, the product of the two quantities is the interaction rate per unit of time.
The instantaneous luminosity will be discussed in the context of the experimental environment for the Large Hadron Collider in section 3.1. The theoretical calculation of cross sections is based on Fermi's golden rule

$$\sigma \propto |\mathcal{M}|^2 \cdot \rho \quad , \tag{2.9}$$

where $|\mathcal{M}|^2$ is the quantum mechanical transition amplitude and ρ is the phase space [37]. In jargon $|\mathcal{M}|^2$ is also called (squared) matrix element (ME). The ME is calculated from $|\mathcal{M}|^2 = |\langle\psi_\mathrm{f}|V|\psi_\mathrm{i}\rangle|^2$ for a given process from the initial state $|\psi_\mathrm{i}\rangle$ to the final state $\langle\psi_\mathrm{f}|$ using the interaction potential V from the Lagrangian density. The calculations are then performed in perturbation theory, whose visualizations are the well-known Feynman diagrams [38].

2.2.2 Physics of partons

In an experimental environment of a hadron collider, the initial condition of the collision of two hadrons, in the following illustrated for two protons, is not well-defined. Protons are composite objects and consist of smaller components called partons. In the case of a proton, it consists of two up quarks and one down quark, which are also called valence quarks. These quarks can emit and absorb gluons, which in turn can split into more quark-antiquark pairs via the strong interaction according to the laws of QCD. Thus, there is an extensive structure in the proton, which can be resolved and studied via deep inelastic scattering processes. The deep inelastic scattering measurements are mainly performed using electron-proton collisions, since the electron does not have an inner structure [39]. The results are called parton distribution functions (PDF), which describe the probability density for finding a particle with a longitudinal proton momentum fraction x at a resolution scale Q^2. The obtained PDFs are universal, which allows them to be transferred to other collision processes with protons, yet they are energy dependent (Q^2). Figure 2.1 shows PDF sets at two different energy scales. At low energy in Figure 2.1a it can be seen how the valence quarks dominate the high momentum fractions, whereas at



(a) Lower energy scale $Q^2 = 10\,\mathrm{GeV}^2$

(b) Higher energy scale $Q^2 = 10^4\,\mathrm{GeV}^2$

Figure 2.1: Example PDF sets as a function of the momentum fraction x at two different energy scales ($\mu^2 = Q^2$): in Figure 2.1a at a lower energy scale ($10\,\mathrm{GeV}^2$), in Figure 2.1b at a higher energy scale ($10^4\,\mathrm{GeV}^2$) [40] are shown. At both energies, the valence quarks dominate the high momentum fractions while the gluons dominate the lower scale. The gluon distributions are divided by a factor of 10 and thus dominate to a significantly stronger extent at low Q^2 than visually represented. However, the distributions are shifting at high energies, the probabilities for finding the valence quarks diminish and the sea quarks as well as the gluons become relevant. In fact, even the b quark can be noticed (black). The PDF sets will be used in Chapter 6.

high energies (Figure 2.1b) also the sea quarks as well as the gluons become relevant at high momentum fractions. Although the PDFs cannot be determined from first principles and are determined from deep inelastic scattering, renormalization group equations can be used to describe the evolution of the quark and gluon PDFs with respect to Q^2. The mathematical representation of these renormalization group equations is formulated in the Dokshitzer-Gribov-Lipatov-Altarelli-Parisi (DGLAP) equations [41]. The resulting PDFs are applied for the description of the proton constituents in hadron colliders, as explained in the following.

For collisions of protons in a hadron collider at high energies, two cases must be differentiated. The collision energy of the two partons is high, resulting in small values of the coupling constant α_S. However, other sub-processes in the proton occur at small energies and therefore at large values of the coupling constant α_S. Only at small coupling constant values perturbation theory can be applied. The concrete implications for the calculation and a simulation will be specified in section 5.2, at this point only the general concept is introduced. The distinction between the two cases, namely soft interactions at large coupling constants and hard interactions at small coupling constants, is accomplished via the factorization theorem. The factorization theorem introduces a new scale μ_F that separates the two cases. With the help of this approach, the hard partonic process can now be calculated in perturbation theory, while sub-processes such as soft and collinear gluon radiation are included in the PDFs.

For a proton-proton collision process into an arbitrary state X, the QCD factorization is given as

$$\sigma_{pp\to X} = \sum_{jk} \int dx_j dx_k f_j\left(x_j, \mu_\mathrm{F}^2\right) f_k\left(x_k, \mu_\mathrm{F}^2\right) \cdot \hat{\sigma}_{\hat{p}_j \hat{p}_k \to X}\left(x_j p_1, x_k p_2, \mu_\mathrm{F}^2, \alpha_S\left(\mu_\mathrm{R}^2\right)\right) \quad , \tag{2.10}$$

where $f_{j,k}$ denote the PDFs of the two partons $\hat{p}_{j,k}$ with a momentum fraction $x_{j,k}$ in the two protons with the momenta $p_{1,2}$. The hard process of the partons calculated in perturbation theory is $\hat{\sigma}$. It can be seen from the equation how the hard partonic process is eventually weighted with the PDFs and summed for both partons to finally obtain an observable cross section. The coupling constant α_S is a function of the renormalization scale μ_R. Similar to the factorization scale μ_F, the renormalization scale μ_R is introduced to resolve artifacts of pertubative QCD. However, the origin is different. Due to vacuum polarization effects in QCD resulting from the color charge of gluons, there are likewise renormalization group equations for α_S. The reference scale for the evolution of α_S is μ_R, therefore $\alpha_S\left(\mu_\mathrm{R}^2\right)$ depends on this scale. These scales must be chosen, resulting in uncertainties due to the effect of the chosen scale and calculations in finite orders in perturbation theory. Both scales μ_R and μ_F and their particular choice will become relevant in the studies for dedicated event simulations in Chapter 6.

3 Experimental environment

In this chapter the experimental environment of the CERN Large Hadron Collider and the CMS experiment is briefly introduced. The chapter starts with the introduction of the particle accelerator in 3.1. In section 3.2, the main components of the detector are described. The chapter concludes with a description of key kinematic quantities for physics at hadron colliders (section 3.3).

3.1 The Large Hadron Collider

The Large Hadron Collider (LHC) is the most powerful accelerator at the site of the European Organization for Nuclear Research (CERN) [42]. The LHC has a circumference of 27 km and is the last element in a chain of accelerators. Since the LHC is a synchrotron and not designed to accelerate particles at low energies, a series of pre-accelerators are needed. The entire complex can be seen in Figure 3.1. Protons and lead nuclei are accelerated to center-of-mass energies of up to 14 TeV and 5.6 TeV/nucleon, respectively. In the most recent run (Run-II), the LHC was operated at a center-of-mass energy of 13 TeV [43]. Initially, protons are obtained by ionization of hydrogen. Thereafter, the protons are injected into the first accelerator, the linear accelerator LINAC2. Already at this point the protons are bunched and subsequently accelerated to energies of 50 MeV. Next, the protons pass through the Proton Synchrotron Booster and reach an energy of about 1.4 GeV. In the adjacent Proton Synchrotron, the protons obtain an energy of 25 GeV before being further accelerated to 450 GeV in the Super Proton Synchrotron [44]. Ultimately in the LHC, the proton bunches can be accelerated in opposite directions in two different beam pipes to 7 TeV each.
Besides the center-of-mass energy, the expected number of particle collisions per time interval is the most important characteristic of a particle collider. This quantity is defined by the instantaneous luminosity, which is given by

$$L_{\text{inst}} = f_{\text{rev}} \cdot \frac{n_b N_b^2}{4\pi\sigma_x\sigma_y} \quad , \tag{3.1}$$

where f_{rev} describes the revolution frequency, N_b the number of particles per bunch and n_b the number of bunches per beam. In the denominator of the equation σ_x and σ_y describe the width of the protons' spatial distribution in a bunch perpendicular to the beam axis. The LHC is designed to reach an instantaneous luminosity of $10^{34}\,\text{cm}^{-2}\text{s}^{-1}$ [44].

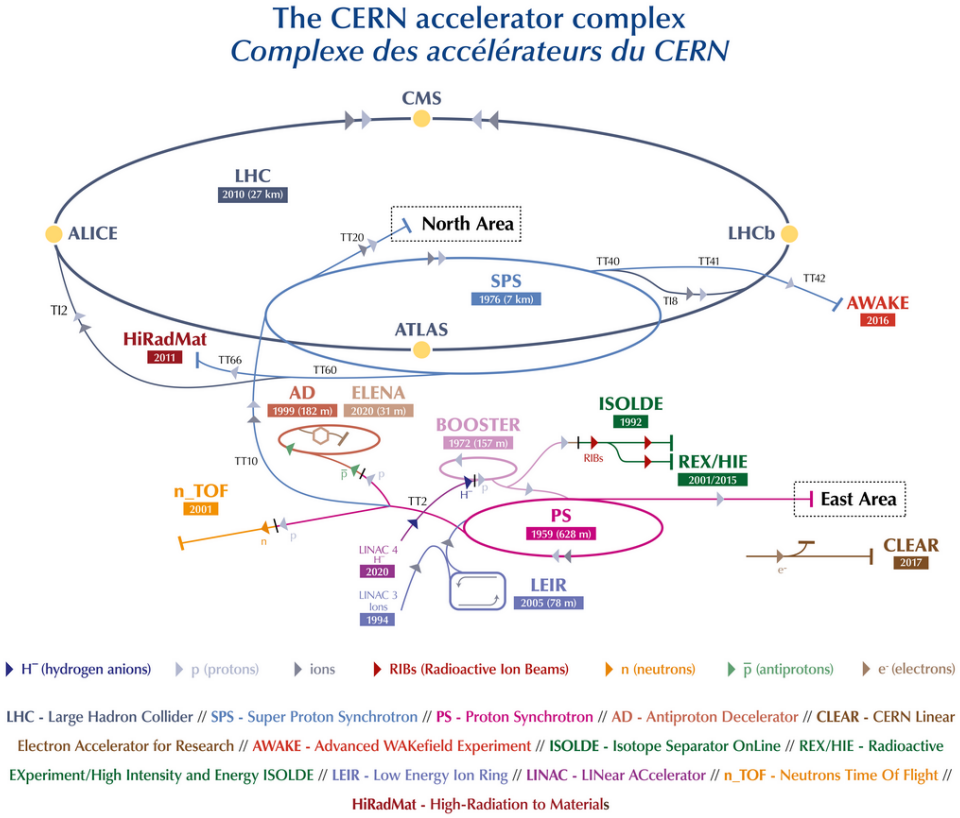


Figure 3.1: The CERN accelerator complex, taken from [42]. The Large Hadron Collider is the last element in the chain of accelerators (dark blue). Also, the experiments at CERN are indicated at the respective positions of the accelerators (yellow dots).

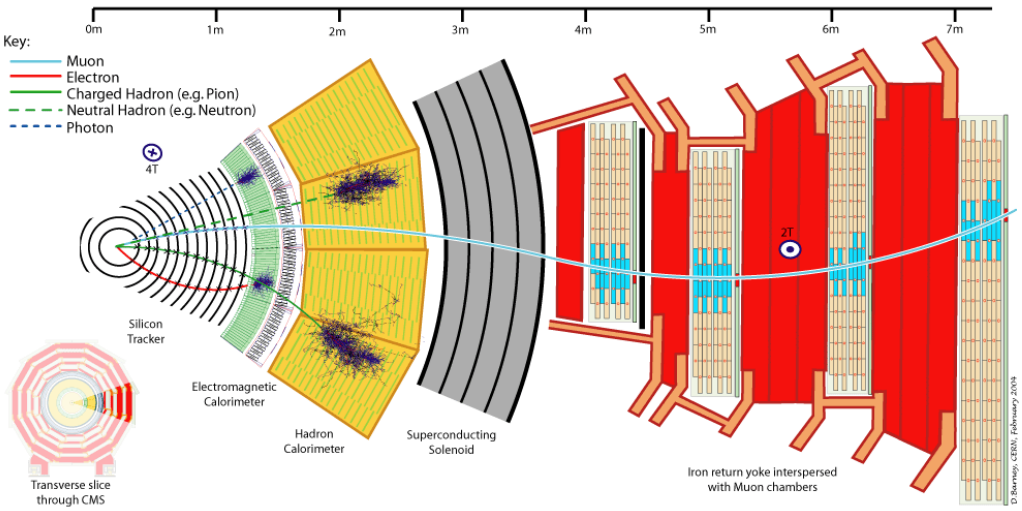


Figure 3.2: A transverse slice through the CMS experiment [49]. On the left side the collision point can be seen from which exemplary particles emerge in the transverse direction. The particles penetrate the various subdetectors with different reach depending on the particle type.

The particles collide at four different positions of the LHC, which are also shown in Figure 3.1. The detectors of the four experiments ALICE [45], ATLAS [46], CMS [47] and LHCb [48] are located at these positions. The ALICE detector is specialized on heavy-ions and analyses the strong-interaction sector of the SM. The LHCb experiment is a general purpose detector in forward direction and follows a versatile physics program. The ATLAS and the CMS experiment are also general purpose detectors with a comprehensive scope, but surround the entire collision point with an enclosed detector. Although the two experiments focus on a similar program, the two detectors differ significantly in the design of each individual detector component. In section 3.2 the main components of the CMS detector are briefly explained.

3.2 The Compact Muon Solenoid experiment

As a multi purpose detector, the Compact Muon Solenoid (CMS) experiment has to meet a number of requirements to detect a variety of particles at different momenta and energies. For this reason, the detector is built in an onion-like shell structure around the beam pipe of the LHC. Different detector layers aim at different goals, which are outlined from inside to outside in the following. Unless otherwise stated, the summary is based on the detailed description in [47].
A slice of the CMS detector with all layers is shown in Figure 3.2.

The **tracking system** is placed in the innermost section of the detector. The role of the tracking system is to determine the trajectories of charged particles. For this purpose, it is

located as close as possible to the interaction point. Through this approach the position of particle collisions and possible secondary vertices can be identified (see section 4.1). Since the tracker has to meet unique requirements for granularity, speed and radiation hardness, it is made of silicon. The tracking system consists of four layers of silicon pixel detectors and several layers of silicon strip detectors. If a charged particle successively passes through the layers of the tracking system, the charge depositions in each layer (''hits'') must subsequently be combined to form a coherent trajectory.
Since the tracking system is placed within a strong magnetic field, the transverse momentum of a charged particle is given by

$$p_{\mathrm{T}} = mv = qBr \quad . \tag{3.2}$$

In this case q is the electric charge of the particle, B is the magnetic flux density and r denotes the bending radius of the trajectory [50]. With geometrical considerations, the transverse momentum can be determined through the measurement of the sagitta s and the relation

$$s = \frac{qBl^2}{8p_{\mathrm{T}}} \quad , \tag{3.3}$$

where l denotes the path length in the magnetic field [35].

The **electromagnetic calorimeter** (ECAL) is dedicated to measure the energy of electrons/positrons and photons. It is a homogeneous calorimeter and is composed of lead tungstate ($PbWO_4$) crystals. Mainly two processes occur in the ECAL resulting in electromagnetic showers. At high energies, electrons and positrons deposit their energy in the ECAL mainly through bremsstrahlung (e.g. $e^- \rightarrow \gamma e^-$). Photons with an energy greater than twice the electron mass can generate positrons and electrons in the material via pair production (e.g. $\gamma \rightarrow e^+e^-$). If the energy is no longer sufficient for the aforementioned processes, only ionization takes place in the detector material. Finally, photodetectors (here avalanche photodiodes) are used to infer the deposited energy. The energy resolution can be parameterized as

$$\left(\frac{\sigma_E}{E}\right)^2 = \left(\frac{a}{\sqrt{E}}\right)^2 + \left(\frac{b}{E}\right)^2 + c^2 \quad , \tag{3.4}$$

where a is the stochastic term, b the detector noise term and c the constant term due to non-uniformity of longitudinal light collection, intercalibration errors and leakage of energy. The unit of the energy E is GeV. The parameters a, b and c are material dependent and can determined in beam tests.

The **hadron calorimeter** (HCAL) is the key to measuring the energy of particles with strong force interactions. This sub-detector is a sampling calorimeter and is composed of alternating layers of brass and scintillators. Since hadrons have a longer nuclear interaction length than the corresponding radiation length of electromagnetic interacting particles, strong interacting particles deposit their energy mainly in the HCAL. Charged hadrons can also start showering in the ECAL, however, this effect is small due to the small hadronic stopping power of the ECAL. If electrons are produced during the processes in the HCAL, electromagnetic subcascades may arise in the HCAL. Neutral hadrons can only be detected in the HCAL, since they do not leave a signature in the tracking system or ECAL. For the reconstruction of neutrinos and exotic particles, the HCAL plays a crucial role, since these particles can only be deduced indirectly from missing transverse energy (see section 4.5).

The **superconducting solenoid** encases all previous detector elements. The magnet is designed to achieve a magnetic flux density of up to 4 T. The setup of the magnet differs substantially from the ATLAS detector [46]. The presence of the magnetic field allows to

determine the particle's sign of electric charge and to reconstruct the transverse momentum through the curvature as described above.

The **muon system** is responsible for the identification, momentum measurement and triggering of muons. Muons at the LHC are produced in an energy regime in which they deposit only a small amount of energy in the form of ionization [35]. For this reason they are also called minimum-ionizing particles. Thus, muons pass through all previous detector layers with minimal interaction and are detected in the dedicated muon system. The muon system is located in the return yoke of the solenoid, which has a magnetic flux density of 2 T.

The **trigger system** of the detector is crucial to handle the LHC's high data rates of over 1 GHz in proton-proton collisions [51]. Since the data rates can neither be stored nor every single event contains relevant information for the analysis, the CMS detector is equipped with a two-level trigger system. The first level trigger (L1) is a hardware trigger and limits the output rate down to 100 kHz. The L1 trigger rapidly selects events that contain promising candidates, for instance ionization deposits that are consistent with a muon. The second level is the high-level trigger (HLT). The HLT is a software trigger and reduces the rate to 400 Hz for offline event storage. For each event, objects are reconstructed that might be interesting for the ensuing analysis. The HLT runs on a large processor farm [51].

3.3 Kinematic quantities

The origin of the CMS detector's coordinate system is defined to be in the center of the collision point. The x-axis points radially inwards to the center of the LHC, the y-axis points perpendicularly upwards. This determines the direction of the z-axis, which points along the beam axis towards the Jura mountains at the French-Swiss border. Given the detector's geometry and the rotational invariance of particle collisions, cylindrical coordinates are established. As the azimuthal angle ϕ is measured from the x-axis in the x, y-plane, the radial coordinate in this plane is designated as r. Starting from the z-axis, the polar angle θ is measured.
The rapidity, a quantity of velocity is defined by

$$y = \frac{1}{2} \ln \frac{E + p_z}{E - p_z} \quad , \tag{3.5}$$

where p_z denotes the momentum along the z-axis. For $m \ll p$, the rapidity can be substituted by the more convenient pseudo rapidity. The pseudo rapidity η is used to specify the direction of a particle, which is associated with the angle θ via the relation

$$\eta = -\ln \tan \left(\frac{\theta}{2} \right) \quad . \tag{3.6}$$

Thus, small values for η point perpendicular to the z-axis, while large values point towards the z-axis.
The spatial distance ΔR_{ij} between two objects with pseudo rapidity $\eta_{i,j}$ and azimuthal angle $\phi_{i,j}$ is defined by

$$\Delta R_{ij} = \sqrt{(\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2} \tag{3.7}$$

in the η, ϕ-plane and plays an important role in Chapter 6 and 7.
Since the momenta of the colliding partons in the center-of-mass frame are unknown, transverse quantities in the x, y-plane are defined. For example, the transverse momentum is given as

$$p_{\mathrm{T}} = \sqrt{p_x^2 + p_y^2} \quad . \tag{3.8}$$

Another important quantity is the Missing Transverse Energy, abbreviated as MET. Since certain particles, e.g. neutrinos and exotic neutral particles, cannot be reconstructed by the detector, MET provides indications to the presence of these particles in an event. This is possible since all other particles are detected, leaving neutrinos and exotic particles. A more detailed description of this quantity is given in section 4.5 in the context of the object definition.

4 Object and event definition

In Chapter 3 it was demonstrated how distinct particles deposit energy in each detector component. However, the energy depositions must be translated into physics objects for analyses. The primary reconstruction methods are presented in the following.

4.1 Track and vertex reconstruction

As described in section 3.2, the innermost layer of the detector is the tracking system. The goal of track reconstruction is to combine hits into a coherent trajectory of a particle passing the tracker. The CMS experiment applies an adaptation of the combinatorial Kalman filter called Combinatorial Track Finder (CTF) [52]. A detailed description of the CTF can be found in [53].
Tracking is based on a computationally expensive and iterative process. First, a dedicated *seed generation* is used to identify initial track candidates. Second, a Kalman filter-based *track finding* technique is used to calculate seed trajectories along the expected flight direction of the particle and to check whether hits are present in the corresponding region. Due to the magnetic field, the track of the charged particle is helix-shaped. Third, a *track fitting* module is used, which determines the best parameters for each trajectory. Finally, the quality of track candidates is evaluated by a *track selection* and bad candidates are discarded according to pre-defined criteria [53].
Once the tracks of the particle candidates in the tracking system have been reconstructed, it is possible to extrapolate the tracks to the origin of the proton-proton collision in the beam pipe. This allows to reconstruct vertices. The simultaneous proton-proton collision of two bunches can produce more than 50 interaction vertices. This effect is known as pileup. The vertex with the highest p_{T}^2 sum is set as primary vertex. Further vertices away from collision region are an indication of particle decays.

4.2 Particle-flow

Figure 3.2 shows the different sub-detectors of the CMS and the signal generation of specific particles. For each collision, the final-state particles must be identified from the generated signals. In order to combine the entire information from all sub-detectors, the particle-flow algorithm is applied [54].

Electrons are reconstructed through the ECAL and a trajectory in the tracker. For this purpose, a track matching the energy deposition in the ECAL must be found in the tracker.

Accordingly, the energy and momentum of electrons can be measured. Moreover, photons from bremsstrahlung in the tracker are considered.

Photons are also reconstructed in the ECAL. Photons deposit their energy in the ECAL as do electrons, but unlike the latter they do not have a track in the first sub-detector. However, the production of electron-positron pairs from photons in the tracker must be taken into account.

Muons are reconstructed in the muon system. In addition, it is verified whether a matching track can be found in the tracking system.

Charged hadrons are reconstructed through energy depositions in the ECAL and HCAL as well as an compatible track in the tracking system.

Neutral hadrons deposit their energy exclusively in the HCAL. After all previous signatures have been reconstructed to physics objects, the remaining energy depositions in the HCAL are assigned to neutral hadrons.

4.3 Jet reconstructions

Due to the confinement of QCD, color-charged particles such as quarks and gluons are not allowed to exist freely at lower energies. Therefore, color-neutral particles are created through hadronization immediately after the formation. A parton from the hard scattering process thus generates a number of other color-neutral particles in the direction of flight, which is called a particle jet. Reaching the HCAL, the already color-neutral hadrons of the particle jet can, for example, radiate additional gluons through hadronic excitation. The anti-k_T algorithm is applied at CMS in order to combine all particles of a common flight direction to a jet [55]. Due to the correlated initial flight direction, the particles approximately form the shape of a cone.
The anti-k_T algorithm fulfills two important requirements for a jet algorithm: infrared and collinear safety. Infrared safety implies the invariance of a jet under radiation of low-energy gluons. If the jet remains unchanged despite the radiation of gluons at small angles, the jet is collinear safe. The anti-k_T algorithm uses the distance measure

$$d_{ij} = \min\left(p_{\mathrm{T},i}^{-2}, p_{\mathrm{T},j}^{-2}\right) \frac{\Delta R_{ij}^2}{R^2} \tag{4.1}$$

between two particles i and j. In the equation ΔR_{ij} denotes the spatial distance from eq. 3.7. A radius parameter of $R = 0.4$ is commonly applied in CMS analyses to identify jets in MC simulation and data. The distance between a particle i and the beam is defined by

$$d_{iB} = p_{\mathrm{T},i}^{-2} \quad . \tag{4.2}$$

The anti-k_T algorithm clusters the two particles whose d_{ij} is smallest into a new pseudo particle. From eq. 4.1 it can be seen that a hard particle and a soft particle are clustered before combining two soft particles at identical distance. Accordingly, clustering soft particles with hard particles is preferred over clustering two soft particles. With this algorithm, cone-like jets are obtained [55].

4.4 b tagging

Jets resulting from B hadrons possess unique characteristics that distinguish them from other jets. Essentially, B hadrons have a slightly longer lifetime compared to other hadrons. As a result, B hadrons travel off the collision point before their decay [35]. Theoretically,

the extended lifetime is motivated by the suppressed CKM matrix elements that describe the possible transitions of the b quark. Experimentally, this is reflected by a secondary decay vertex shifted with respect to the primary interaction vertex. Technically, the identification of a b jet in this thesis is realized via the DeepJet algorithm [56]. The process of identifying jets as b jets is called b tagging. The DeepJet algorithm uses deep neural networks (DNNs) with 1D convolutional layers on a number of features of the physics objects from the particle-flow algorithm. The concept of DNNs is described in section 7.3.2. Three of the outputs nodes of the multiclassification DNN are then combined into a value between 0 and 1, which is interpreted as the b jet-ness of the jet candidate. This value is called b tag value. A medium working point of 0.277 is specified for the data sets analyzed in this thesis [57]. The medium working point results from the definition of 1 % mistag efficiency. This means a tagging efficiency of about 79 % for DeepJet for data taken in 2018 [56]. As a consequence, if the b tag value of a jet is equal to or greater than 0.277, it is assumed to be a b jet.

4.5 MET

Solely weakly interacting, electrically neutral particles produced in the collision cannot be observed by the detector. Among these undetectable particles are not only neutrinos, but also hypothetical exotic neutral particles such as Dark Matter candidates. Nevertheless, the presence of the particles can be inferred indirectly by calculating the sum of transverse momenta. Via the considerations of the laws of conservation of energy and conservation of momentum a statement

$$\text{MET} = \left| - \sum_{\text{detected}} \vec{p}_{\text{T},i} \right| = \left| \sum_{\text{undetected}} \vec{p}_{\text{T},i} \right| \quad , \tag{4.3}$$

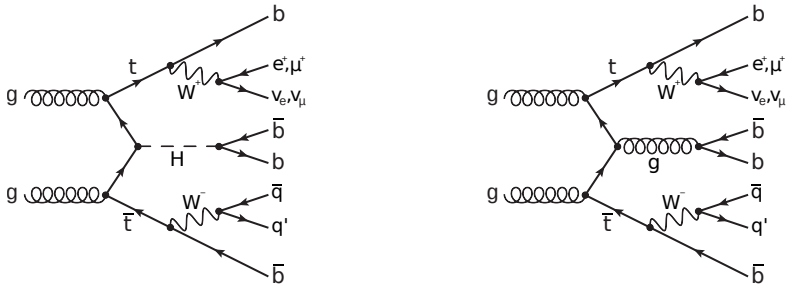
about the energy of the undetectable particles can be made [58]. There are two important aspects to be taken into account: First, the transverse energy of all particles must be reconstructed entirely. Second, it is unknown among how many undetectable particles the missing transverse energy is distributed. Hence, MET depends strongly on sub-detector resolutions, mismeasurements of all reconstructed particles and detector artifacts.

5 $t\bar{t}$+$b\bar{b}$ production

The beginning of this chapter outlines the necessity of studying the $t\bar{t}$+$b\bar{b}$ process. Subsequently, section 5.2 describes the event generation process and delineates the simulation levels for the later studies. Section 5.3 defines the event topologies and concludes with the $t\bar{t}$+$b\bar{b}$ definition. This chapter forms the foundation for the analyses presented in Chapter 6 and 7.

5.1 Motivation

As already discussed in Chapter 2, the measurement of the coupling of the Higgs boson to the top quark with a Yukawa-type interaction is an important test of the SM. It also constrains models of physics beyond the SM (BSM) which predict different coupling strengths. The associated production of a top quark-antiquark pair and a Higgs boson ($t\bar{t}$+H) allows a direct probe of the top-Higgs Yukawa coupling and is the most favorable production mode for a direct measurement. Even if this production mode represents only approx. 1 % of the total Higgs boson production cross section, the top quarks have a distinguishable signature and several Higgs boson decay channels can be accessed [59]. The decay of a Higgs boson into a bottom quark-antiquark pair with a leading-order (LO) branching ratio Br $\left(\mathrm{H} \to \mathrm{b\bar{b}}\right)$ of about 58 % is the largest of all decay modes [60]. However, with a predicted inclusive cross section of approx. 0.30 pb, a $t\bar{t}$+H and $\mathrm{H} \to \mathrm{b\bar{b}}$ process (commonly abbreviated as $t\bar{t}\mathrm{H}(\mathrm{b\bar{b}})$) at a center-of-mass energy of $\sqrt{s} = 13\,\mathrm{TeV}$ is rare compared to other processes at the LHC [60]. Both ATLAS and CMS performed searches for the SM Higgs boson in the $t\bar{t}\mathrm{H}(\mathrm{b\bar{b}})$ channel in proton-proton collisions at $\sqrt{s} = 13\,\mathrm{TeV}$ [59, 61]
Another process that also contains a $b\bar{b}$ pair in association with a top quark-antiquark pair but without a Higgs boson is the $t\bar{t}$+$b\bar{b}$ process [62]. This process leads to the same final state, but has an approx. eight times larger predicted inclusive cross section of 2.27 pb [60]. In section 2.1.3 the different channels of $t\bar{t}$ decays were introduced. This thesis focuses on the single-lepton channel. At LO, two b quarks originating from the top quark decay, two light flavor quarks, one charged lepton and one neutrino are expected from the $t\bar{t}$ system in this channel. The single-lepton channel is a good compromise, since events can be triggered well due to the lepton, in contrast to the all-hadronic channel, and since it has a higher branching fraction, compared to the dilepton channel. Both the $t\bar{t}\mathrm{H}(\mathrm{b\bar{b}})$ and the $t\bar{t}$+$b\bar{b}$ processes in the single-lepton channel are shown in Figure 5.1 where the identical initial and final states can also be seen very well.



(a) $t\bar{t}H(b\bar{b})$ process (b) $t\bar{t}+b\bar{b}$ process

Figure 5.1: Examples of LO Feynman diagrams for the $t\bar{t}H(b\bar{b})$ and $t\bar{t}+b\bar{b}$ processes.

Besides, the $t\bar{t}+b\bar{b}$ process is not only an essential background for the $t\bar{t}H(b\bar{b})$ process. The $t\bar{t}+b\bar{b}$ process itself is particularly interesting from a theoretical and experimental perspective, because two processes occur at different QCD scales [6]. The underlying $pp \to t\bar{t}$ process appears at scales around 500 GeV. This process is accompanied by the production of b jets, which takes place on scales of a few ten GeV [63]. Although the respective QCD NLO calculations are available for both processes, the two different scales lead to large uncertainties in the choice of the factorization and renormalization scale [6].

To summarize, the $t\bar{t}+b\bar{b}$ process is a large irreducible background in $t\bar{t}H(b\bar{b})$ searches at the LHC which are important to test the SM and to constrain BSM physics. Furthermore, the $t\bar{t}+b\bar{b}$ process itself is of particular interest due to its multiscale QCD nature. Hence, a detailed $t\bar{t}+b\bar{b}$ study is crucial.

5.2 Event generation and simulation levels

The production of simulated events follows a sequence of mathematical and technical procedures to successively embed a variety of physical phenomena until the events are described as detailed as necessary. The theoretical foundations of QCD factorization described in section 2.2.2 constitute the starting point of the simulation process. Accordingly, the matrix elements (ME) of the hard scattering process of the partons in the proton are calculated first. These calculations are computed to a fixed order of perturbation theory, for example LO or next-to-leading-order (NLO), to account for higher-order QCD and electroweak corrections. Already in the definition of the ME there exist several ways to model a given process. Among other details, this difference in modeling is examined for different generators in Chapter 6. According to the factorization theorem, the initial states of the partons in the ME can now be sampled from the PDFs. The PDFs describe the probability densities of finding the partons with certain longitudinal proton momentum fractions x at a given energy scale Q^2 in the proton.
Now, the final-state partons of the ME are further processed and QCD fragmentation in the form of parton showers (PS) as well as initial state radiation (ISR) are simulated. These PS are caused by coherent radiation of collinear gluons and their further splitting into quark-antiquark pairs. With the decrease in the PS energy and the resulting increase of the strong coupling constant α_S, the hadronization processes must then be simulated. As the PS energy scale drops below 1 GeV, the strong coupling constant becomes $\alpha_S \approx 1$. Thus, it is no longer possible to apply perturbation theory. As an alternative, phenomenological models for hadronization are required. There are two major models for hadronization, the Lund string model [64] and the cluster model [65]. The hadronization finally results in color-neutral hadrons.

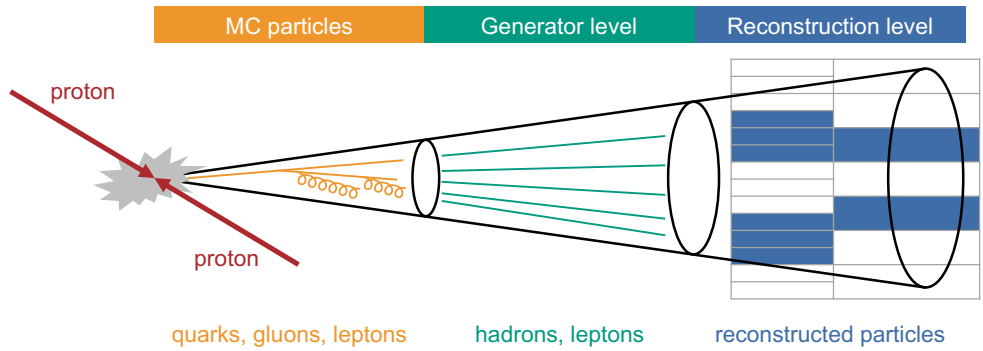


Figure 5.2: Visualization of different simulation levels of a jet. Generator level is shown in the middle of the figure, reconstruction level is shown on the right side. Exemplary physics objects are correspondingly listed at the bottom.

The full simulation of events of a given process consequently follows a multi-level simulation chain. In this thesis physics objects are analyzed at two different simulation levels. These two distinct simulation levels are introduced and defined in section 5.2.1 and 5.2.2 to clearly distinguish them before the actual analyses. The two levels will then be analyzed and discussed individually in Chapter 6 and 7. The two simulation levels are based on the generated MC particles. The exact calculation is specific to the used MC code, which is described in detail for some MC generators in [66]. The first level of the simulation, which is studied in this thesis, is the generator level simulation. Sometimes the term ''MC truth'' is also used to describe this level, since detector effects are not incorporated. The second level is the reconstruction level. At reconstruction level the simulated events are also processed through detector simulations and thus reflect a prediction of real data. The decisive difference can be noticed already at this point: while the generator level contains the true information about the event (i.e. final state partons), but does not correspond to the objects measurable in an event, reconstruction level does not contain this true information, but includes detector effects. Yet these events should not be confused with real reconstructed data from actual data taking. The differences of the simulation levels are also demonstrated in Figure 5.2.

5.2.1 Events at generator level

Considering events on generator level allows for a comparison between various MC event generators, simulated data from different experiments and theory predictions. Such a comparison is performed in Chapter 6 using a comprehensive framework.
At generator level, an event is characterized by its remaining particles after hadronization. Even if these particles are termed stable in this context, this merely means that they do not decay before reaching the detector. The following generator level object definitions are based on [67].
Prompt charged leptons are particles arising from electroweak interactions, i.e. not associated with hadrons or τ leptons. To define prompt charged leptons a jet algorithm is used to cluster all photons in close proximity to the lepton. With this approach QED radiation effects are considered. The contribution of ISR photons in this case is negligibly small. Hence, mainly FSR photons remain. On a technical level, all photons are added to the lepton's four momentum to take photon radiation into account. For this, an anti-k_{T} algorithm is used as described in section 4.3. Typically, the radius parameter is set to a small value of $R = 0.1$ in this case. Additionally, for the analysis in Chapter 6 and 7, prompt charged leptons are only considered if their p_{T} passes a $p_{\mathrm{T,min}}$ threshold and their

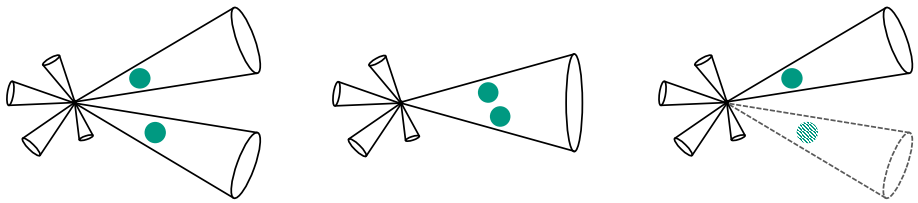
pseudorapidity η is within the interval $|\eta| < |\eta_{\text{max}}|$.
Jets are similarly defined at generator level using the anti-k_{T} algorithm. In this procedure all generator level particles are collected into a cone according to the algorithm. Subsequently, prompt electrons, muons and neutrinos that are not associated to hadrons are removed. Finally, the remaining stable particles are clustered. Many analyses use a radius parameter of $R = 0.4$. Furthermore, for the analysis jets must pass a p_{T} threshold to be selected and their pseudorapidity must lie in an interval of $|\eta| < |\eta_{\text{max}}|$.
If the full event history is available, it is possible to assign a jet flavor on generator level. This is achieved via a method called ghost-association or ghost-matching [68]. In this method, the B hadrons' momenta are scaled to infinitely small values before the jet clustering. These super soft B hadrons are then clustered into the jet as ''ghosts''. As the jet algorithm is infrared and collinear safe, this does not affect the actual jet clustering. The occurrence of such a ghost then determines the flavor of the jet which now can be labeled as a b jet.

5.2.2 Events at reconstruction level

The simulation level considered so far remains theoretical, since events at generator level cannot be associated with real data from the experiment. In order to make this possible, the response of the particles in the CMS detector must also be simulated, which is done in the detector simulation. This level is called reconstruction level and the corresponding events are discussed in Chapter 7. The physics objects are defined according to the definitions in Chapter 4. Hence, a direct assignment of the particles in an event between the two simulation levels is not seamlessly possible and the true information from generator level is no longer available. To solve this issue and still allow an assignment between the two simulation levels, a matching algorithm is introduced in Chapter 7.

5.3 Event topologies and $t\bar{t}$+$b\bar{b}$ definition

The $t\bar{t}$+$b\bar{b}$ process mentioned above is a part of the superset $t\bar{t}$+jets of all $t\bar{t}$ processes. By definition, $t\bar{t}$+jets events are divided in three mutually exclusive topologies, depending on the flavor of the generator level hadrons found within the jets that are not associated with the $t\bar{t}$ system (called additional jets in the following). The definition follows the method described in [69]. Events that contain at least one additional b jet are generally denoted as $t\bar{t}$+B while events which contain at least one additional c jet but no b jets are denoted as $t\bar{t}$+$c\bar{c}$. Finally, all other events which consequently contain only light-flavour jets or no additional jets at all are called $t\bar{t}$+lf.
Furthermore, three cases for $t\bar{t}$+B events are distinguished, namely $t\bar{t}$+$b\bar{b}$, $t\bar{t}$+2b and $t\bar{t}$+b, which are depicted in Figure 5.3. All of them require two B hadrons each in addition to the $t\bar{t}$ system. In this thesis, the term $t\bar{t}$+$b\bar{b}$ *selection* will refer to the case in Figure 5.3a, i.e. a selection of events on generator level that contains a $t\bar{t}$ system and two additional b jets, meaning that the two additional B hadrons are assigned to two separate jets. However, in the $t\bar{t}$+2b case, the two B hadrons are very close to each other such that they are both assigned to the same jet. As in the first case, the two B hadrons in the $t\bar{t}$+b case are sufficiently far apart from each other. Therefore, they are not assigned to the same jet, but one of the two jets is now out of acceptance.
The $t\bar{t}$+$b\bar{b}$ case should not be confused with $t\bar{t}$+$b\bar{b}$ as a name for simulated data samples in Chapter 6 and 7, because physics objects in the $t\bar{t}$+$b\bar{b}$ sample have no such pre-selection applied on generator level, but only two additional b jets at ME level.



(a) $t\bar{t}+b\bar{b}$ (b) $t\bar{t}+2b$ (c) $t\bar{t}+b$

Figure 5.3: Illustration of $t\bar{t}$+B events. The $t\bar{t}$ system is not shown in this visualization, the differences only pertain to the additional jets. Figure 5.3a shows two separate jets, each of them contains a B hadron (green circle). Figure 5.3b shows both B hadrons very close to each other which are therefore merged into a single jet. Figure 5.3c also shows two separate jets with a B hadron in each jet, but one jet is out of acceptance (dashed).

6 Generator level study

For the simulation of events with $t\bar{t}$+$b\bar{b}$ processes, different MC generators as well as settings may be used by the ATLAS and CMS collaborations. In this chapter, different simulation strategies and settings of existing simulations are described and compared at generator level. For this purpose, the setting for the study is initially described in sections 6.1 to 6.5. Based on this, exclusively CMS simulations are analyzed in section 6.6. This is followed by a comparison between ATLAS and CMS simulations which are performed in the context of the LHC Higgs working group [70] in section 6.7. The chapter concludes with a summary in section 6.8.

6.1 Objectives and approach

In section 5.1, the necessity of a study of the $t\bar{t}$+$b\bar{b}$ process was presented. In this chapter, different available MC simulations are compared with each other at generator level. The goal is to identify differences in the modeling and in the associated uncertainties. This comparison is first performed for $t\bar{t}$ and $t\bar{t}$+$b\bar{b}$ simulations generated by the CMS collaboration corresponding to the experimental conditions of the data taken in the year 2018 during the LHC Run-II. In a second analysis, the CMS simulations are compared with the $t\bar{t}$ and $t\bar{t}$+$b\bar{b}$ simulations of the ATLAS experiment. This fosters the design of a common strategy between ATLAS and CMS for background modeling uncertainties in $t\bar{t}H(b\bar{b})$ measurements.

Based on these objectives, the approach is as follows. First, the existing simulations are described at a technical level and the main differences in parameters (configurations) are pointed out. Second, an object and event selection is defined for the analysis. Third, distinct (kinematic) features of the generated events (validation observables) are defined by which the simulations will be compared. Fourth, the analysis routine is written in a framework that provides an interface between the experiments. Fifth, a visual comparison between the simulations is performed with histograms.

6.2 $t\bar{t}$ and $t\bar{t}$+$b\bar{b}$ simulations

In the following sections, the simulation strategies studied in this thesis are described from a technical point of view. In particular, the MC generators and shower programs used are addressed and the parameter values applied are stated. Likewise, key differences in

modeling are highlighted. The content of this chapter explicitly does not include a detailed description of the functioning of the respective simulation programs for the generation of the events. Furthermore, not every parameter choice is motivated, since this is beyond the scope of this thesis and follows the work of the authors of the simulations. The details can be found in the references provided. Rather, the focus in the following is on describing the exact settings of the $t\bar{t}$ and $t\bar{t}+b\bar{b}$ simulation environments and the comparison based on the validation observables (section 6.4). The CMS simulations are described initially, followed by a brief description of the differences between these simulations and the ATLAS simulations.

6.2.1 $t\bar{t}$ Powheg+Pythia8 simulation

The first simulation approach is called $t\bar{t}$ POWHEG+PYTHIA8 simulation. This abbreviation refers to the MC event generator POWHEG v2 interfaced with PYTHIA8 [71–74]. The matrix elements (ME) calculation order is next-to-leading-order (NLO). Since only the top quarks and additional NLO radiation are calculated in the ME of the $t\bar{t}$ simulation approach, the additional b jets mainly originate from the parton shower (PS). The inner structure of the two protons is described by the parton distribution function (PDF) set NNPDF 3.1 at NLO [40]. The PS and the hadronization is generated by PYTHIA8.230 which produces a multi-particle final state from the hard scattering process [75]. PYTHIA8 comes with a whole set of parameters available for adjustments, which define the response of the modeling. The set of adjustable parameters is called tune. In this simulation approach the "CMS PYTHIA8 tune 5" (CP5) is applied. The detailed specifications of the CP5 tune are defined in [76]. The top quark mass is set to $m_t = 172.5\,\text{GeV}$. The h_{damp} parameter, which regulates the real emissions in POWHEG, is set to 1.379 times the top quark mass [77]. The renormalization and factorization scales are also determined by quantities of the top quark and scaled to

$$\mu_{R,F} = \sqrt{m_t^2 + p_{T,t}^2} \quad , \tag{6.1}$$

where m_t denotes the top quark mass and $p_{T,t}$ the transverse momentum of the top quark. The $t\bar{t}$ POWHEG+PYTHIA8 simulation approach has been used in $t\bar{t}H(b\bar{b})$ analyses published by CMS for the modeling of the irreducible $t\bar{t}$ background so far [61, 78].

The most important difference of this simulation approach compared to the following $t\bar{t}+b\bar{b}$ POWHEG+PYTHIA8 simulation approach is the description of the b quarks. Crucial for the effects observed in this study is where the additional b quarks, which are not associated with $t\bar{t}$ decay, stem from. These additional b quarks can originate either from ME calculations, or from PS.

The difference of the additional b quarks' origin can be implemented in two different ways, called 5-flavor scheme (5FS) and 4-flavor scheme (4FS) [79]. The primary difference between the two flavor schemes is whether or not the b quark is contained in the initial state in the proton, i.e. as a part of the PDF. In the 5FS, the b quark is one of the five quark flavors (u, d, c, s, b) present in the proton, which is indicated in the naming. In this representation, the b quarks are assumed to be massless. Thus, the b quarks can be included in the calculations in the initial state, simplifying them and allowing for resummation of large initial state logarithms into the b quark PDF as explained in the following. The $t\bar{t}$ POWHEG+PYTHIA8 simulation approach is performed in the 5FS.
The calculations in pertubation theory of the process are influenced by two different mass or energy scales, the hard scale Q and the b quark mass m_b. In these calculations logarithms and exponents with the quotient m^2/Q^2 occur. Particularly the occurring logarithm spoils the accuracy of the calculations in case of $Q \gg m$. These logarithms arise from gluon splitting into b quark-antiquark pairs, which can appear in the initial

or final state. The final state logarithm can be resummed in pertubative fragmentation functions, for example. The emerging difficulties of initial state logarithms can be solved by defining a b quark PDF. This is possible for large Q at small mass m_{b} of the b quarks. As a result, the logarithms are resummed to all orders in the strong coupling constant α_S via the Dokshitzer-Gribov-Lipatov-Altarelli-Parisi (DGLAP) evolution equations [80]. By assuming massless b quarks, $Q \gg m_{\mathrm{b}} = 0$ is accomplished.

The advantage of using the 5FS is smaller scale uncertainties. This results in more precise predictions for inclusive observables, e.g. total rates. The disadvantage is less accurate predictions for differential distributions. For this purpose, 4FS NLO calculations are applied instead [79]. The following $t\bar{t}$+$b\bar{b}$ simulation approaches are calculated in the 4FS.

In this and all other simulation approaches, events are generated in the dilepton channel as well as in the single-lepton channel (see section 2.1.3). Although this thesis focuses on the single-lepton channel (see section 5.1), the later analysis will demonstrate that events simulated in the dilepton channel can satisfy the selection of the routine due to lepton misidentification and thus influence the result, whereas the all-hadronic channel does not.

6.2.2 $t\bar{t}$+$b\bar{b}$ Powheg+Pythia8 simulation

The second simulation approach is referred to as $t\bar{t}$+$b\bar{b}$ Powheg+Pythia8 simulation. The MC generator used for this simulation approach is Powheg-Box-Res combined with OpenLoops [63, 81, 82]. The ME calculations are performed at NLO accuracy. In contrast to the $t\bar{t}$ simulation approach, the ME calculation does not only include $t\bar{t}$ production, but $t\bar{t}$+$b\bar{b}$. The PDF set used is NNPDF 3.1 at NLO with $\alpha_S = 0.118$, albeit in the 4FS. For the shower and hadronization, the identical Pythia configuration is adopted as for the $t\bar{t}$ Powheg+Pythia8 simulation approach. As before, the CP5 tune is applied and the top quark mass is set to $m_{\mathrm{t}} = 172.5\,\mathrm{GeV}$. Also, the h_{damp} parameter is set to $1.379 m_{\mathrm{t}}$ in this simulation approach, since it is governed by the tune. Due to the presence of two different scales and the different ME, the choice of $\mu_{\mathrm{R,F}}$ changes to

$$\mu_{\mathrm{R}} = \frac{1}{2} \cdot \sqrt[4]{m_{\mathrm{T,t}} \cdot m_{\mathrm{T,}\bar{\mathrm{t}}} \cdot m_{\mathrm{T,b}} \cdot m_{\mathrm{T,}\bar{\mathrm{b}}}} \quad \text{and} \quad \mu_{\mathrm{F}} = \frac{\sum (M_{\mathrm{T}})}{4} \quad , \tag{6.2}$$

where $\sum (M_{\mathrm{T}}) = m_{\mathrm{T,t}} + m_{\mathrm{T,}\bar{\mathrm{t}}} + m_{\mathrm{T,b}} + m_{\mathrm{T,}\bar{\mathrm{b}}} + p_{\mathrm{T},g}$ is the sum of the pertinent transverse quantities (see section 3.3). The scale choice incorporates the two widely differing scales for the $t\bar{t}$+$b\bar{b}$ production and geometrically averages the relevant scales of the top and b quark masses.

In case b quarks are considered together with the top quarks in the ME, the 4FS is the preferred choice for the simulation approach [63]. The b quarks are now massive and no longer part of the PDF set. The b quark mass is set to $m_{\mathrm{b}} = 4.75\,\mathrm{GeV}$. Following the reasoning in the previous section, the occurring logarithms can no longer be resummed with massive b quarks. The resulting calculation is handled by splitting the ME into three parts

$$\mathcal{M}_{t\bar{t}+b\bar{b}} = \mathcal{M}_{\mathrm{IS},t\bar{t}+b\bar{b}} + \mathcal{M}_{\mathrm{FS},t\bar{t}+b\bar{b}} + \mathcal{M}_{\mathrm{rem},t\bar{t}+b\bar{b}} \quad , \tag{6.3}$$

where $\mathcal{M}_{\mathrm{IS},t\bar{t}+b\bar{b}}$ and $\mathcal{M}_{\mathrm{FS},t\bar{t}+b\bar{b}}$ denote the initial and final state $g \to b\bar{b}$ splittings, respectively. It is found that the numerical impact of the remaining ME $\mathcal{M}_{\mathrm{rem},t\bar{t}+b\bar{b}}$ is negligible [63]. Despite a more complicated calculation and the impossibility of resummation, $t\bar{t}$+$b\bar{b}$ ME offer significant advantages. By describing the protons in the 4FS, the dependence on PS modeling can be minimized and $g \to b\bar{b}$ splittings are free of collinear singularities [63]. Due to the occurrence of two relevant scales in this process, larger scale uncertainties are expected.

6.2.3 $t\bar{t}$+$b\bar{b}$ Sherpa simulation

The $t\bar{t}$+$b\bar{b}$ SHERPA simulation approach is executed with the MC generator SHERPA 2.2.4 and OPENLOOPS at NNLO [83]. As indicated in the simulation name, this is a simulation approach with $t\bar{t}$+$b\bar{b}$ in the ME and thus structurally similar to the previous simulation approach. The ME calculations are realized at NLO. The general simulation approach in SHERPA is slightly different from those in the previous programs. In this approach, merging schemes are used in the simulation that consistently merge NLO ME of different multiplicities. This procedure requires specific algorithms that merge ME calculations of different orders with various multiplicities, namely LO and NLO, and subsequently allows them to be matched with the PS. As PDF set NNPDF 3.0 is used at NLO with $\alpha_S = 0.118$ [84]. The PS is also realized in SHERPA. The tune is the default setting according to the authors [85]. Both the top quark mass with $m_\mathrm{t} = 172.5\,\mathrm{GeV}$ as well as the b quark mass with $m_\mathrm{b} = 4.75\,\mathrm{GeV}$ are chosen as in the previous simulations. The renormalization scale and factorization scale are defined as

$$\mu_\mathrm{R} = \sqrt{m_{\mathrm{T,t}} \cdot m_{\mathrm{T},\bar{\mathrm{t}}} \cdot m_{\mathrm{T,b}} \cdot m_{\mathrm{T},\bar{\mathrm{b}}}} \quad \text{and} \quad \mu_\mathrm{F} = \frac{\sum (M_\mathrm{T})}{4} \quad . \tag{6.4}$$

Accordingly, while the renormalization scale is different compared to μ_R of the $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8 simulation approach, the factorization scale μ_F is the same.

6.2.4 $t\bar{t}$+$b\bar{b}$ MG5aMC(NLO) simulation

The last simulation approach is the $t\bar{t}$+$b\bar{b}$ MG5AMC(NLO) simulation. The MC generator used for this simulation approach is MADGRAPH5_AMC@NLO 2.4.2 [86]. As with the previous two simulation approaches, MG5AMC(NLO) also calculates $t\bar{t}$+$b\bar{b}$ in the ME. The ME calculations are processed at NLO. The multijet merging setup described for SHERPA is also implemented in this simulation program. The PDF set NNPDF 3.1 is used at NLO with $\alpha_S = 0.118$. The PS is implemented in PYTHIA as in the $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8 simulation approach. CP5 is uniformly used as tune. The top quark mass is set to $m_\mathrm{t} = 172.5\,\mathrm{GeV}$ as in the other simulation approaches, whereas the b quark mass with $m_\mathrm{b} = 4.70\,\mathrm{GeV}$ is slightly different. The renormalization scale and factorization scale are chosen differently in this simulation approach compared to the previous ones with

$$\mu_\mathrm{R,F} = \frac{\sum (M_\mathrm{T})}{2} \quad . \tag{6.5}$$

The main settings of the simulation approaches examined are condensed in Table 6.1. The scales used are summarized in Table 6.2.

6.2.5 ATLAS simulations

The individual settings of the ATLAS simulation approaches are not discussed at length as they were for the CMS simulation approaches. Essentially, ATLAS uses slightly different parameters or combines other versions of MC generators with different PDF sets that have already been discussed. The applied configurations are shown along with the CMS simulation approaches in Table 6.1 for direct comparison. Unlike the CMS simulations, no simulated events with a MG5AMC(NLO) simulation approach are available. Instead, another $t\bar{t}$ simulation approach is analyzed, in which events are generated with SHERPA where the ME for events with $t\bar{t}$ and zero or one additional jets are calculated at NLO while the ME for events with two to four additional jets are calculated at LO. Analogous to the CP5 tune from CMS, ATLAS uses its own tune called A14 [87]. The renormalization scales and factorization scales used are shown in Table 6.2 alongside the scales from CMS. ATLAS applies identical scales in the $t\bar{t}$ simulation approaches, but applies different scales in the $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8 and $t\bar{t}$+$b\bar{b}$ SHERPA simulation approach.

Table 6.1: Overview of the configurations used in the MC simulation approaches analyzed in this chapter. In section 6.6 the CMS simulated events are compared among each other, in section 6.7 the POWHEG+PYTHIA8 and SHERPA simulated events of ATLAS and CMS are compared in relation to the events simulated with $t\bar{t}$ POWHEG+PYTHIA8 from ATLAS. The column "process" describes which process is calculated in the ME.

	Process	MC generator	ME order	Shower	Tune	PDF set	m_t	m_b
CMS	$t\bar{t}$	POWHEG v2	NLO	PYTHIA8	CP5	5FS NNPDF 3.1 NLO	172.5 GeV	0
	$t\bar{t}$+$b\bar{b}$	POWHEG-BOX-RES	NLO	PYTHIA8	CP5	4FS NNPDF 3.1 NLO $\alpha_S = 0.118$	172.5 GeV	4.75 GeV
	$t\bar{t}$+$b\bar{b}$	SHERPA 2.2.4	NLO	SHERPA	default	4FS NNPDF 3.0 NNLO $\alpha_S = 0.118$	172.5 GeV	4.75 GeV
	$t\bar{t}$+$b\bar{b}$	MG5AMC(NLO) 2.4.2	NLO	PYTHIA8	CP5	4FS NNPDF 3.1 NLO $\alpha_S = 0.118$	172.5 GeV	4.70 GeV
ATLAS	$t\bar{t}$	POWHEG v2	NLO	PYTHIA8	A14	5FS NNPDF 3.0 NLO	172.5 GeV	0
	$t\bar{t}$+$b\bar{b}$	POWHEG-BOX-RES	NLO	PYTHIA8	A14	4FS NNPDF 3.0 NLO $\alpha_S = 0.118$	172.5 GeV	4.75 GeV
	$t\bar{t}$+$b\bar{b}$	SHERPA 2.2.1	NLO	SHERPA	default	4FS NNPDF 3.0 NNLO $\alpha_S = 0.118$	172.5 GeV	4.75 GeV
	$t\bar{t}$	SHERPA 2.2.1	tt+0,1@NLO+2,3,4@LO	SHERPA	default	5FS NNPDF 3.0 NNLO $\alpha_S = 0.118$	172.5 GeV	0

Table 6.2: Scale choices of the simulation approaches considered for ATLAS and CMS. The $t\bar{t}$ POWHEG+PYTHIA8 and $t\bar{t}$ SHERPA simulation approaches use identical scales. The $t\bar{t}+b\bar{b}$ POWHEG+PYTHIA8 and $t\bar{t}+b\bar{b}$ SHERPA simulation approaches use different scales in ATLAS and CMS, the differences are highlighted. $\sum(M_T)$ is defined as $\sum(M_T) = m_{T,t} + m_{T,\bar{t}} + m_{T,b} + m_{T,\bar{b}} + p_{T,g}$ and H_T as $H_T = \sum p_{T,t} + p_{T,\bar{t}} + p_{T,b} + p_{T,\bar{b}} + p_{T,g}$.

Simulation	Scale CMS	Scale ATLAS
$t\bar{t}$ POWHEG+PYTHIA8	$\mu_{R,F} = \sqrt{m_t^2 + p_{T,t}^2}$	$\mu_{R,F} = \sqrt{m_t^2 + p_{T,t}^2}$
$t\bar{t}+b\bar{b}$ POWHEG+PYTHIA8	$\mu_R = \frac{1}{2} \cdot \sqrt[4]{m_{T,t} \cdot m_{T,\bar{t}} \cdot m_{T,b} \cdot m_{T,\bar{b}}}$, $\mu_F = \frac{\sum(M_T)}{4}$	$\mu_R = 1 \cdot \sqrt[4]{m_{T,t} \cdot m_{T,\bar{t}} \cdot m_{T,b} \cdot m_{T,\bar{b}}}$, $\mu_F = \frac{\sum(M_T)}{2}$
$t\bar{t}+b\bar{b}$ SHERPA	$\mu_R = \sqrt[4]{m_{T,t} \cdot m_{T,\bar{t}} \cdot m_{T,b} \cdot m_{T,\bar{b}}}$, $\mu_F = \frac{\sum(M_T)}{4}$	$\mu_R = \sqrt[4]{m_{T,t} \cdot m_{T,\bar{t}} \cdot m_{T,b} \cdot m_{T,\bar{b}}}$, $\mu_F = \frac{H_T}{2}$
$t\bar{t}+b\bar{b}$ MG5AMC(NLO)	$\mu_{R,F} = \frac{\sum(M_T)}{2}$	no simulation
$t\bar{t}$ SHERPA	no simulation	$\mu_{R,F} = \frac{1}{4}\sqrt{m_{T,t}^2 + m_{T,\bar{t}}^2}$

6.2.6 Uncertainties

In addition to the nominal distributions, the incorporation of uncertainties is also part of the analysis. These uncertainties are variations of the renormalization scale μ_R and factorization scale μ_F defined above specifically for each simulation approach. Regardless of the scale definition for each simulation approach, the procedure for a scale variation is identical for all simulations. The renormalization scale and the factorization scale are varied independently from each other. One of the two scales is doubled or halved respectively, whereas the other one is fixed. This results in a total of four variations, which are $1\mu_R, 0.5\mu_F$; $1\mu_R, 2\mu_F$; $0.5\mu_R, 1\mu_F$ and $2\mu_R, 1\mu_F$. These variations are also called $\mu_{R,F}$ up/down variations. To account for the up/down scale variations of μ_R and μ_F in the distributions of the validation observables, in each bin it is checked to determine which two variations deviate up resp. down from the nominal value. The plotted scale uncertainty is then calculated by the sum of squares of the two independent variations

$$v_{\text{up/down}} = v_{\text{nominal}} \pm \sqrt{\Delta\mu^2_{\text{R,up/down}} + \Delta\mu^2_{\text{F,up/down}}} \quad , \tag{6.6}$$

where v denotes the value in a given bin and $\mu_{\text{R/F,up/down}}$ the scale variations listed above. No variations are plotted for the SHERPA simulated events since they could not be accessed.

In the comparisons of the CMS simulations with the ATLAS simulations (see section 6.7), the scale variations are calculated in a different way. In this calculation, the two scales are not varied independently from each other as before, but simultaneously. This results in only two variations, $0.5\mu_R, 0.5\mu_F$ and $2\mu_R, 2\mu_F$. These variations are also referred to as ME scale variations in the following.

In addition to the ME variations, also PS variations are considered. In the PYTHIA8 parton showering process, splitting processes occur that depend on the choice of α_S. Thus, the occurring additional gluon radiation at this stage of the simulation also depends on the renormalization scale, since α_S is associated with the renormalization scale (see section 2.2.2). PYTHIA8 distinguishes between a contribution of PS uncertainties arising from gluons from ME simulations and a contribution from gluons from the actual PS process. The uncertainties pertaining to the initial-state radiation (ISR) and final-state radiation (FSR) are used for the analysis in the default configuration ($\text{ISR}_{\text{up/down}}$, $\text{FSR}_{\text{up/down}}$) [88, 89].

6.3 Object and event selection

The object definition for comparing the different simulated events is based on the generator level definition described in section 5.2.1. The jets are defined according to the anti-k_T algorithm with a radius parameter of $R = 0.4$. In addition, the jets must exceed a p_T value of 25 GeV and lie within a pseudorapidity range of $|\eta| < 2.5$. The b jets are identified using ghost matching as described in section 5.2.1. The B hadrons are required to have a p_T value of at least 5 GeV. The leptons considered are electrons and muons. As a consequence, in the following leptons always denote electrons or muons. They must pass a p_T threshold of $p_T \geq 27$ GeV and, like the jets, must lie within a pseudorapidity interval of $|\eta| < 2.5$. The leptons are removed if they are found within a distance of $R < 0.4$ from a jet.

An event is selected for further analysis if exactly one lepton is present, corresponding to a selection for the single-lepton channel. In addition, at least four jets in the event must meet the above criteria. This selection is lower than the number of jets that would be expected at LO for a $\text{t}\bar{\text{t}}+\text{b}\bar{\text{b}}$ process in the single-lepton channel (see Chapter 5). Although six jets are expected at LO, the selection is chosen to account for the fact that not all jets

may meet the thresholds of the object selection. In addition to the selection of the number of jets, the events are divided into two categories. The first category additionally requires exactly 3 b jets. By analogous reasoning, one of the four b jets may not be identified as such. The second category corresponds to the signal region and additionally requires at least four b jets. In summary, the analysis is performed on two categories, ''1 lepton, ≥ 4 jets, 3 b jets'' and ''1 lepton, ≥ 4 jets, ≥ 4 b jets'', respectively. The motivation for introducing the two categories is the fact that in the ≥ 4 b jets case the full event with all b jets is present, whereas the 3 b jets case is better suited for the study of acceptance effects in related analyses, e.g. $t\bar{t}H(b\bar{b})$ analyses by CMS [61, 78].

6.4 Validation observables

For the comparison of the simulated events, quantities must be defined against which the actual comparison can be performed. These quantities are referred to as validation observables in the following. The first validation observable is the number of jets (N_{jets}) in an event, which is also referred to as jet multiplicity. The second validation observable is the number of b jets ($N_{\text{b jets}}$) in the event, which is a subset of N_{jets}.
In the observable p_{T} (all b jets) the individual p_{T} values of all b jets are recorded in an event. The observables p_{T} (highest p_{T} b jet), p_{T} (second-highest p_{T} b jet), p_{T} (third-highest p_{T} b jet) and p_{T} (fourth-highest p_{T} b jet) compare the corresponding b jet sorted by the p_{T} value of all b jets in an event.
The observable H_{T} is defined by the scalar sum of the transverse momenta of the examined particles

$$H_{\text{T}} = \sum_i p_{\text{T},i} \quad , \tag{6.7}$$

where i denotes the particle considered. In total, three different H_{T} observables are defined. H_{T} (jets) is the scalar sum of all jet p_{T} values in an event. H_{T} (jets+lepton) further adds the lepton's p_{T} value to the previous observable. H_{T} (b jets) only considers the scalar sum of the transverse momenta of all b jets in the event. The detailed study of the numerous observables, which investigate the p_{T} values of the jets and also of the lepton in different ways, allow the verification of whether certain simulation approaches tend to produce softer (lower momentum) or harder (higher momentum) objects.
The distance between b jets in the η, ϕ-plane according to equation 3.7 is analyzed in three observables. In the observable ΔR (average between b jets) is calculated as the arithmetic mean of ΔR between all pairs of b jets in an event. The observable $\Delta R\,(\text{bb})$ (closest) determines the distance between the two b jets that are closest to each other in an event, while $\Delta R\,(\text{bb})$ (leading) indicates the distance between the two b jets with the two highest p_{T} values. The analysis of the distance observables allows to investigate whether b jets tend to be closer to each other in simulation approaches with $t\bar{t}$ or $t\bar{t}$+$b\bar{b}$ in the ME.
Another validation observable is the invariant mass, which is determined for the two closest b jets ($m\,(\text{bb})$ (closest)) and for the two highest p_{T} b jets ($m\,(\text{bb})$ (leading)). This validation observable is an important observable in studies of $t\bar{t}H(b\bar{b})$ production for which $t\bar{t}$+$b\bar{b}$ is a crucial background (see Chapter 5). The invariant mass of the two b jets in the $H \to b\bar{b}$ process should correspond to the invariant Higgs boson mass, therefore also the background in this observable must be modeled accurately.
Similar to the invariant mass, the p_{T} value of the system with the two closest b jets ($p_{\text{T}}\,(\text{bb})$ (closest)) and the two hardest b jets ($p_{\text{T}}\,(\text{bb})$ (leading)) is also analyzed. The validation observables are summarized in Table 6.3.

Table 6.3: List of all validation observables used for the comparison of the $t\bar{t}$ and $t\bar{t}+b\bar{b}$ simulation approaches.

Observable	Description
N_{jets}	Number of jets in the event
$N_{\text{b jets}}$	Number of b jets in the event
All b jet p_{T}	p_{T} of all b jets in the event
Leading b jet p_{T}	p_{T} of b jet with largest p_{T} in the event
Sub-leading b jet p_{T}	p_{T} of b jet with second largest p_{T} in the event
Third b jet p_{T}	p_{T} of b jet with third largest p_{T} in the event
Fourth b jet p_{T}	p_{T} of b jet with fourth largest p_{T} in the event
H_{T}	Scalar p_{T} sum of all jets and lepton in the event
H_{T} (jets)	Scalar p_{T} sum of all jets in the event
H_{T} (b jets)	Scalar p_{T} sum of all b jets in the event
ΔR (bb) (average)	Average ΔR of all two b jet combinations in the event
ΔR (bb) (closest)	ΔR of the two b jets which are closest in ΔR in the event
ΔR (bb) (leading)	ΔR of the two largest p_{T} b jets in the event
m (bb) (closest)	Invariant mass of the two b jets closest in ΔR in the event
m (bb) (leading)	Invariant mass of the two largest p_{T} b jets in the event
p_{T} (bb) (closest)	p_{T} of the system with the two b jets closest in ΔR in the event
p_{T} (bb) (leading)	p_{T} of the system with the two largest p_{T} b jets in the event

6.5 Analysis routine

After the data sets have been described, the uncertainties specified, the object and event selection defined and the validation observables chosen, a suitable framework is required. In order to not only compare CMS internal simulations, but also to be able to compare them with ATLAS simulations, the framework Rivet is chosen for the analysis. A detailed description of the Rivet framework can be found in reference [90]. Rivet is an acronym for ''Robust Independent Validation of Experiment and Theory'' and provides all required features for a generator level study between CMS only simulations, but also with ATLAS simulations. In particular, Rivet version 3 and higher offers the possibility to include uncertainties [91]. The $t\bar{t}+b\bar{b}$ routine developed for this analysis is available in [92].

6.6 Comparison: CMS

In this section, the results of the analysis routine applied to the four CMS simulation approaches $t\bar{t}$ Powheg+Pythia8 (sec. 6.2.1), $t\bar{t}+b\bar{b}$ Powheg+Pythia8 (sec. 6.2.2), $t\bar{t}+b\bar{b}$ Sherpa (sec. 6.2.3) and $t\bar{t}+b\bar{b}$ MG5aMC(NLO) (sec. 6.2.4) are discussed. In the labels of the histograms Powheg+Pythia8 is abbreviated with PP8 and MG5aMC(NLO) is shortened to aMC. Since comparisons are performed for each validation observable in Table 6.3 for each of both categories, a complete discussion of all observables exceeds the limits of this thesis. The distributions of the observables in the ''1 lepton, ≥ 4 jets, 3 b jets'' region can be found in Appendix A. Also, all distributions of the ''1 lepton, ≥ 4 jets, ≥ 4 b jets'' category not discussed in this section can be found there.

The first validation observable is the number of jets in an event, which is shown in Figure 6.1. This figure, like any in this section, is divided into three parts. The upper part shows the distribution of the examined observable, normalized to an integral value of 1. In addition, each bin content is divided by the bin width to account for non-uniform

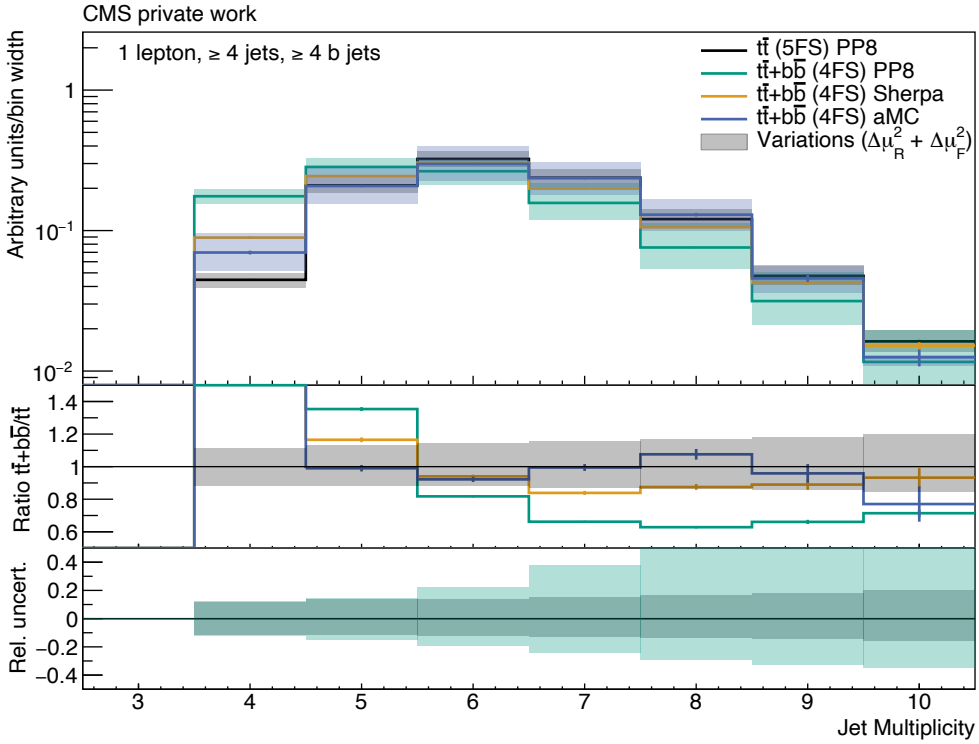


Figure 6.1: Jet multiplicity of the $t\bar{t}$ POWHEG+PYTHIA8 (black), $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8 (green), $t\bar{t}$+$b\bar{b}$ SHERPA (orange) and $t\bar{t}$+$b\bar{b}$ MG5-AMC(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_R and μ_F are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, $\geq$ 4 b jets".

bin widths, if applicable. Variations are shown as bands around the nominal distributions (except for events simulated with SHERPA) as explained in section 6.2.6. In the middle panel of the figure, the ratio of the distributions of the events simulated with all three $t\bar{t}$+$b\bar{b}$ simulation approaches relative to the $t\bar{t}$ simulated events is displayed. Additionally, the scale variation of the $t\bar{t}$ simulated events is shown as a gray uncertainty band. The lower panel of the figure compares the relative uncertainties of the distributions of the two POWHEG+PYTHIA8 simulation approaches. The other two distributions of the events simulated with the $t\bar{t}$+$b\bar{b}$ SHERPA and $t\bar{t}$+$b\bar{b}$ MG5AMC(NLO) simulation approaches are not included in this part for better visibility.

In Figure 6.1, it can be seen that the distributions of the jet multiplicity differ in all four simulation approaches. Nevertheless, the trends of the distributions of the events simulated with the three $t\bar{t}$+$b\bar{b}$ simulation approaches are similar with respect to the distribution of the $t\bar{t}$ simulated events. In the events simulated with the $t\bar{t}$+$b\bar{b}$ simulation approaches, there tend to be fewer jets in an event compared to the $t\bar{t}$ simulated events. This trend depends strongly on the simulated processes under review. If only a simulation of the single-lepton channel is examined instead of a more inclusive simulation (i.e. also the dilepton channel), a reverse trend can be observed. Even if the selection in the analysis requires precisely one lepton in the event and thus corresponds strictly to the single-lepton channel, events generated in the dilepton channel can have an impact on the result. If, for example, one of the two simulated leptons in the dilepton channel does not meet the required threshold values for leptons, this event may comply with the requirement ‘‘1 lepton, ≥ 4 jets, ≥ 4 b jets’’. The deviation of the distribution by adding or omitting the dilepton channel despite the explicit selection on the single-lepton channel in the analysis routine will be demonstrated later in this section.
In the lower part of Figure 6.1 it is shown that the variation of the events simulated with the $t\bar{t}$+$b\bar{b}$ simulation approach is much larger than the variation of the events simulated with the $t\bar{t}$ simulation approach as expected (see section 6.2.2). In addition, it can be seen that the scale variations are asymmetric.

The number of b jets, a subset of the previously analyzed jet multiplicity, can be seen in Figure 6.2. The b jet multiplicity tends to be higher in events of the $t\bar{t}$+$b\bar{b}$ simulation approach than in the events of the $t\bar{t}$ simulation approach. All distributions agree well with each other for exactly four b jets.

Examining the p_{T} values of all b jets in the event, no major effect can be observed as in the jet or b jet multiplicity. The distribution is shown in Figure 6.3. Starting from a b jet p_{T} of about 150 GeV the statistical uncertainty and thus statistical fluctuation of the distribution increases due to the limited number of simulated events. The effect is counteracted by increasing the bin width. In order to retain a smooth distribution, the bin contents are divided by the bin width as mentioned above.

The validation observable H_{T} of the jets, in contrast, differs remarkably in the simulated events of all considered MC generators. This is illustrated in Figure 6.4. The quantity H_{T} (jets) is defined according to equation 6.7 and thus correlated with the p_{T} values of all b jets in the event (Figure 6.3). It can be seen that for small values up to about 300 GeV of H_{T}, the three distribution of events with the $t\bar{t}$+$b\bar{b}$ simulation approaches predict significantly more events than the $t\bar{t}$ simulated events. Above about 400 GeV, the distributions of the events simulated with the $t\bar{t}$+$b\bar{b}$ SHERPA and POWHEG+PYTHIA8 simulation approach show a similar behavior with clearly smaller values for H_{T} (jets) compared to the events simulated with the $t\bar{t}$ POWHEG+PYTHIA8 simulation approach. Thus, H_{T} (jets) is distinctly shifted to smaller values for the events simulated with these two simulation approaches. This kind of consistent behavior is not observed for the events

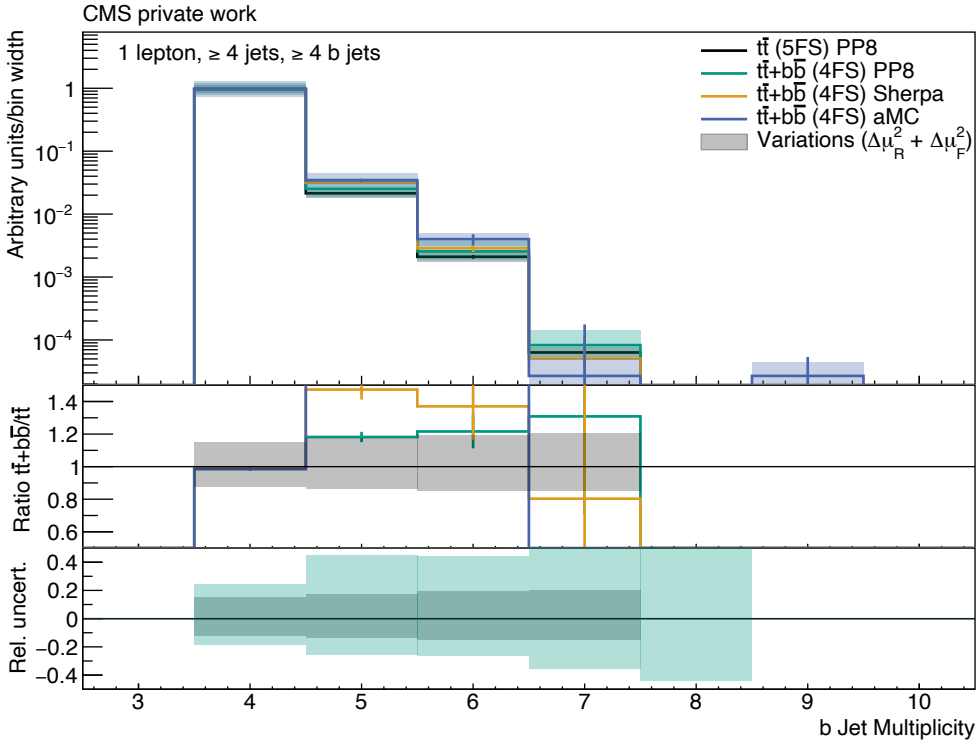


Figure 6.2: b jet multiplicity of the $t\bar{t}$ POWHEG+PYTHIA8 (black), $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8 (green), $t\bar{t}$+$b\bar{b}$ SHERPA (orange) and $t\bar{t}$+$b\bar{b}$ MG5-aMC(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_R and μ_F are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, $\geq$ 4 b jets".

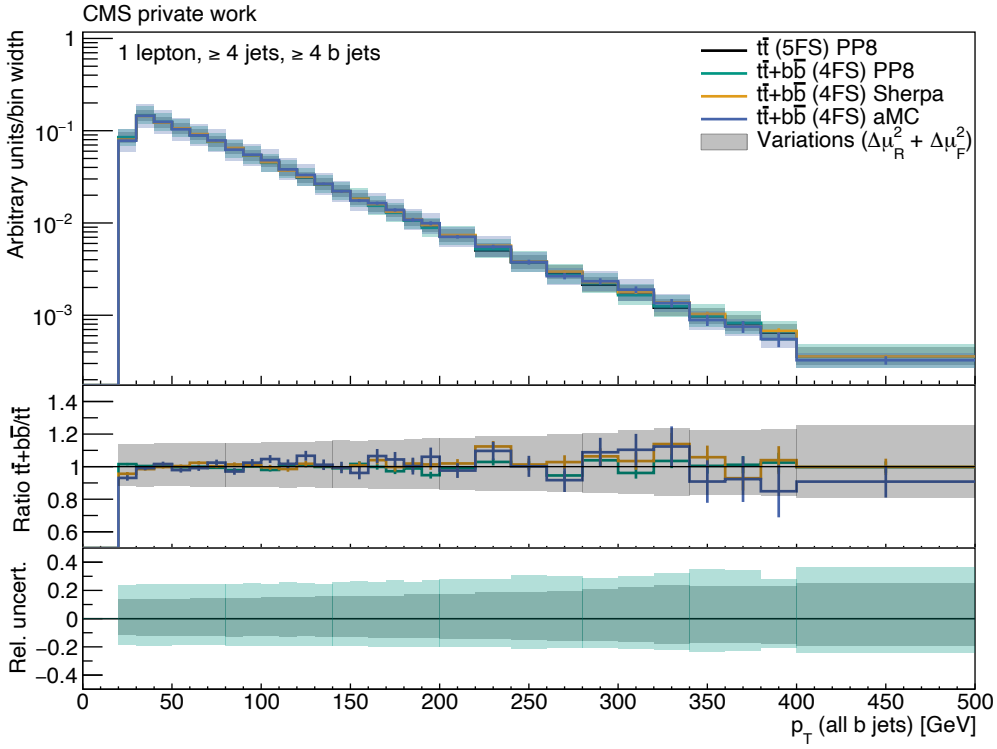


Figure 6.3: p_{T} (all b jets) of the $\mathrm{t\bar{t}}$ Powheg+Pythia8 (black), $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ Powheg+Pythia8 (green), $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ Sherpa (orange) and $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ MG5aMC(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_{R} and μ_{F} are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, $\geq$ 4 b jets".

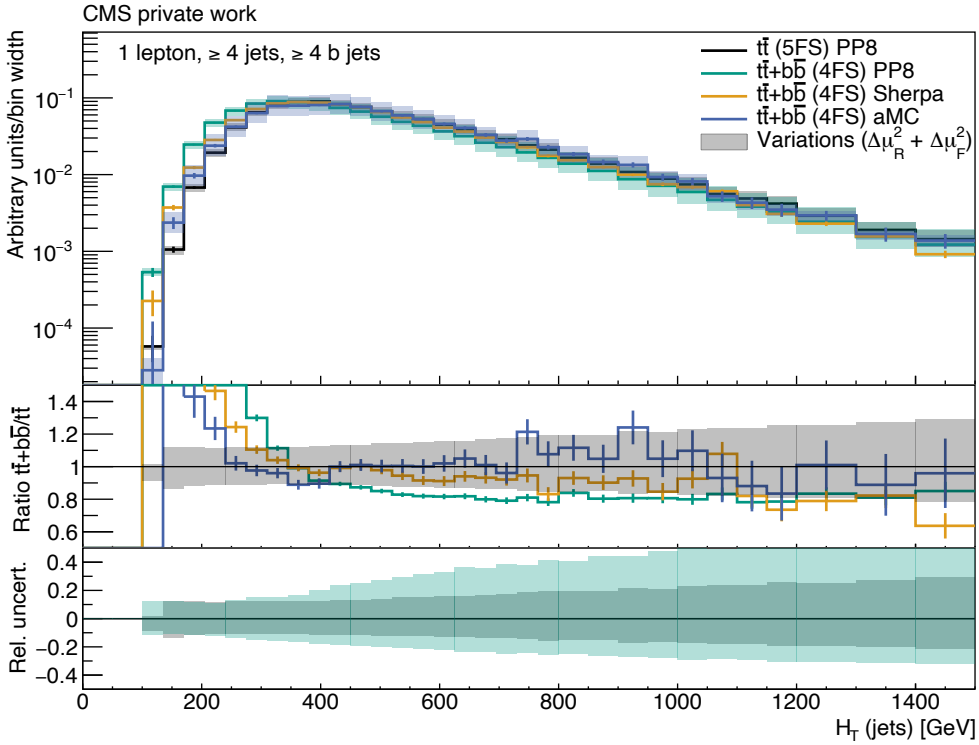


Figure 6.4: H_{T} (jets) of the $\mathrm{t\bar{t}}$ POWHEG+PYTHIA8 (black), $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ POWHEG+PYTHIA8 (green), $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ SHERPA (orange) and $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ MG5AMC(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_{R} and μ_{F} are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, $\geq$ 4 b jets".

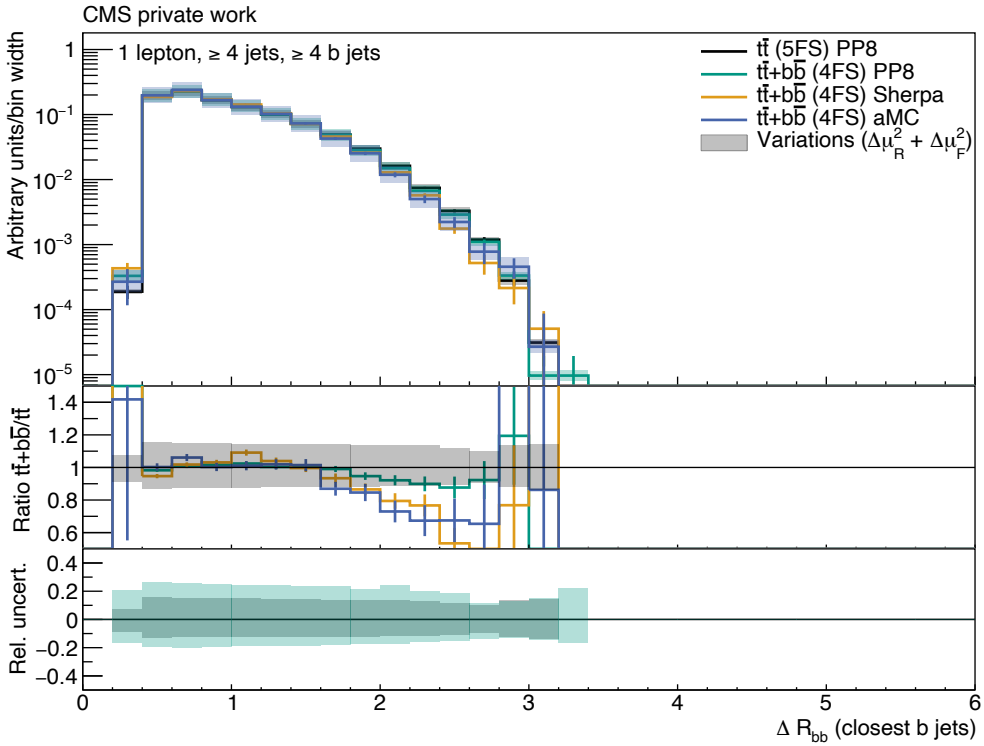


Figure 6.5: $\Delta R\,(\mathrm{bb})$ (closest) of the $\mathrm{t\bar{t}}$ Powheg+Pythia8 (black), $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ Powheg+Pythia8 (green), $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ Sherpa (orange) and $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ MG5-aMC(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_R and μ_F are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection ''1 lepton, $\geq$ 4 jets, $\geq$ 4 b jets''.

simulated with MG5aMC(NLO). The apparent effect is similarly prominent when the p_T value of the lepton is added to H_T (see Figure A.21). However, the effect is not visible if only b jets are considered in H_T (see Figure A.22).

An interesting characteristic of the events simulated with the three $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ simulation approaches can be seen in Figure 6.5. In this figure, the distance between the two closest b jets in the η, ϕ-plane is shown. It is identifiable that the $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ simulation approaches tend to predict smaller values in the ΔR (closest b jets) validation observable. The behavior can be attributed to the additional b jets, which do not originate from the top quark decays. When describing not only the $\mathrm{t\bar{t}}$ system but also the additional b quarks via the ME calculation, i.e. in the $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ simulation approaches, they show a more collinear characteristic compared to b jets from gluon splittings in the $\mathrm{t\bar{t}}$ simulation approaches.

As previously noted, the distributions can be affected by different $\mathrm{t\bar{t}}$ decay channel simulations, although a selection is applied to the single-lepton channel in the analysis routine. The effect is demonstrated for the ΔR (closest b jets) validation observable in Figure 6.6. As in the figures above, distributions of the events simulated with $\mathrm{t\bar{t}}$ Powheg+Pythia8 in a simulation including the dilepton channel and the single-lepton channel (black) and $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ Powheg+Pythia8 in a simulation including the dilepton channel and the single-lepton channel (green) can be seen. For exemplification, events

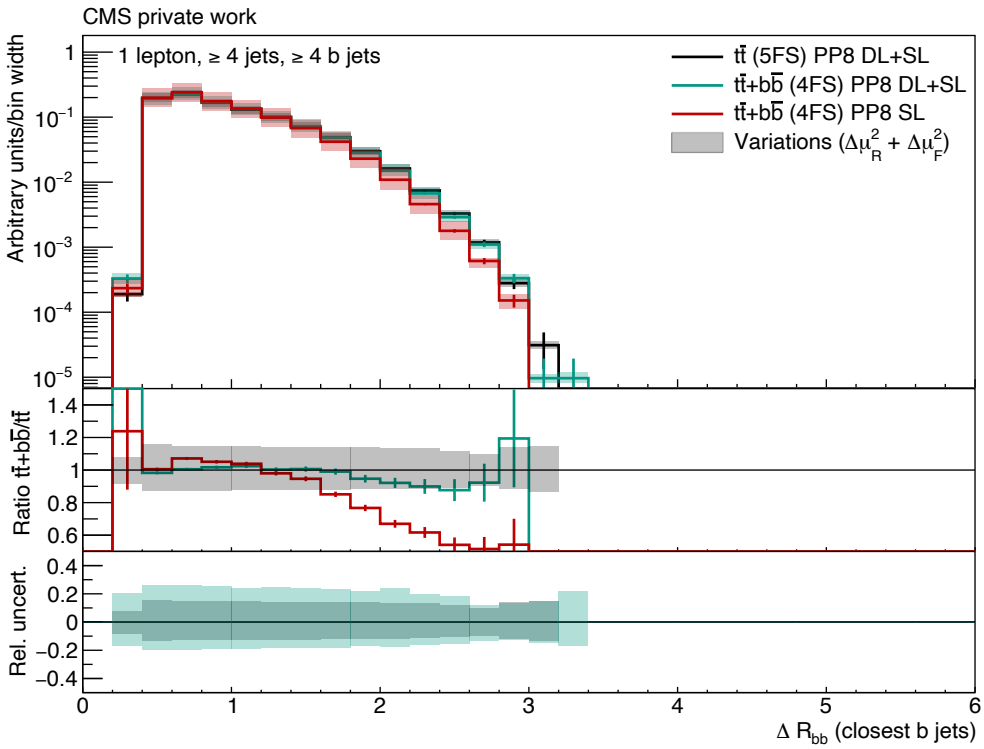


Figure 6.6: $\Delta R\,(\mathrm{bb})$ (closest) of the $\mathrm{t\bar{t}}$ POWHEG+PYTHIA8 (black), $\mathrm{t\bar{t}}+\mathrm{b\bar{b}}$ POWHEG+PYTHIA8 simulated events in a simulation of the dilepton channel and the single-lepton channel (green), $\mathrm{t\bar{t}}+\mathrm{b\bar{b}}$ POWHEG+PYTHIA8 simulated events in a simulation of the single-lepton channel only (red). The statistical uncertainties are shown as bars. The quadratically summed variations of μ_R and μ_F are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, ≥ 4 jets, ≥ 4 b jets".

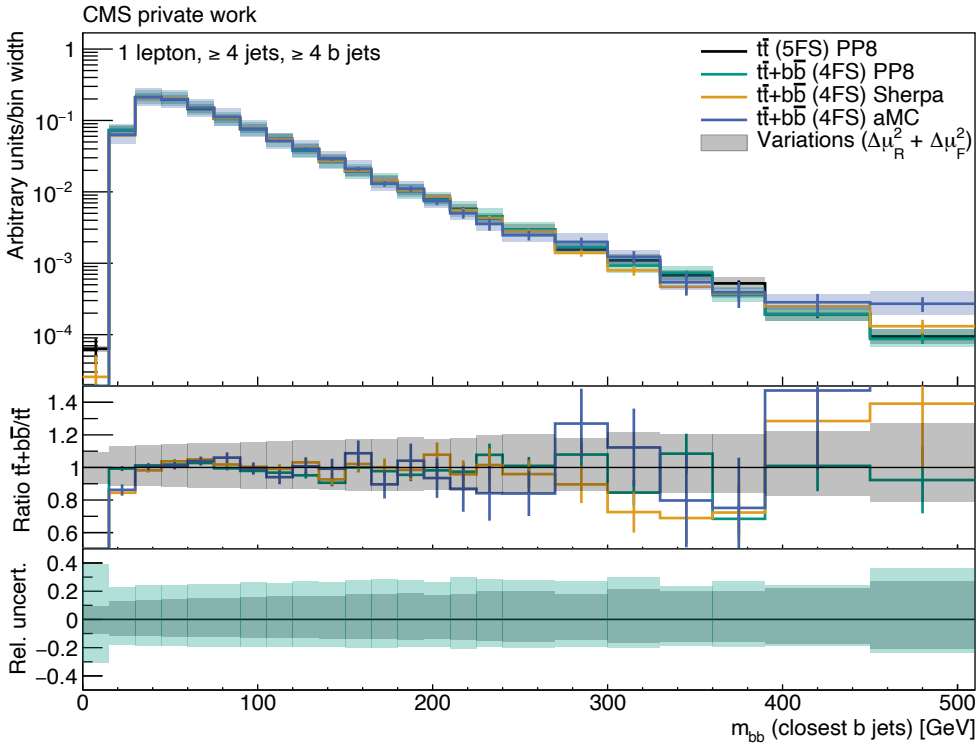


Figure 6.7: $m\,(\mathrm{bb})$ (closest) of the $\mathrm{t\bar{t}}$ Powheg+Pythia8 (black), $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ Powheg+Pythia8 (green), $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ Sherpa (orange) and $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ MG5-aMC(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_R and μ_F are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, $\geq$ 4 b jets".

simulated with a $\mathrm{t\bar{t}}$ Powheg+Pythia8 simulation approach in a simulation including the single-lepton channel only, which has not been shown before but is also investigated in this thesis, is displayed in red. The extent to which the inclusion of the dilepton channel simulation changes the distributions in this variable can be clearly seen. Therefore, despite the selection of events of the single-lepton channel in the routine of the analysis, it is highly important to carefully consider which simulated events are used in the analysis, i.e. adding the dilepton channel or not. For consistency, both $\mathrm{t\bar{t}}$ decay channels are included at every point in this comparison, with the exception of this illustrative example.

The mass of the two closest b jets $m\,(\mathrm{bb})$ (closest) is shown in Figure 6.7. The validation observable of the distributions of the simulated events with the three $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ simulation approaches varies around the distribution of the $\mathrm{t\bar{t}}$ simulation approach. In the range 60 GeV to about 100 GeV, a slight excess in all three distributions of the $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ simulation approaches can be observed. The $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ Powheg+Pythia8 simulated events show a decreasing trend compared to the events simulated with the $\mathrm{t\bar{t}}$ Powheg+Pythia8 simulation approach, which can be seen up to about 200 GeV. Above this value, the statistical fluctuations increase due to the small number of events at these high masses. No trend is discernible in the events simulated with Sherpa and MG5aMC(NLO).

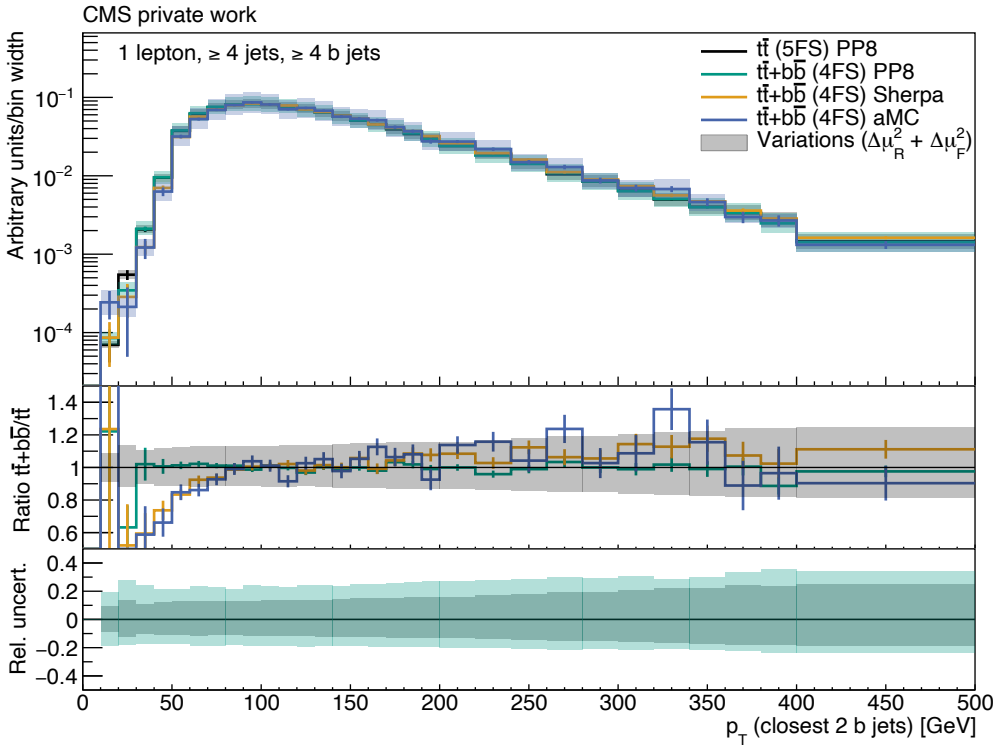


Figure 6.8: p_{T} (bb) (closest) of the $\mathrm{t\bar{t}}$ POWHEG+PYTHIA8 (black), $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ POWHEG+PYTHIA8 (green), $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ SHERPA (orange) and $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ MG5-AMC(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_{R} and μ_{F} are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection ''1 lepton, $\geq$ 4 jets, $\geq$ 4 b jets''.

After comparing the distance and mass of the closest two b jets in Figures 6.5 and 6.7, the p_{T} values of these jets are investigated in Figure 6.8. For small p_{T} values up to nearly 100 GeV for p_{T} (bb) (closest), the $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ events simulated with SHERPA and MG5AMC-(NLO) generators deviate significantly from the $\mathrm{t\bar{t}}$ POWHEG+PYTHIA8 simulated events. Also, both distributions show an increasing trend and thus a shift to higher p_{T} values, i.e. generally harder jets for the closest two b jets. In contrast, the two distributions of the events simulated with POWHEG+PYTHIA8 simulation approach agree well.

6.7 Comparison: ATLAS and CMS

The comparison of the $\mathrm{t\bar{t}}$ and $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ simulation approaches between ATLAS and CMS is a common effort of the LHC Higgs Working Group to foster a joint strategy across the two experiments. The figures discussed in this section contain the previously discussed CMS distributions, which are created for this thesis and the corresponding distributions created by the ATLAS Collaboration. The figures in this section are created by the ATLAS Collaboration after the CMS distributions have been provided for the purpose of a comparison. The distributions are based on the same analysis routine in the RIVET framework [92].
In the following section, $\mathrm{t\bar{t}}$ and $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ simulation approaches are compared between ATLAS

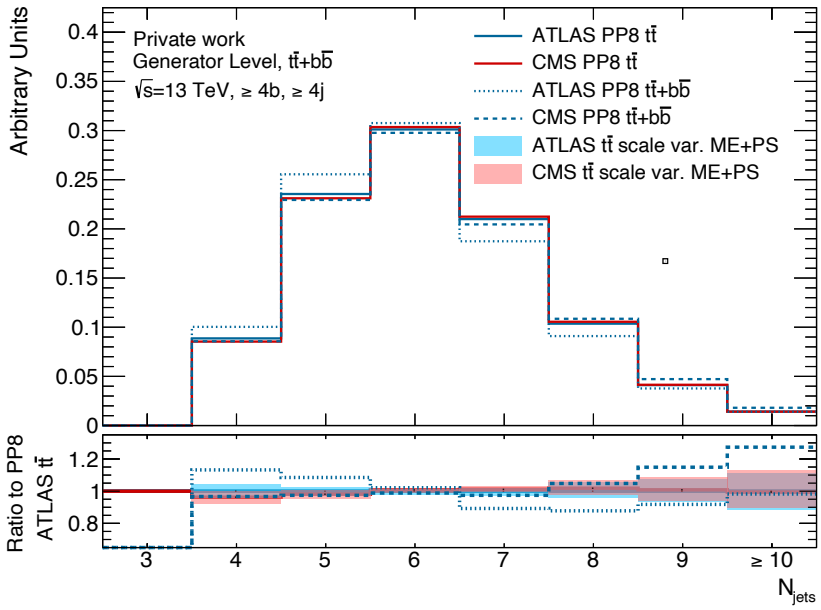
and CMS at generator level. The configurations used for the MC simulations are shown in Table 6.1. For the generation of the distributions, events are analyzed with the simulation of the dilepton channel as well as the single-lepton channel. This comparison enables to identify differences in the simulation approaches and their uncertainties between the two experiments and, if necessary, to adjust them afterwards.

The distributions of the events simulated with $t\bar{t}$ and $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8 simulation approaches from ATLAS and CMS are compared in Figure 6.9a for the jet multiplicity observable. The upper panel of the figure shows the distributions of the simulated events normalized to the integral value 1 as in the previous section. Unlike the distributions in the CMS simulation analysis, the vertical axis is not logarithmic in the following. The ME+PS uncertainties as defined in section 6.2.6 for the simulated events with $t\bar{t}$ POWHEG+PYTHIA8 simulation approaches are also visualized as uncertainty bands. In the lower panel of the figure the ratios of the distributions with respect to the $t\bar{t}$ POWHEG+PYTHIA8 simulated events from ATLAS can be seen. Clearly, the two distributions of simulated events with the $t\bar{t}$ POWHEG+PYTHIA8 simulation approach from ATLAS and CMS are in very good agreement. Also, the scale variations agree well. The simulated events of the two $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8 simulation approaches differ notably from each other. While the $t\bar{t}$+$b\bar{b}$ ATLAS simulation approach shows a shift to lower jet multiplicities, the $t\bar{t}$+$b\bar{b}$ CMS simulation approach shows an opposite pattern with an increased number of simulated events at higher jet multiplicities. This behavior was not observed in the comparison for CMS in Figure 6.1. There, the trend is rather similar to the $t\bar{t}$+$b\bar{b}$ ATLAS graph in Figure 6.9a. This effect may be attributed to taking into account different $t\bar{t}$ decay channels. Within the previous study, this effect was investigated for CMS and it was found that the trend behaves as shown in Figure 6.9a, if only events simulated in the single-lepton channel are examined. However, it is stated about the distributions that they include the dilepton channel as well as the single-lepton channel for all simulations.

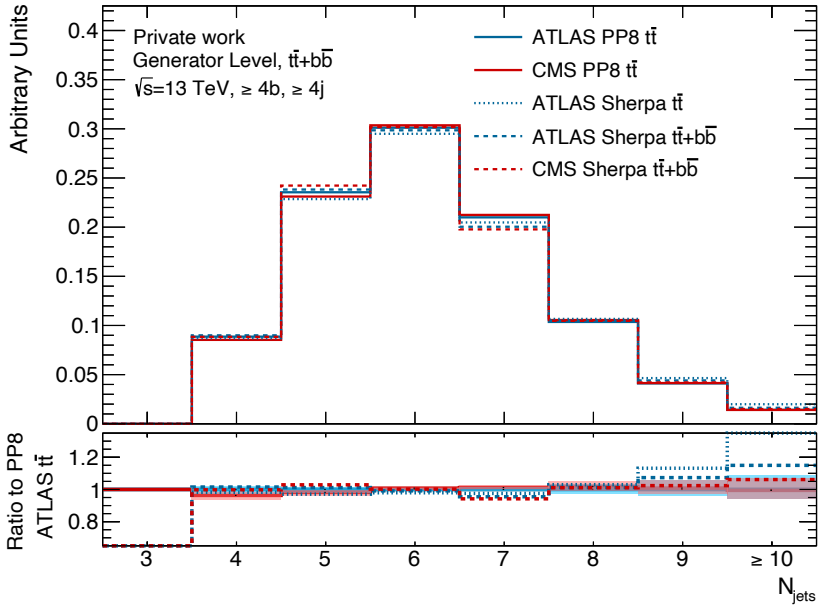
In Figure 6.9b, the distributions of the simulated events with the three SHERPA simulation approaches are compared with the previously analyzed $t\bar{t}$ POWHEG+PYTHIA8 simulated events. Among these three distributions of events simulated with the SHERPA generator are two simulation approaches from ATLAS, a $t\bar{t}$ and a $t\bar{t}$+$b\bar{b}$ simulation approach, and a $t\bar{t}$+$b\bar{b}$ simulation approach from CMS. It can be seen how well the $t\bar{t}$+$b\bar{b}$ SHERPA simulated events match with the $t\bar{t}$ POWHEG+PYTHIA8 simulated events. However, the distribution of the events with the $t\bar{t}$ simulation approach from ATLAS differs significantly from the other distributions for events with a large number of jets.

Figure 6.10a shows the validation observable H_T (jets) for the events simulated with POWHEG+PYTHIA8. Good agreement can be observed in the distributions of the $t\bar{t}$ simulation approaches from ATLAS and CMS in this observable. Also the sizes of the scale variations shown agree well. The $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8 simulated events from CMS differs slightly from the $t\bar{t}$ POWHEG+PYTHIA8 simulated events from ATLAS. The events simulated with the $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8 simulation approach from ATLAS deviates clearly. The deviation of these simulated events is particularly prominent around the region of the peak at about 400 GeV. Additionally, the simulated events reveal a tendency towards smaller values for H_T (jets) in the $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8 simulation approach from ATLAS.

In Figure 6.10b, the $t\bar{t}$+$b\bar{b}$ SHERPA simulated events are presented in comparison to the $t\bar{t}$ POWHEG+PYTHIA8 simulated events. Both distributions of the events simulated with the $t\bar{t}$+$b\bar{b}$ SHERPA simulation approach from ATLAS and CMS show an excess around the peak compared to the $t\bar{t}$ POWHEG+PYTHIA8 simulated events. Again, the distributions of the events simulated with the $t\bar{t}$+$b\bar{b}$ simulation approaches feature a shift to smaller p_T

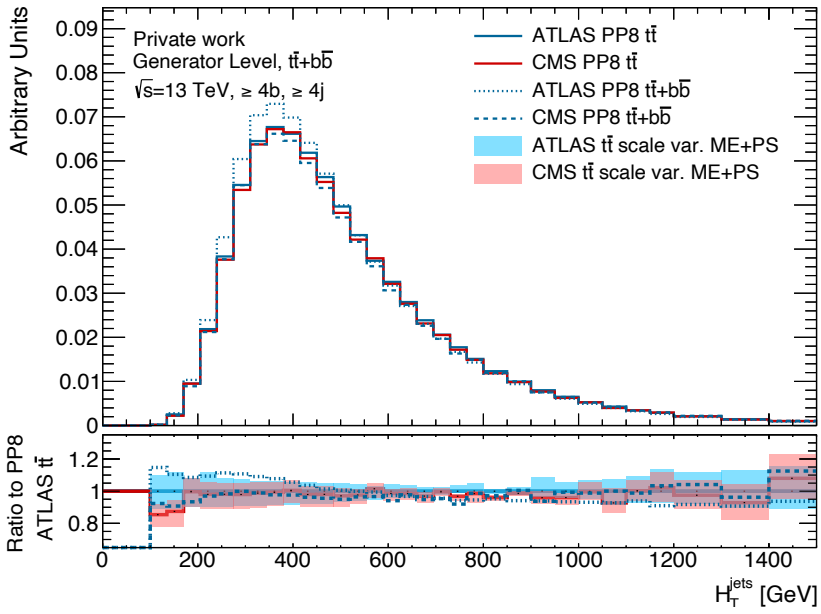


(a) $t\bar{t}$ Powheg+Pythia8 and $t\bar{t}$+$b\bar{b}$ Powheg+Pythia8 simulations

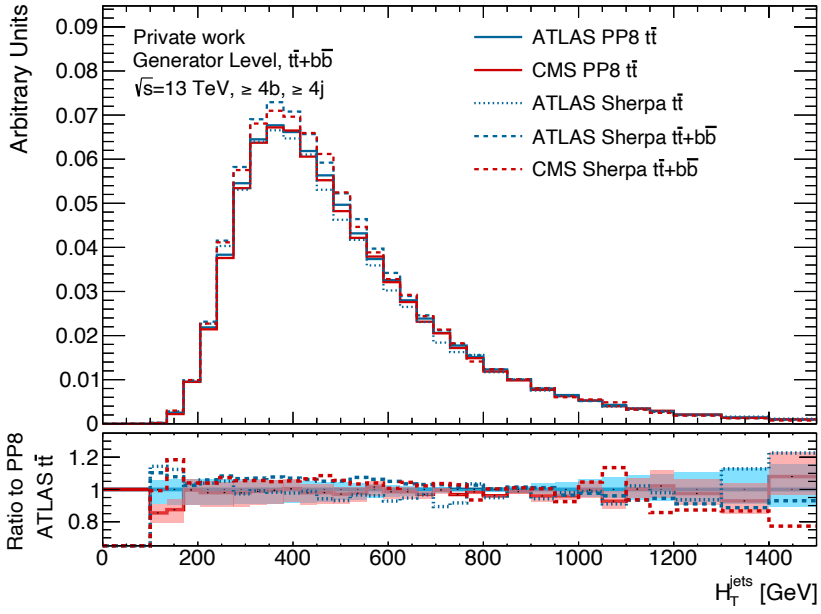


(b) $t\bar{t}$ Powheg+Pythia8 and $t\bar{t}$/$t\bar{t}$+$b\bar{b}$ Sherpa simulations

Figure 6.9: Jet multiplicity for the $t\bar{t}$ Powheg+Pythia8 and $t\bar{t}$+$b\bar{b}$ Powheg+Pythia8 simulated events (top) and for the $t\bar{t}$ Powheg+Pythia8 and $t\bar{t}$/$t\bar{t}$+$b\bar{b}$ Sherpa simulated events (bottom) for ATLAS and CMS. The distributions are normalized to an integral value of 1. The ratios are determined with respect to the $t\bar{t}$ PP8 simulated events from ATLAS. The plotted uncertainty bands show the scale variation of ME+PS. Plotting by A. Knue.



(a) $t\bar{t}$ Powheg+Pythia8 and $t\bar{t}$+$b\bar{b}$ Powheg+Pythia8 simulations



(b) $t\bar{t}$ Powheg+Pythia8 and $t\bar{t}$/$t\bar{t}$+$b\bar{b}$ Sherpa simulations

Figure 6.10: H_T (jets) for the $t\bar{t}$ Powheg+Pythia8 and $t\bar{t}$+$b\bar{b}$ Powheg+Pythia8 simulated events (top) and for the $t\bar{t}$ Powheg+Pythia8 and $t\bar{t}$/$t\bar{t}$+$b\bar{b}$ Sherpa simulated events (bottom) for ATLAS and CMS. The distributions are normalized to an integral value of 1. The ratios are determined with respect to the $t\bar{t}$ PP8 simulated events from ATLAS. The plotted uncertainty bands show the scale variation of ME+PS. Plotting by A. Knue.

values as previously observed. Being consistent with this trend, the $t\bar{t}$ SHERPA simulated events from ATLAS does not show this characteristic.

Figure 6.11a depicts the distribution of the distance ΔR between the two closest b jets $\Delta R\,(\mathrm{bb})$ (closest). Both $t\bar{t}$ POWHEG+PYTHIA8 simulated events from ATLAS and CMS agree very well in this observable. The ME+PS scale variations have identical dimensions in both distributions. Of particular interest is the strong and unambiguous difference to the distributions of the events simulated with the $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8 simulation approach. The closest two b jets in the event are therefore substantially more collinear in the simulation approaches in which the additional b jets are described in the ME than in the simulation approaches in which they are described through a PS. This effect was also identified in the analyses of the $t\bar{t}$+$b\bar{b}$ simulation approaches relative to the $t\bar{t}$ POWHEG+PYTHIA8 simulation from CMS in the previous section (see Figure 6.5). The validation observable $\Delta R\,(\mathrm{bb})$ (closest) shows the strongest difference of all observables examined between the $t\bar{t}$ and $t\bar{t}$+$b\bar{b}$ simulation approaches.

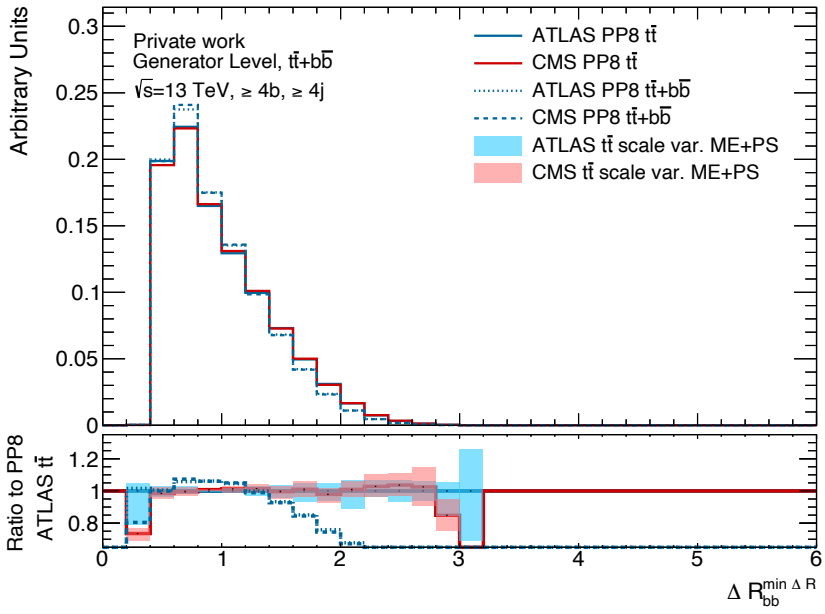
The distributions of the events simulated with the three SHERPA simulation approaches for $\Delta R\,(\mathrm{bb})$ (closest) are presented in Figure 6.11b. The $t\bar{t}$+$b\bar{b}$ SHERPA simulated events from both ATLAS and CMS, along with the $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8 simulated events, show the behavior of comparably close b jets. A striking feature is the fact that the distributions of the events simulated with the $t\bar{t}$ SHERPA simulation approach from ATLAS also demonstrates this strong trend.

In the following paragraphs the two b jets with the largest p_{T} values are analyzed. Figures 6.12a shows the mass of these two b jets. Comparing the two distributions of the events simulated with the $t\bar{t}$ POWHEG+PYTHIA8 simulation approach from ATLAS and CMS, a slight trend towards smaller masses can be discerned for the simulation approaches by CMS relative to the simulation approaches by ATLAS. Also the two distributions of the events simulated with the $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8 simulated events show this shift, whereas the deviation in distributions of the events simulated with the simulation approaches by CMS is larger.

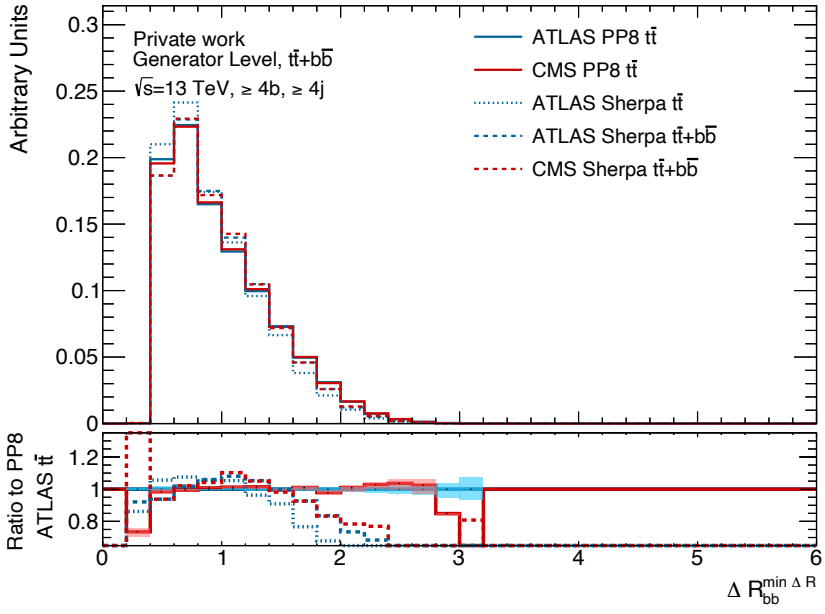
The observable $m\,(\mathrm{bb})$ (leading) for the events simulated with SHERPA generator is shown in Figure 6.12b. The two distributions of the events simulated with the $t\bar{t}$+$b\bar{b}$ SHERPA simulation approach, like the distribution of the $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8 simulated events, tend to yield smaller masses for the $m\,(\mathrm{bb})$ (leading) validation observable. The $t\bar{t}$ SHERPA simulated events from ATLAS shows a clear deviation from the $t\bar{t}$ POWHEG+PYTHIA8 simulated events, especially around the peak at about 180 GeV and smaller masses.

The p_{T} value of the two b jets with the highest p_{T} value is compared in Figure 6.13a for the events simulated with POWHEG+PYTHIA8. The two distributions of the events simulated with the $t\bar{t}$ POWHEG+PYTHIA8 simulation approaches indicate a good agreement. However, the $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8 simulated events demonstrate a decreasing trend compared to the $t\bar{t}$ POWHEG+PYTHIA8 simulated events from ATLAS. Thus, the jets tend to be marginally softer in this observable for the $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8 simulated events. Nevertheless, the deviations are considerably smaller than in the other validation observables.

Figure 6.13b shows the comparison of the three distributions of the events simulated with the SHERPA generator with respect to the $t\bar{t}$ POWHEG+PYTHIA8 simulated events from ATLAS. Also in these simulated events, the trend towards lower p_{T} values for the p_{T} (leading) validation observable can be seen.

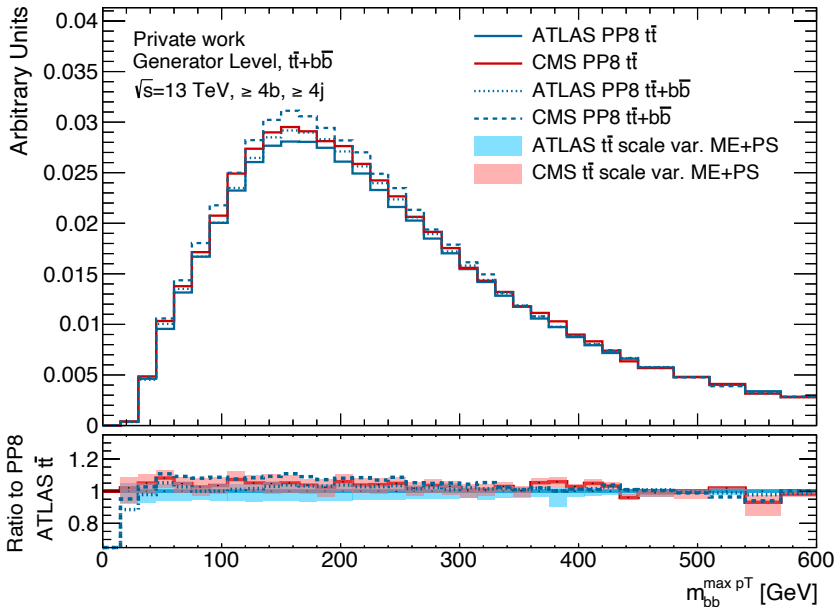


(a) $t\bar{t}$ Powheg+Pythia8 and $t\bar{t}$+$b\bar{b}$ Powheg+Pythia8 simulations

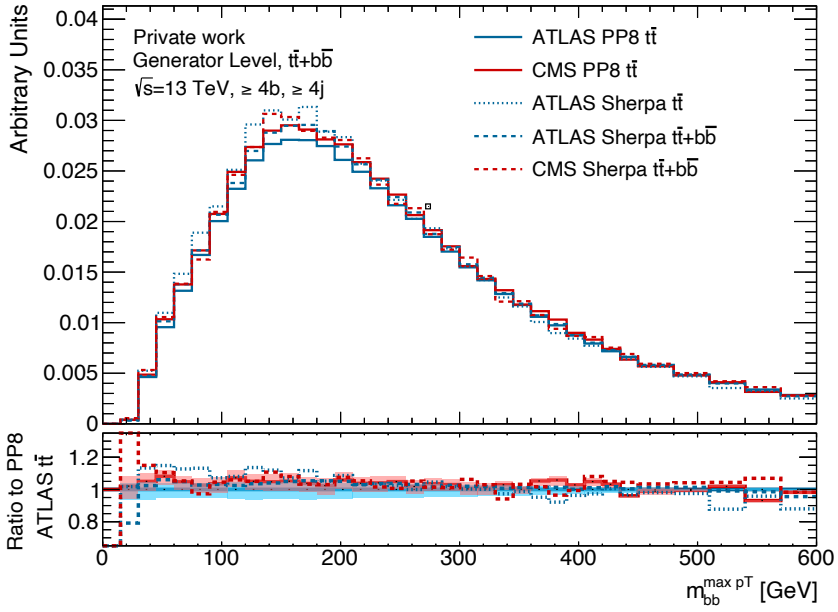


(b) $t\bar{t}$ Powheg+Pythia8 and $t\bar{t}$/$t\bar{t}$+$b\bar{b}$ Sherpa simulations

Figure 6.11: $\Delta R\,(\mathrm{bb})$ (closest) for the $t\bar{t}$ Powheg+Pythia8 and $t\bar{t}$+$b\bar{b}$ Powheg+Pythia8 simulated events (top) and for the $t\bar{t}$ Powheg+Pythia8 and $t\bar{t}$/$t\bar{t}$+$b\bar{b}$ Sherpa simulated events (bottom) for ATLAS and CMS. The distributions are normalized to an integral value of 1. The ratios are determined with respect to the $t\bar{t}$ PP8 simulated events from ATLAS. The plotted uncertainty bands show the scale variation of ME+PS. Plotting by A. Knue.

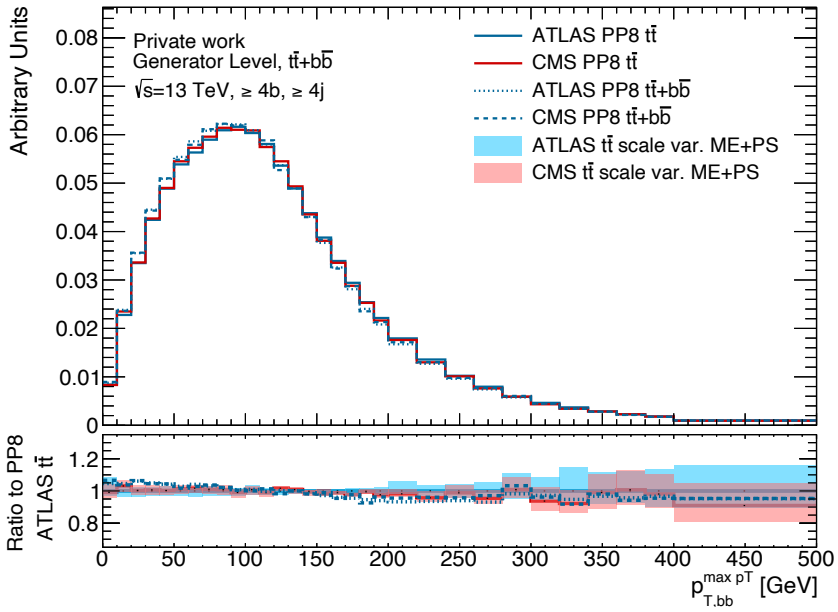


(a) $t\bar{t}$ Powheg+Pythia8 and $t\bar{t}$+$b\bar{b}$ Powheg+Pythia8 simulations

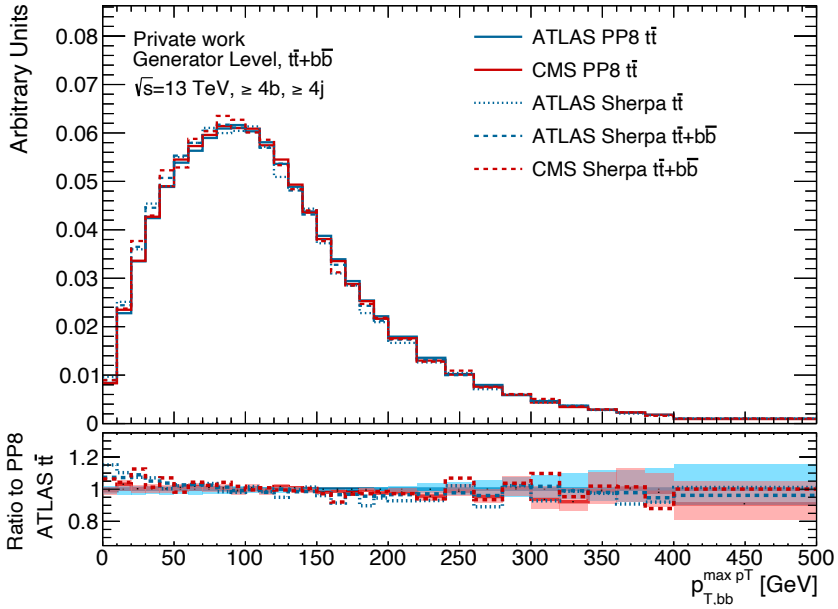


(b) $t\bar{t}$ Powheg+Pythia8 and $t\bar{t}$/$t\bar{t}$+$b\bar{b}$ Sherpa simulations

Figure 6.12: m (bb) (leading) for the $t\bar{t}$ Powheg+Pythia8 and $t\bar{t}$+$b\bar{b}$ Powheg+Pythia8 simulated events (top) and for the $t\bar{t}$ Powheg+Pythia8 and $t\bar{t}$/$t\bar{t}$+$b\bar{b}$ Sherpa simulated events (bottom) for ATLAS and CMS. The distributions are normalized to an integral value of 1. The ratios are determined with respect to the $t\bar{t}$ PP8 simulated events from ATLAS. The plotted uncertainty bands show the scale variation of ME+PS. Plotting by A. Knue.



(a) $t\bar{t}$ Powheg+Pythia8 and $t\bar{t}$+$b\bar{b}$ Powheg+Pythia8 simulations



(b) $t\bar{t}$ Powheg+Pythia8 and $t\bar{t}$/$t\bar{t}$+$b\bar{b}$ Sherpa simulations

Figure 6.13: p_T (bb) (leading) for the $t\bar{t}$ Powheg+Pythia8 and $t\bar{t}$+$b\bar{b}$ Powheg+Pythia8 simulated events (top) and for the $t\bar{t}$ Powheg+Pythia8 and $t\bar{t}$/$t\bar{t}$+$b\bar{b}$ Sherpa simulated events (bottom) for ATLAS and CMS. The distributions are normalized to an integral value of 1. The ratios are determined with respect to the $t\bar{t}$ PP8 simulated events from ATLAS. The plotted uncertainty bands show the scale variation of ME+PS. Plotting by A. Knue.

6.8 Summary

In this chapter, a comparison on four simulation approaches each from ATLAS and CMS was performed on events with $t\bar{t}$+$b\bar{b}$ processes in the single-lepton channel. The configurations of the simulation approaches are stated in Table 6.1 and the scale choices are presented in Table 6.2. The validation observables by which the simulated events were compared are defined in Table 6.3. The analysis routine is available in [92]. ATLAS and CMS partly use other versions and parameter settings of the MC generators. Also, ATLAS uses a renormalization scale and factorization scale that is twice as large compared to CMS in the $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8 simulation approaches. For the SHERPA simulation approaches, different $\mu_{R,F}$ scales are chosen in ATLAS compared to CMS.
The largest difference between $t\bar{t}$ and $t\bar{t}$+$b\bar{b}$ simulation approaches is the distance between the closest two b jets. The $t\bar{t}$+$b\bar{b}$ simulated events feature considerably smaller values in the $\Delta R\,(\mathrm{bb})$ (closest) observable than the $t\bar{t}$ simulated events. Moreover, the jet multiplicity and H_{T} distributions differ considerably. Other observables such as p_{T} or $m\,(\mathrm{bb})$ of the two leading or two closest b jets deviate rather moderately.
The differences in the examined observables between the distributions of the events simulated with the $t\bar{t}$ POWHEG+PYTHIA8 simulation approach from ATLAS and CMS are found to be small. Also, the $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8 simulated events from ATLAS and CMS show similar trends among themselves and also appear comparable with the $t\bar{t}$+$b\bar{b}$ SHERPA simulation approach in this regard. However, the distributions of the events simulated with the $t\bar{t}$+$b\bar{b}$ simulation approaches also reveal the aforementioned different characteristics in comparison with the $t\bar{t}$ simulation approaches. The distributions of the events simulated with the $t\bar{t}$ SHERPA simulation approach by ATLAS differs unexpectedly strongly from the $t\bar{t}$ POWHEG+PYTHIA8 simulated events and partly features characteristics of a $t\bar{t}$+$b\bar{b}$ simulation approach.

7 Reconstruction level study

The goal of the study presented in this chapter is the identification of the additional b jets (as defined in Chapter 5) in $t\bar{t}+b\bar{b}$ events at reconstruction level. Initially, the motivation of this study is discussed in section 7.1 and two main analysis methods are briefly introduced. The section also creates a common basis that goes beyond the previous definitions and eventually allows both analysis methods to be compared among each other. In sections 7.2 and 7.3 the two analysis methods are discussed in detail.

7.1 Overview

As pointed out in Chapter 5, the $t\bar{t}+b\bar{b}$ process is an important background in $t\bar{t}H(b\bar{b})$ measurements. The example Feynman diagrams of both processes in Figure 5.1 illustrate that the $t\bar{t}$ system does not differ in either case. The only difference is the origin of the two additional b jets, which do not stem from the top quark decays. Hence, it is key to understand the additional b jets as good as possible in order to gain a solid understanding of the $t\bar{t}+b\bar{b}$ process. Looking only at the objects visible in the detector, at LO one expects six jets and a lepton in the single-lepton channel. The neutrino from the leptonic W boson decay is not visible in the detector as explained in Chapter 4. At LO four of the six jets should originate from b quarks and are thus expected to be b tagged. However, due to b tagging inefficiencies, mistagging and acceptance effects, deviations may occur.
It is unknown which b jets in the final state originate from top quarks and which are the additional b jets in an event at reconstruction level. The study in this chapter therefore aims to answer this question: Which method allows for an assignment of the b jets to their origin and how accurate is the assignment? Since in the $t\bar{t}+b\bar{b}$ process the assignment of the additional b jets is crucial, the following analyses will focus on the assignment of these b jets.

To assign the additional b jets, two main analysis methods are studied. The first analysis method (section 7.2) aims at keeping the procedure as simple as possible and examines the most characteristic observables for the additional b jets. In contrast, the second analysis method (section 7.3) follows a more sophisticated approach. Using deep neural networks (DNNs), the additional b jets are reconstructed in a complex analysis process.

In order to technically realize the assignment of the additional b jets the true information of the jet origin is needed, which is not available at reconstruction level (see section 5.2.2). To be able to use the true information of the jets, which is only available at generator

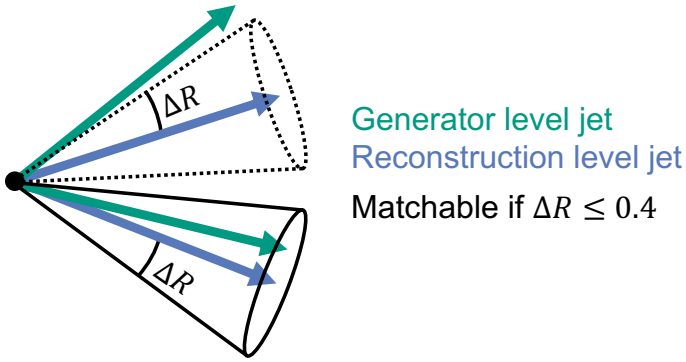


Figure 7.1: Visualization of the matching algorithm between a jet at generator level and a jet at reconstruction level. In the upper case the distance between the two jets is above the ΔR threshold. In the lower case the distance between the two jets is smaller than the given threshold, therefore these two jets are called matchable.

level, a **matching algorithm** between the jets at generator and reconstruction level is implemented. In this matching algorithm for each pair of reconstruction level and generator level jet, the distance ΔR between the two is calculated. In the next step, the two jets from both simulation levels whose distance is the smallest are assigned to each other. If the calculated ΔR value is equal to or below the threshold of 0.4, the two jets are called "matched" or "matchable". In other words, at both generator level and at reconstruction level a jet was found with a distance so small they can be assumed to be the same jet. As a result, this technique allows the true information from the generator level jet to be transferred to the reconstruction level jet whose origin is now known. The ΔR threshold for the matching is set to the radius parameter value of the anti-k_{T} algorithm, which determines the jet size (see section 4.3). The matching criterion between a jet at generator level and a jet at reconstruction level is illustrated in Figure 7.1. No conclusions can be drawn for the jets in an event where the threshold is exceeded. Summarized, the matching algorithm is an integral method to determine a jet's origin in an event at reconstruction level. A matched jet can thereby be labeled with its origin at reconstruction level.

To make the two jet identification methods comparable, a predefined metric is necessary. For this purpose the considered phase space is set to the $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ signal region ($\geq$ 6 jets, $\geq$ 4 b tagged jets). Furthermore, only those events are considered which fulfill the selection criteria for $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ events at generator level as defined in section 5.3. It is also required that the additional b jets are matchable. This results in the metric

$$a_{\text{evaluated method}} = \frac{N_{\text{events with correctly assigned additional b jets}}}{N_{\text{total events after selection}}} \tag{7.1}$$

which describes the assignment accuracy for an evaluated method, e.g. the accuracy of an observable to identify the additional b jets.

In the following analyses, specific terms are introduced for the jets and objects in a $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ event. This allows to differentiate the objects terminologically and to indicate their origin uniquely in the name at the same time. The names are allocated to the objects in Figure 7.2. The designations are characterized by the decay of the top quark. The top quark with the subsequently leptonically decaying W boson is called LepTop and the associated b jet is referred to as LepTopB. Accordingly, the top quark initiating the hadronic decay channel

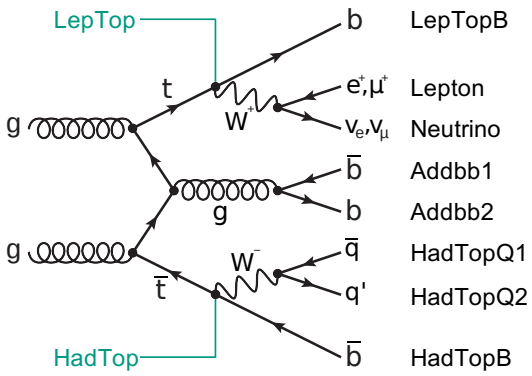


Figure 7.2: Example $t\bar{t}$+$b\bar{b}$ Feynman diagram with names introduced for the objects.

is called HadTop. The associated b jet is designated as HadTopB. The light flavor jets are named HadTopQ1 and HadTopQ2, but no further distinction is made between the two jets. The additional b jets are designated as Addbb1 and Addbb2, which are not further distinguished.

The study is applied to samples of the $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8 simulation and $t\bar{t}$ POWHEG+PYTHIA8 simulation, which were examined in detail in Chapter 6. In the following, the designation POWHEG+PYTHIA8 is omitted and the sample names are reduced to $t\bar{t}$+$b\bar{b}$ and $t\bar{t}$.

7.2 Distinctive observable method

The goal of this analysis method is to find the best observable identifying the additional b jets. Hence, the observable requires a feature that is different for additional b jets compared to b jets originating from the $t\bar{t}$ system. The results presented in this chapter are reduced to the best two distinctive observables.
From theoretical considerations, it is expected that the additional b jets tend to be closer to each other compared to all other distances ΔR between two b jets in a $t\bar{t}$+$b\bar{b}$ event. This is due to the fact that the additional b jets predominantly result from a collinear gluon splitting while the other b jets have no such common origin. In order to calculate the observable ΔR_{min} of all b jets and to test if the jets associated with that distance are the additional b jets, two separate steps are performed: First, all distances ΔR between all b jets in the event are calculated and the smallest value is selected. Second, the matching algorithm is used to determine whether this value is associated to the two additional b jets. Using the metric a, a value of $a_{\Delta R_{\text{min}},t\bar{t}+b\bar{b}} = 0.41$ is obtained for the $t\bar{t}$+$b\bar{b}$ sample and a value of $a_{\Delta R_{\text{min}},t\bar{t}} = 0.39$ for the $t\bar{t}$ sample.

A second observable is the minimum of the invariant mass of the $b\bar{b}$ system m_{min}. Analogous to the modus operandi of ΔR_{min}, the lowest value of m is determined and verified if it stems from the additional b jets. With the metric a, m_{min} results in $a_{m_{\text{min}},t\bar{t}+b\bar{b}} = 0.37$ for the $t\bar{t}$+$b\bar{b}$ sample and $a_{m_{\text{min}},t\bar{t}} = 0.35$ for the $t\bar{t}$ sample.
The results of the distinctive observable method are shown in Table 7.1. Besides ΔR_{min} and m_{min}, other kinematic observables such as η_{min} or p_{T} combination pairs (e.g. p_{T} of the leading and sub-leading b jet) were studied, but none of these observables showed an accuracy above 0.3.

Summarized, the method of determining ΔR_{min} leads to a correct assignment of the additional b jets in 41 % of the considered events for the $t\bar{t}$+$b\bar{b}$ sample and 39 % for the $t\bar{t}$

Table 7.1: Resulting accuracies according to the metric a (equation 7.1) of the best two distinctive observables for the $t\bar{t}$+$b\bar{b}$ and the $t\bar{t}$ sample.

	$t\bar{t}$+$b\bar{b}$ sample	$t\bar{t}$ sample
ΔR_{min}	0.41	0.39
m_{min}	0.37	0.35

sample and shows therefore the best results of all examined observables. Hence, the study in section 7.3 is always benchmarked against the results of the ΔR_{min} method.

7.3 Deep neural network based method

The aim of this study is to examine the feasibility of a more sophisticated reconstruction method in order to achieve a better assignment accuracy of the additional b jets. For this purpose, deep neural networks (DNNs) as introduced in the following are used. Section 7.3.1 presents three different training strategies and the methods for the assignment of the additional b jets. Section 7.3.2 builds on this and discusses the technical details and techniques of the DNNs that are used. The training strategies and their results are examined in detail in sections 7.3.3, 7.3.4 and 7.3.5.

7.3.1 Analysis strategy

Three different training strategies are applied for the assignment of the additional b jets. In Chapter 6 it was shown that the additional b jets differ in certain variables (e.g. p_{T} and ΔR) depending on the modeling. To be independent from this modeling, the first two strategies take an *indirect* look at the additional b jets. This means the DNN first reconstructs the $t\bar{t}$ system or parts of it. Subsequently, the additional b jets are identified with the help of an assignment method, i.e. a fixed decision rule. In contrast, the third strategy takes a *direct* look at the additional b jets.
The first strategy is the reconstruction of the entire $t\bar{t}$ system. Accordingly, this includes the two b jets from the top quark decay and the two light flavor jets from the hadronic W boson decay. At LO, the two remaining jets can be associated to the additional b jets in case of a correct DNN reconstruction.
The second strategy does not reconstruct the whole $t\bar{t}$ system and focuses solely on the reconstruction of the b jets from the top quark decays. If the correct b jets are identified by the DNN, all other b jets in the event have to be the additional b jets.
In contrast to the first two strategies, the third strategy focuses directly on the DNN reconstruction of the additional b jets. The strategy thus accepts the disadvantage to be dependent on modeling of the additional b jets.

The DNN reconstruction determines which jets in the event are assigned to which partons of a given strategy. As a useful technique, jet p_{T} indices are introduced to structure the jets in an event and facilitate assignments. The introduction of jet indices becomes particularly helpful towards the end of the analysis for an in-depth study, hence it is suitable to introduce and apply them already at this stage.
In each event, all jets are ordered by their p_{T} value. Depending on the strategy followed, the DNN reconstruction designates which jet indices correspond to the jets under consideration. For instance, the DNN determines the jet indices of the four jets resulting from the $t\bar{t}$ system. Based on the DNN's decision, three assignment methods are applied for the first two strategies, which assign the additional b jets to free indices in each event. The first

Table 7.2: Exemplary event with 9 jets and 5 b tagged jets. The jets are ordered by their p_T value. The second column shows the b tag value of the corresponding jet. A jet is considered to be b tagged if the b tag value exceeds the medium working point of 0.277 (see section 4.4). The DNN assigns the p_T indices 1, 4, 5 and 6 to the jets from the $t\bar{t}$ system. The last three columns show different results from the assignment methods.

#	p_T value	b tag value	DNN reco.	p_T assign.	$p_{T,\text{b tag}}$ assign.	b tag assign.
0	488 GeV	0.93		Addbb1	Addbb1	Addbb1
1	179 GeV	0.99	HadTopB			
2	145 GeV	0.01		Addbb2		
3	94 GeV	0.61			Addbb2	
4	61 GeV	0.02	HadTopQ1			
5	53 GeV	0.99	LepTopB			
6	42 GeV	0.13	HadTopQ2			
7	38 GeV	0.72				Addbb2
8	31 GeV	0.04				

assignment method selects the two jets with highest p_T whose indices are still unassigned after the DNN reconstruction (called p_T assignment method). The second method proceeds in the same way as the first method, but additionally requires that the jets assigned to the additional b jets are also b tagged (called $p_{T,\text{b tag}}$ assignment method). Contrary to the first two methods, the third method does not decide by p_T value ordering. In this method the jets with the highest b tag values are selected (called b tag assignment method). An exemplary event is shown in Table 7.2. Each of the three assignment methods in this event leads to a different result.

The overall input type for the DNN training is identical for all three strategies. First, it is evaluated whether all jets are matchable (see section 7.1) which are to be reconstructed by the DNN according to the strategy under scrutiny. For the first strategy this implies that the four jets of the $t\bar{t}$ system are matchable. If so, the event can be used for the training because the true information is now available. Two hypotheses are generated from this event, a signal and a background hypothesis. In the signal hypothesis the jets assigned to the $t\bar{t}$ system are used and the jets are assigned to the correct p_T indices. From the same event also a background hypothesis is created, which has an almost random (and consequently wrong) assignment to the p_T indices. As a constraint for both signal and background jet combination hypotheses it is specified that the b jet candidates associated with the b jets from the top quark decays must be b tagged. This procedure is performed on every event of the sample.

The entire analysis process from the sample to the final assignment of the additional b jets is depicted in Figure 7.3. Initially, the two hypotheses are created from the sample for each applicable event. These hypotheses are passed to the DNN as input and are trained against one another. The details of the training are covered in section 7.3.2. Next, the trained model is exported and evaluated with the events of the sample. The result for a single event at this point corresponds to the fourth column of Table 7.2. The monitored result is a summary of all events which shows how often the examined jets were assigned correctly. Afterwards, the assignment methods for the additional b jets are applied and also evaluated over all events. The result is the assignment accuracy of the additional b jets according to metric a (eq. 7.1).

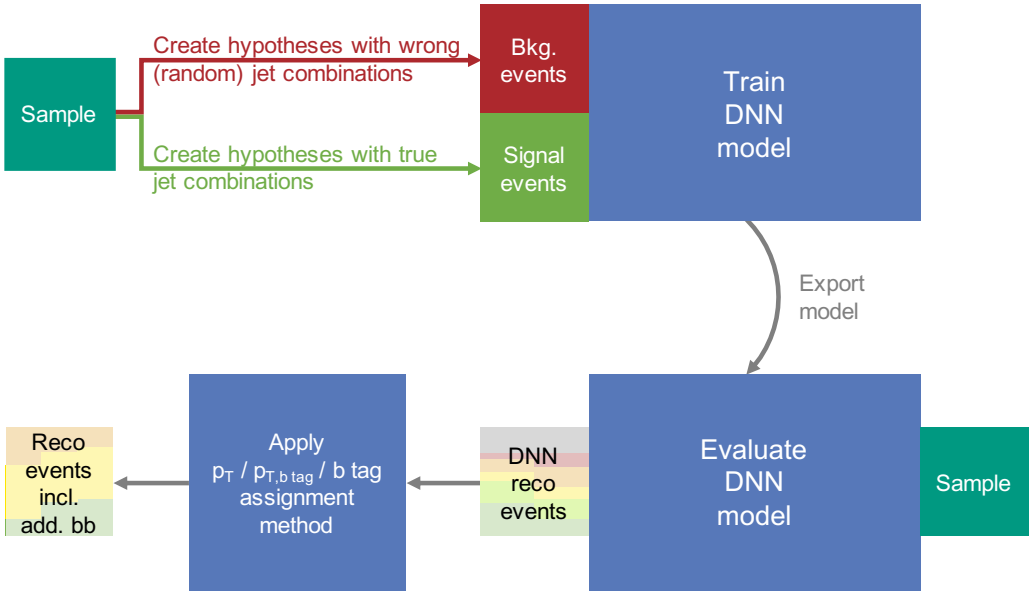


Figure 7.3: Illustration of the analysis process of the DNN based method. The upper part of the figure shows the creation of the hypotheses and the training of the DNN. The lower part of the graphic shows the evaluation of the sample, the application of the assignment methods and the final result, the assignment of the additional b jets.

7.3.2 Deep neural networks

Artificial neural networks are derived from the concept of biological neural networks [93]. To understand the functioning of an entire network it is helpful to begin with a single neuron. A neuron has n inputs x_i with $i = 1, \ldots, n$. These inputs x_i are multiplied with associated weights w_i to consider some inputs stronger or weaker than others. The sum of the products $x_i w_i$ is the argument of a non-linear activation function f, which results in an output value o

$$f\left(\sum_{i=1}^{n} x_i w_i\right) = o \quad . \tag{7.2}$$

Using the output o the single neuron can already make a decision. As long as the output is smaller or equal to the threshold b, it is assigned to a class A. If the value o is greater than the threshold b, the neuron classifies the input as class B. To change the result of the classification, the weights w_i and the threshold b can be modified, which is part of the training.
However, a single neuron can only execute rather simple classifications. If many of these neurons are arranged in a network-like structure, more complex objects can be classified. In DNNs this network-like structure is composed of several layers and a multitude of neurons in each layer as depicted in Figure 7.4. While there are many different types of DNNs, fully connected DNNs are used in this thesis. These fully connected DNNs posses the property that all neurons in a layer are connected to the previous layer and the subsequent layer. The first layer is the input layer. In this layer the initial information is fed into the network, analogous to the single neuron. An exemplary input feature x_i is given with the weight $w_{ji}^{(1)}$ to the j-th neuron in the first hidden layer. For instance, the input feature x_i could be the p_T value of the hypothesized b jet of the leptonic top quark decay. A network can contain several of these hidden layers, the figure shows only two layers and six neurons

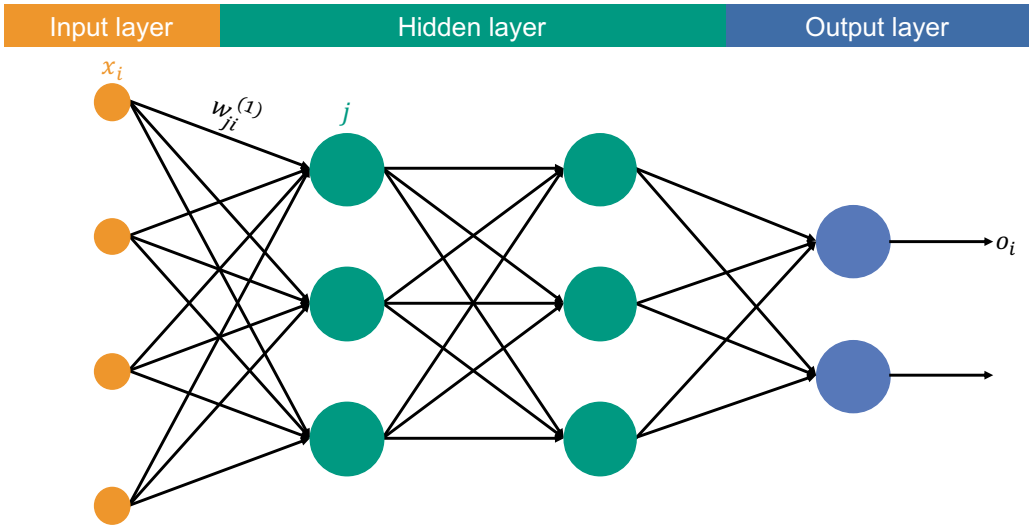


Figure 7.4: Visualization of a deep neural network. On the left side the inputs x_i are shown in the input layer (orange). Two hidden layers are depicted in the center (green). An exemplary weight from the i-th input to the j-th neuron is indicated with $w_{ji}^{(1)}$. On the right side the output layer with the output value o_i is illustrated (blue).

for better visualization. The generalization of equation 7.2 leads to

$$f\left(\sum_{i=1}^{n} x_i^{(l)} w_{ji}^{(l)}\right) = x_j^{(l+1)} \quad , \tag{7.3}$$

where the index l denotes the layer of the network. The output of a neuron of the l-th layer is one of the inputs for the $(l+1)$-th layer in the network. The leaky rectified linear unit (LeakyReLU)

$$f(x') = \begin{cases} x' & \text{if } x' > 0 \\ 0.3x' & \text{if } x' \leq 0 \end{cases} \quad , \tag{7.4}$$

is used as activation function f in this thesis [94]. The LeakyReLU function is used due to its simplicity stemming from the ReLU function, but unlike the latter, the LeakyReLU function does not have a slope of zero for negative values.

Since the network used in this thesis only takes a binary decision, there is only one neuron in the output layer. Based on the output value o, a binary decision is made whether the jet assignment hypothesis is true or false. The sigmoid function

$$\text{sig}(t) = \frac{1}{1+e^{-t}} \quad , \tag{7.5}$$

is applied as output activation function [95]. The choice is based on the good performance of the function for binary classifications [96].
To improve the accuracy of the DNN classification, the weights must be adjusted. This procedure is called training of a DNN. During the training, the hypotheses are propagated through the network and the output o is calculated. Subsequently, the output o is compared with the true output $\hat{o}$. If the hypothesis is true, i.e. an event with correct jet assignments, the true output value is $\hat{o}_{\text{true}} = 1$. In contrast, if the hypothesis is wrong, the true output

value is set to $\hat{o}_{\text{false}} = 0$. The quality of the classification is evaluated via the loss function. The binary cross-entropy function

$$L\left(o, \hat{o}\right) = -\left[\hat{o} \cdot \ln\left(o\right) + \left(1 - \hat{o}\right) \cdot \ln\left(1 - o\right)\right] \quad , \tag{7.6}$$

is used as loss function. The greater the deviation between the DNN output o and the true value $\hat{o}$, the worse the classification quality. The binary cross-entropy function is chosen because it is a well established standard for discriminative approaches [97]. The function mathematically describes the distance of the network's class prediction relative to the true class. Hence, the goal of the training is to minimize the loss function by adjusting the weights according to

$$w_{ji}^{(l)} \leftarrow w_{ji}^{(l)} - \rho \frac{\partial L}{\partial w_{ji}^{(l)}} \quad , \tag{7.7}$$

with the hyper parameter ρ. This tunable hyper parameter ρ is commonly called learning rate. According to equation 7.7, the extent to which the weights are changed is controlled by the learning rate. The procedure to update the weights is referred to as gradient descent [93]. In this thesis the method ADADELTA is applied for the gradient descent [98]. An advantage of this method is that the learning rate does not have to be set manually. Furthermore, ADADELTA is robust to large gradients, noise and the architecture choice [99].

Depending on the method, the weights can be updated after each hypothesis or after the entire set of hypotheses have been fully processed. A balanced approach is to define a batch size [100]. The batch size divides the set of available hypotheses into smaller packages with a fixed number of hypotheses. After a batch is completely processed, the weights are updated according to the optimizer ADADELTA. The complete processing of all batches and thus all hypotheses available for training is referred to as an epoch. The entire training of the DNN covers a large number of epochs, which are defined in advance. The number of training epochs can be chosen too large without consequences, since an early stopping criterion can be specified. The early stopping ends the training if there are no improvements in the loss function within a specific number of epochs [98].

When adjusting the large number of weights of the DNN, there is a risk called over-fitting. In case of over-fitting, the weights are adjusted to a degree that the generalization capabilities of the network are not preserved. Without generalization of the DNN the loss function is small and therefore the correct classification of training data is high, whereas the correct classification for unknown data is low. For this reason, the available data set is divided into two sets. The first subset is the training data set, on which the training is actually performed. The second subset is called validation data set. With the validation data set only the classification quality is monitored, i.e. the loss function is calculated, but no weights are adjusted. Over-fitting can be identified if the the loss function values of the training data set are continuously improving, but the loss function of the validation data set does not improve or even worsens [100].

Two methods are applied to avoid over-fitting. In order to avoid specific weights affecting the result too heavily, the L1 and L2 regularization is applied. With these two regularization methods, two additional terms are added to the loss function (eq. 7.6)

$$L_{\text{L1}} = \lambda_{\text{L1}} \cdot \sum_i \left|w_i\right| , \quad L_{\text{L2}} = \lambda_{\text{L2}} \cdot \sum_i w_i^2 \quad , \tag{7.8}$$

where $\lambda_{\text{L1}}, \lambda_{\text{L2}}$ are hyper parameters with $\lambda_{\text{L1}}, \lambda_{\text{L2}} < 1$ [98]. Hence, whenever the weights are too large, the loss function will be penalized.
Another method used to prevent over-fitting is the technique called DROPOUT [101]. With the DROPOUT method, a percentage of neurons including all associated connections in the

Table 7.3: Configuration of the DNNs

parameter	settings
number of hidden layers	4
number of neurons per hidden layer	100
activation function	LeakyReLU (eq. 7.4)
output activation	Sigmoid (eq. 7.5)
loss function	binary crossentropy (eq. 7.6)
optimizer	Adadelta (default settings) [98]
batch size	128
number of epochs	max. 2000
early stopping	no loss value improvement in 20 epochs
$\lambda_{L1}, \lambda_{L2}$	10^{-4} (eq. 7.8)
Dropout rate	0.2

DNN is randomly disabled from the network for each batch. This prevents a too strong impact of single neurons.
The configuration of the DNNs used in the following sections is summarized in Table 7.3. Each DNN parameter was varied independently in the trainings. The number of hidden layers was varied between two and ten. Also, the performance of the DNN was examined with $25, 50, 100, 200$ and 500 neurons per hidden layer. The Dropout rate was changed to $0.05, 0.1, 0.2$ and 0.4. Batch sizes of $128, 512$ and 2048 were examined in the tests. Moderate deviations from this configuration, such as doubling the neurons per hidden layer, adding an additional hidden layer, etc., showed minor deviations of a few percent. Large variations as for example a network configuration with 500 neurons per hidden layer and a total of ten hidden layers could no longer result in reasonable classifications. The configuration listed in Table 7.3 represents the identified optimal network design.

7.3.3 $t\bar{t}$ system reconstruction strategy

As briefly introduced in section 7.3.1, the first strategy is a DNN-based reconstruction of the $t\bar{t}$ system. At LO, if the $t\bar{t}$ system jets are correctly identified, only the additional b jets remain in an event. Effects beyond LO and how exactly the methods deal with them is demonstrated towards the end of this chapter. To elucidate the idea of the approach, only LO is mentioned at this point. However, the $p_{T,\text{b tag}}$ assignment method as well as the b tag assignment method can also assign the additional b jets in events with high jet multiplicities.
For a detailed analysis of the strategy's accuracy, three different aspects are evaluated. At first, the maximum possible efficiency which the assignment methods p_T, $p_{T,\text{b tag}}$ and b tag can achieve is calculated, assuming that the DNN always classifies correctly. As a second step, only the classification accuracy of the DNN is evaluated. The aim is to determine how well the DNN reconstructs the $t\bar{t}$ system on which the assignment methods will be based. Finally, the actual accuracies of the three assignment methods of the additional b jets are calculated.

The maximum possible efficiency of the three assignment methods is shown in Figure 7.5. For this purpose the $t\bar{t}$ system is determined first with the help of the true information through the matching algorithm. After the identification of the true $t\bar{t}$ system the three assignment methods are applied to determine the additional b jets. Again, the matching algorithm is used to verify whether these are the true additional b jets. In this calculation

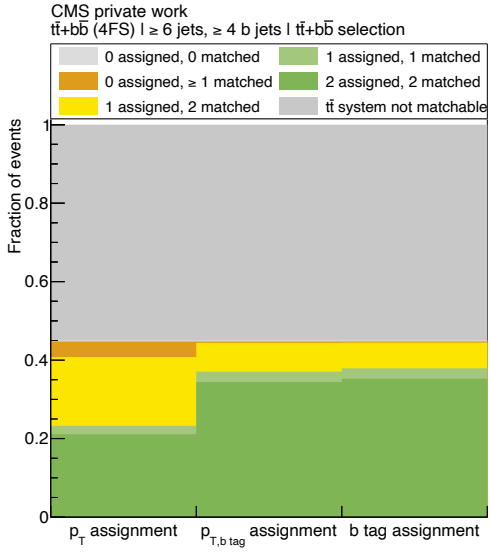


Figure 7.5: Maximum possible efficiency of the p_{T}, $p_{\mathrm{T,b\ tag}}$ and b tag assignment method for the $\mathrm{t\bar{t}}$ system reconstruction strategy assuming a perfect DNN reconstruction. It is shown how many additional b jets are assigned in which fraction of events depending on the number of jets assigned via the matching algorithm.

the $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ sample is evaluated. According to the metric a defined in equation 7.1, a selection is applied to ≥ 6 jets, ≥ 4 b tagged jets and only $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ events are selected (see Figure 5.3a). Figure 7.5 lists how often the methods assign the additional b jets correctly. A precise distinction is made as to how many of the additional b jets can actually be identified with the matching algorithm. For example, the class ''1 assigned, 1 matched'' denotes that an additional b jet was correctly identified, but at the same time only one additional b jet could be identified using the matching algorithm. The class ''1 assigned, 2 matched'' denotes that only one of the two additional b jets was correctly identified by the assignment method. No statement can be made for the relatively large gray area in the figure, since the full $\mathrm{t\bar{t}}$ system cannot be identified with the matching algorithm in these events. These events are referred to as ''$\mathrm{t\bar{t}}$ system not matchable''. The figure shows that the b tag method is the most accurate assignment method. The fraction of events in which the b tag assignment method correctly identifies the additional b jets, but without the events in which no statement can be made, is 0.84. The b tag assignment is closely followed by the $p_{\mathrm{T,b\ tag}}$ method, the p_{T} method is far less accurate.

For the $\mathrm{t\bar{t}}$ system strategy two DNNs are trained, which differ in the set of input variables. The input variables of the first DNN are shown in Table 7.4. Essentially, the quantities p_{T}, M, E, η, ϕ, b tag value and the p_{T} index of each of the four jets of the $\mathrm{t\bar{t}}$ system are fed into the DNN. In addition, hypothetical top quark-like objects are assembled from the corresponding jets and the lepton. In the hadronic case the four-vector sum of the HadTopB, HadTopQ1 and HadTopQ2 candidates is calculated and interpreted as HadTop. In the leptonic case only LepTopB and the Lepton are added, since the neutrino does not exist as an object. In principle, the MET could also be incorporated to add an energy associated with the neutrino (see section 4.5). This would allow for the reconstruction of a hypothetical LepTop via the top mass and the boundary condition of a quantity associated with a transverse W mass. However, this is not necessary here, since already the combination of a LepTopB candidate and the lepton is very distinct in the correct combination compared to a wrong combination, even if the LepTop mass is shifted in this

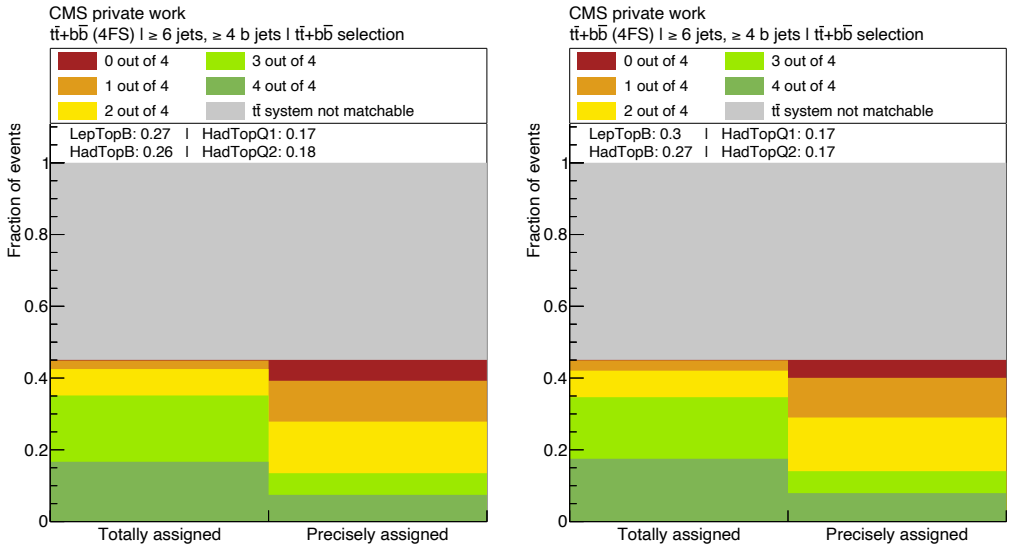
Table 7.4: Input variables of the DNN training on the $t\bar{t}$ system.

HadTopB	LepTopB	HadTopQ1	HadTopQ2	HadTop	LepTop	$t\bar{t}$
p_T	p_T	p_T	p_T	p_T	p_T	p_T
M	M	M	M	M	M	M
E	E	E	E	E	E	E
η	η	η	η	η	η	η
ϕ	ϕ	ϕ	ϕ	ϕ	ϕ	
b tag value	b tag value	b tag value	b tag value			
p_T index	p_T index	p_T index	p_T index			
						Δp_T
						$\Delta\eta$
						$\Delta\phi$
						open. angle

case. HadTop and LepTop are combined to form a $t\bar{t}$ system. From the combined objects again input variables are generated, which are also listed in Table 7.4. The distributions of these input variables for the signal and background hypotheses are provided in Appendix B. The second DNN gets all inputs of the first DNN and a large number of additional inputs. Supplementary, all possible jet and lepton combinations are formed, analogous to the assembled objects HadTop and LepTop. With these objects intentionally logical objects arise, for example from HadTopQ1 and HadTopQ2 a hadronic W boson-like object is created, but also presumably uncorrelated connections between objects like HadTopB and the Lepton are generated. It is then up to the DNN to determine which of these objects contribute to the classification and which do not.

The input data set for the DNNs are generated jet combination hypotheses from the $t\bar{t}$ and the $t\bar{t}+b\bar{b}$ sample as described in section 7.3.1. The input variables of both samples were compared and found not to show any significant deviations. To ensure no differences in the performance, separate trainings were executed on the respective samples. No notable differences were observed compared to a joint training. For this reason, and to increase the number of hypotheses for the trainings, a combined data set from the $t\bar{t}$ and the $t\bar{t}+b\bar{b}$ sample is used for training. A total of approximately 25500 events in the desired phase space are available, from which both signal and background hypotheses can be constructed. The DNNs are evaluated individually on the two samples. Since the results are very similar, only the results of the evaluation of the DNNs on the $t\bar{t}+b\bar{b}$ sample are discussed in this chapter.

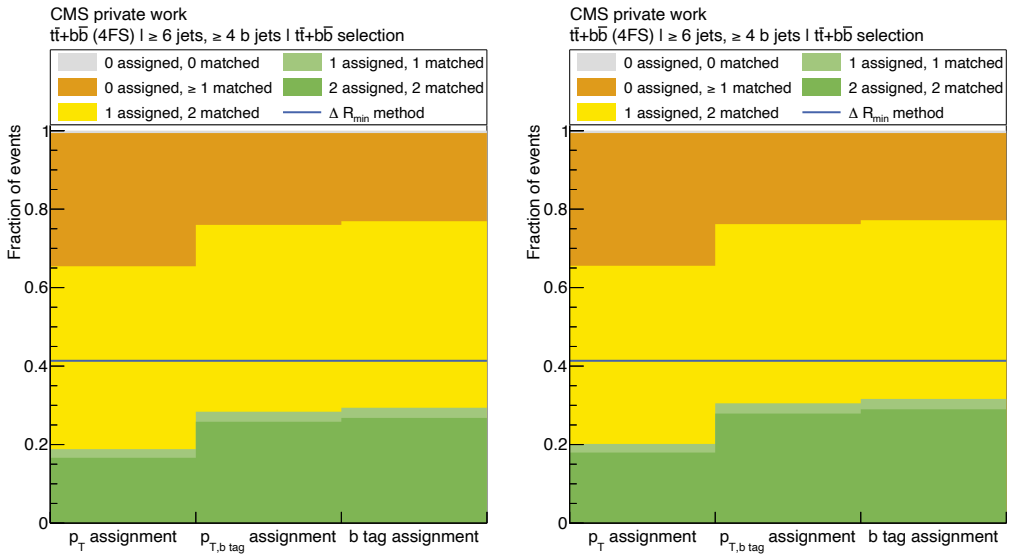
The results of the DNN reconstruction of the $t\bar{t}$ system are presented in Figure 7.6. The figure shows how many of the four jets of the $t\bar{t}$ system were correctly assigned. Two cases are distinguished in the accuracy of the assignment: ‘‘precisely assigned’’ and ‘‘totally assigned’’. The ‘‘precisely assigned’’ case verifies whether a specific jet has also been assigned as this jet, e.g. the HadTopB jet as a HadTopB jet. In contrast to this the ‘‘totally assigned’’ case is invariant to permutations. In this case it is irrelevant whether two or more jets of the $t\bar{t}$ system have been mixed up, e.g. the HadTopB jet with the LepTopB jet. Because even in case of a confusion the assignment methods for the additional b jets give an identical result as long as the $t\bar{t}$ system was recognized in total. The results on the DNN performance show that the two b jets of the $t\bar{t}$ system are detected more often compared to the light flavor jets. Adding the large number of input variables improves the overall DNN performance only slightly. The fraction of correct assignments of b jets



(a) DNN with fewer input variables

(b) DNN with extended input variables

Figure 7.6: DNN performance of the $t\bar{t}$ system reconstruction. It is shown how many jets are correctly assigned by the DNN in which fraction of events. Figure 7.6a shows the DNN performance based on training with the comparably small set of input variables from Table 7.4, Figure 7.6b shows the DNN performance based on the training with the considerably enlarged set of input variables. The label ‘‘precisely assigned’’ denotes the exact assignment of each jet-type, while ‘‘totally assigned’’ is invariant under a permutation of two or more jet-types.



(a) DNN with fewer input variables

(b) DNN with extended input variables

Figure 7.7: Assignment method performance of the $t\bar{t}$ system reconstruction. It is shown how many additional b jets are assigned in which fraction of events depending on the number of jets assigned via the matching algorithm. Figure 7.7a shows the assignment method performance based on the DNN which is trained with the comparably small set of input variables from Table 7.4, Figure 7.7b shows the assignment method performance based on the DNN which is trained with the considerably enlarged set of input variables.

is a bit improved, the frequency of light flavor jets remains almost identical. Hence, the DNN is deep enough to explore possible supplemental information from these additional combinations itself. The fraction of events for the four distinct jets of the $t\bar{t}$ system is also given in the figure.
Based on an input variable's sum of weights of the DNN's first layer a heuristic of the importance can be deduced. The input variables identified via this method which contribute most to correct classification are the invariant masses of the composite objects HadTopB, LepTopB and $t\bar{t}$. The fact that these input variables strongly enhance the classification is reasonable, because for a correct jet combination hypothesis a value around the top quark mass defined in the MC generator ($m_t = 172.5\,\text{GeV}$) can be expected for the hadronic top quark (see section 6.2.2). For the leptonic top quark a value related but not equal to the top quark mass is to be expected due to the missing neutrino. In contrast, a wrong hypothesis, where an additional b jet is assumed as for example HadTopB, should not feature this characteristic.

Based on this DNN reconstruction, the assignment methods for the additional b jets are applied. The results are presented in Figure 7.7. The visualization is analogous to the maximum performance in Figure 7.5, but in this case it is not needed to exclude those events where the $t\bar{t}$ system cannot be found via the matching algorithm. Based on the analysis of the maximum performance, the b tag method was already expected to perform most efficiently. The best result of the strategy is an assignment accuracy of the additional b jets of about $a_{t\bar{t}} = 0.29$ with the b tag method. Figure 7.7 also includes the best result of the distinctive observable method from section 7.2 as a benchmark. The strategy of reconstructing the $t\bar{t}$ system with subsequent selection according to the b tag values is

independent of the modeling of the additional b jets, but performs worse than the selection using the ΔR_{min} observable.
Applying the strategy showed that the b jets of the $t\bar{t}$ system are quite well identified compared to the light flavor jets. Furthermore, the b tag assignment method is the most efficient of the three assignment methods. Therefore, the strategy explored in the following section focuses exclusively on the b jets from the $t\bar{t}$ system.

7.3.4 b jets from the $t\bar{t}$ system reconstruction strategy

The general approach of the strategy in this section is consistent with the procedure of the $t\bar{t}$ system reconstruction strategy. Since the light flavor jets of the $t\bar{t}$ system are reconstructed rather weakly (see section 7.3.3), only the b jets from the $t\bar{t}$ system are considered in this strategy. Furthermore, the confusion of b jets from top decays with additional b jets is expected to be largest, especially due to the fact that the b tag values are the most important criterion for the differentiation between the b jets and all other jets in the event. However, the b tag values are weak metrics to distinguish b jets from the $t\bar{t}$ system and additional b jets. Therefore, this strategy focuses purely on the distinction of the b jets. Figure 7.8 shows the maximum efficiencies of the three assignment methods. The set of events where the entire $t\bar{t}$ system is matchable is a subset of events where just the b jets of the $t\bar{t}$ system are matchable. Therefore, more events are matchable in this strategy and the visualization is only comparable relative to the matchable event fraction of the $t\bar{t}$ system reconstruction strategy in the corresponding Figure 7.5. The p_{T} method performs particularly weakly in this strategy, since only jets up to the third index can be assigned. A more detailed breakdown is presented later in this section. In this strategy the fraction of events in which the b tag assignment method correctly identifies the additional b jets, but without the events in which no statement can be made, is 0.88. Hence, the theoretically achievable maximum accuracy is higher compared to the $t\bar{t}$ system reconstruction strategy (0.84).

To train the DNNs they are given the input variables p_{T}, M, E, η, ϕ, b tag value and the p_{T} index of HadTopB and LepTopB. As in the previous strategy, composite objects are formed from all possible combinations of HadTopB, LepTopB and the lepton. From these objects the quantities M, $\Delta\eta, \Delta\phi$ and ΔR are provided as input variables into the DNNs. In this section the trainings of two different networks are analyzed, which differ only in the selection of the events from which the jet assignment hypotheses are generated. As in section 7.3.3 the first network is trained in the signal region ≥ 6 jets, ≥ 4 b tagged jets. The second DNN is trained on an enlarged selection, which includes all events with ≥ 4 jets, ≥ 2 b tagged jets. In the enlarged region of ≥ 4 jets, ≥ 2 b tagged jets there might not be additional b jets, but since the focus of the training is exclusively on HadTopB and LepTopB it is independent of the additional b jets in any case. A possible advantage is the increased number of available training events. For the DNN, which is trained in the ≥ 6 jets, ≥ 4 b tagged jets region, a total of approximately 44000 events are available. In contrast, with the extended event selection, the DNN has over 7.7 million events available for the training. After the trainings, both DNNs are evaluated on the identical event selection, according to the metric a. The performance of the two DNNs is shown in Figure 7.9. The results of the two DNN trainings do not vary significantly. With approximately 55 % the LepTopB is far better reconstructable than the HadTopB (32 %). It can also be seen that the difference between ‘‘totally assigned’’ and ‘‘precisely assigned’’ in the ‘‘2 out of 2’’ class is relatively small. In other words, LepTopB and HadTopB are rarely confused. Therefore, certain attributes have to exist for the DNNs, which distinguish these two b jets from each other. In a similar manner to the previous strategy, the masses M of the composite objects LepTopB plus Lepton and HadTopB plus Lepton are the best separating variables.

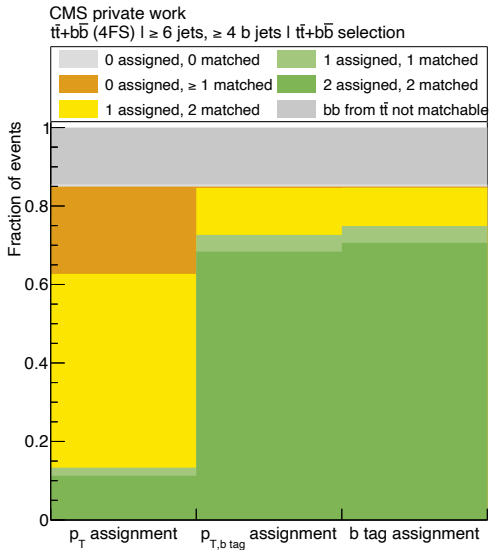
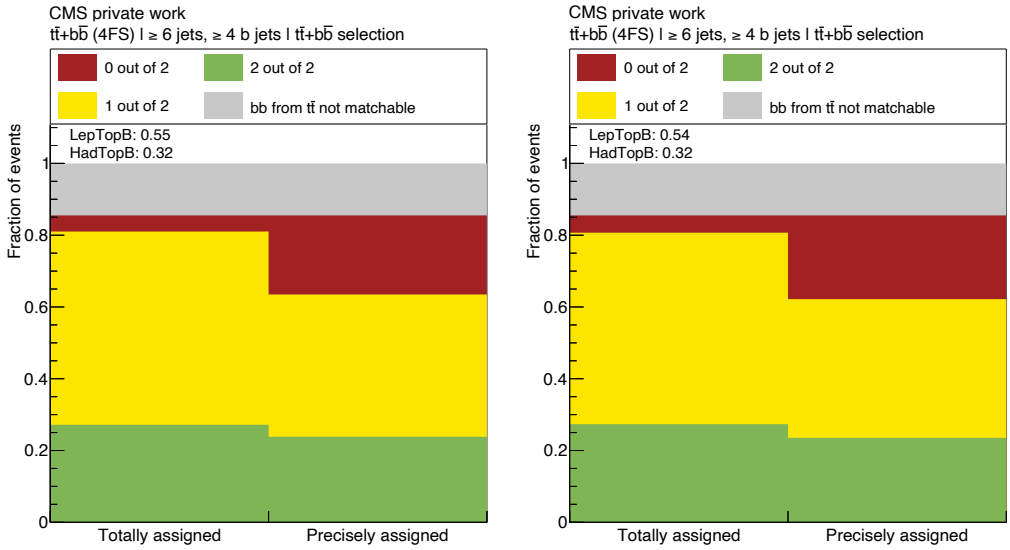


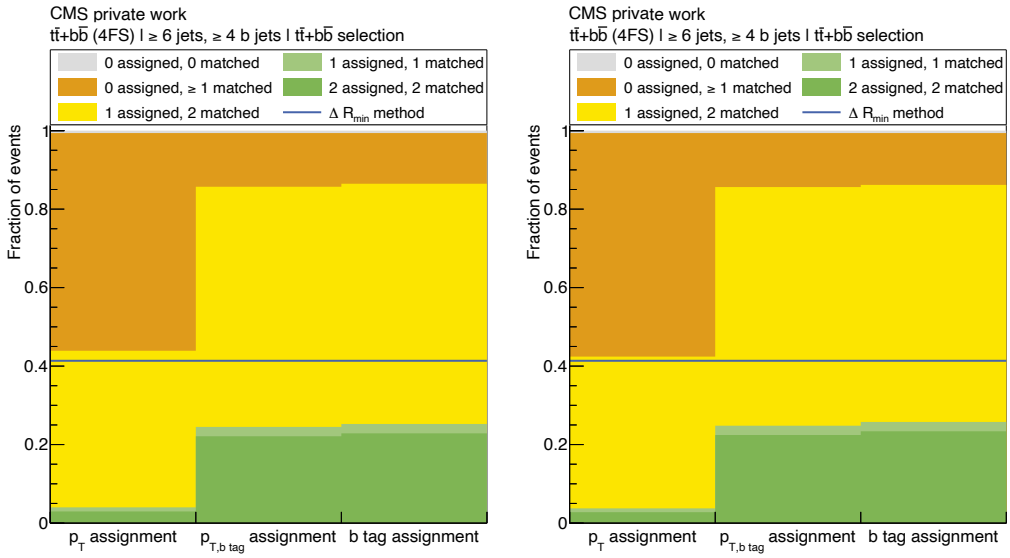
Figure 7.8: Maximum possible efficiency of the p_{T}, $p_{\mathrm{T,b\ tag}}$ and b tag assignment method for the b jets from $\mathrm{t\bar{t}}$ system reconstruction strategy assuming a perfect DNN reconstruction. It is shown how many additional b jets are assigned in which fraction of events depending on the number of jets assigned via the matching algorithm.



(a) DNN trained on $\geq$ 4 jets, $\geq$ 2 b tagged jets

(b) DNN trained on $\geq$ 6 jets, $\geq$ 4 b tagged jets

Figure 7.9: DNN performance of the b jets from $\mathrm{t\bar{t}}$ system reconstruction. Figure 7.9a shows the DNN performance based on training in the $\geq$ 4 jets, $\geq$ 2 b tagged jets region, Figure 7.9b shows the DNN performance based on training in the $\geq$ 6 jets, $\geq$ 4 b tagged jets region. It is shown how many jets are correctly assigned by the DNN in which fraction of events. The label "precisely assigned" denotes the exact assignment of each jet-type, while "totally assigned" is invariant under a permutation of two or more jet-types.



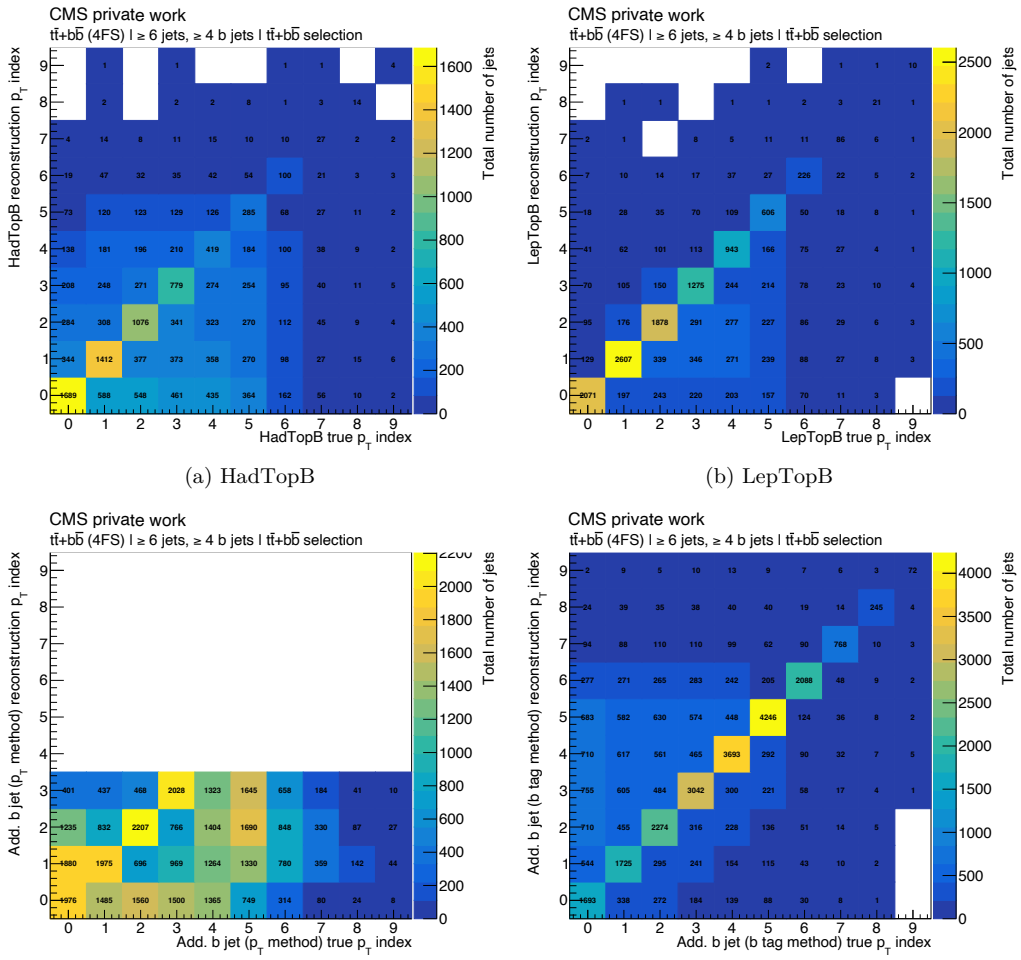
(a) DNN trained on $\geq$ 4 jets, $\geq$ 2 b tagged jets

(b) DNN trained on $\geq$ 6 jets, $\geq$ 4 b tagged jets

Figure 7.10: Assignment method performance of the b jets from $t\bar{t}$ system reconstruction. Figure 7.10a shows the assignment method performance based on the DNN which is trained in the $\geq$ 4 jets, $\geq$ 2 b tagged jets region, Figure 7.10b shows the assignment method performance based on the DNN which is trained in the $\geq$ 6 jets, $\geq$ 4 b tagged jets region. It is shown how many additional b jets are assigned in which fraction of events depending on the number of jets assigned via the matching algorithm.

A reason for the better reconstruction of LepTopB compared to HadTopB is the pairing with the lepton. The lepton can be reconstructed better than the two light flavor jets from the hadronic W decay. Thus, the composite LepTop is also better rebuilt than the composite HadTop, which leads to a better identification of the LepTopB than the HadTopB.
The results of applying the three assignment methods on the trained DNN models are shown in Figure 7.10. The assignment accuracies of the additional b jets are marginally better for the DNN that was trained on the signal region (i.e. $\geq$ 6 jets, $\geq$ 4 b tagged jets). The b tag assignment method performs best. The best accuracy of the correct assignment of the additional b jets with the b jets from $t\bar{t}$ system strategy is $a_{\text{bb from } t\bar{t}} = 0.23$. Hence, the strategy of reconstructing only the b jets from the $t\bar{t}$ system performs worse under this metric than reconstructing the entire $t\bar{t}$ system in a DNN training. The figures of the assignment method performance of the two strategies (Figures 7.7 and 7.10) also reveal that the method described in this section provides better performance to find at least one additional b jet. The $t\bar{t}$ system reconstruction strategy achieves a rate of 0.77 of all events after selection, whereas the strategy in this section achieves a rate of 0.86 for the correct assignment of at least one additional b jet.

To gain deeper insights into which assignments of the jets prevent a better accuracy, all individual jet assignments of the DNNs and the assignment methods are analyzed in the following. For this purpose the reconstructed p_{T} indices of all matchable b jets are plotted over the true p_{T} indices in confusion matrices. The confusion matrices are presented in Figure 7.11. All jets which have been correctly assigned are located on the diagonal. Figures 7.11a and 7.11b show a distinct diagonal in the confusion matrix for HadTopB



(a) HadTopB

(b) LepTopB

(c) Additional b jets, p_{T} assignment method

(d) Additional b jets, b tag assignment method

Figure 7.11: Confusion matrices of b jets for the DNN with training in the signal region. Figures 7.11a and 7.11b show the b jets from the top quark decays assigned directly by the DNN. Based on this, Figures 7.11c and 7.11d show the p_{T} and b tag assignment methods.

Table 7.5: Input variables of the DNN training on the additional b jets.

HadTopB	LepTopB	Addbb1/2	Lepton
p_{T}	p_{T}	p_{T}	p_{T}
M	M	M	
E	E	E	E
η	η	η	η
ϕ	ϕ		ϕ
b tag value	b tag value		
p_{T} index	p_{T} index		
		$\Delta\eta$	
		$\Delta\phi$	
		ΔR	
		Δp_{T}	

and LepTopB. It can also be seen that the diagonal is more prominent for LepTopB than for HadTopB. This is consistent with the finding in Figure 7.9, where it was shown that it is more probable to reconstruct the LepTopB. However, the off-diagonals of the confusion matrices are of particular interest. For both HadTopB and LepTopB the entries outside the main diagonal are asymmetrically distributed. This means the DNN tends to reconstruct high jet indices too low. In other words, in events in which the b jets from the $\mathrm{t\bar{t}}$ system have lower p_{T} values compared to the other jets in the event, the b jets are mistakenly assigned to harder p_{T} jets.
Figure 7.11c displays the confusion matrix of the additional b jets for the p_{T} assignment method. The figure indicates graphically why the p_{T} method performs poorly. Any additional b jet whose true p_{T} index is greater than three can no longer be correctly assigned using the p_{T} assignment method. The figure also demonstrates that the p_{T} method is only useful for events at LO, if at all. This does not apply to the $p_{\mathrm{T,b\ tag}}$ assignment method or the b tag assignment method, since these methods can assign the p_{T} index of the b jets up to arbitrary high jet multiplicities in an event according to the assignment rules. Figure 7.11d shows the confusion matrix of the b tag assignment method, which generally demonstrates a diagonal pattern. The entries outside the main diagonal are now consistently reversed to the previous assignments of the DNN. Accordingly, the applied method mistakenly assigns the additional b jets to jets that possess lower p_{T} values in the event. For an improved allocation of the additional b jets using the b tag method based on DNNs, the next stage should be a study to remove the identified bias of too high indices as too low in the reconstruction of DNN.

7.3.5 Additional b jets reconstruction strategy

In contrast to the first two strategies, the focus in this section is on a direct reconstruction of the additional b jets. This supersedes the three assignment methods and a direct DNN training is performed on the additional b jets. Similarly to the previous strategies, the jet kinematics of the b jets to be classified are fed into the DNN as input variables. In addition, the kinematics of the $\mathrm{b\bar{b}}$ system of the additional b jets are also passed to the DNN. Finally, the kinematics of the lepton are added to the input variables, although this information is not expected to provide a separation at all it may be used in combination with other input features. The full list of input variables is listed in Table 7.5. Some aspects of these input variables for the $\mathrm{t\bar{t}}$ and the $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ sample have already been examined and compared in detail in Chapter 6. Since these variables are indeed different depending on the modeling

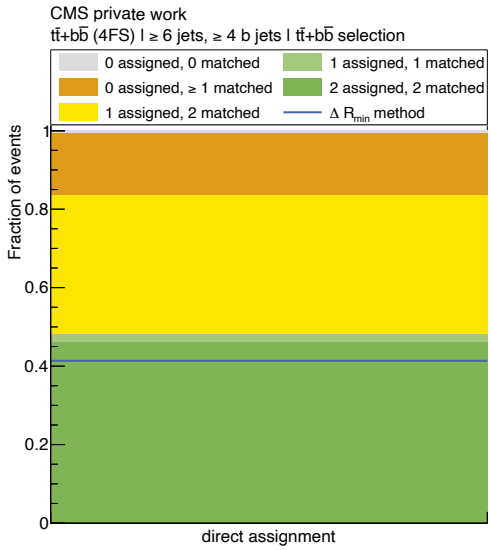


Figure 7.12: Performance of the additional b jet reconstruction with direct DNN training. It is shown how many additional b jets are assigned in which fraction of events depending on the number of jets assigned via the matching algorithm.

Table 7.6: Concise summary of the accuracies according to the metric a (equation 7.1) to assign the additional b jets. The accuracies of the $t\bar{t}+b\bar{b}$ sample are discussed in detail in the indicated sections.

Section	ΔR_{min} 7.2	m_{min} 7.2	DNN ($t\bar{t}$) 7.3.3	DNN (bb from $t\bar{t}$) 7.3.4	DNN (add. bb) 7.3.5
$t\bar{t}+b\bar{b}$ sample	0.41	0.37	0.29	0.23	0.46
$t\bar{t}$ sample	0.39	0.35	0.30	0.25	0.46

and consequently different in both data samples, three DNNs were trained in this strategy. A first DNN was trained only on the $t\bar{t}$ sample, a second DNN only on the $t\bar{t}+b\bar{b}$ sample and a third DNN, analogous to the previous strategies, on combined events of both samples. As the final evaluation of the three DNNs hardly differ from each other, only the last DNN is discussed in the following. The result of the DNN with training on the combined events is shown in Figure 7.12. The strategy of direct training achieves the best accuracy of all examined strategies with a rate $a_{\text{addbb}} = 0.46$ for the correct assignment of the additional b jets. Accordingly, the method also performs better than the ΔR_{min} method discussed in section 7.2. This is reasonable, since the DNN receives ΔR plus additional inputs and is given the chance to decide on the basis of comprehensive information. However, based on the sum of weights of the input variables in the first layer the DNN evaluates the p_{T} values of the two b jets as well as the difference between them even more important than ΔR as input variables.

7.4 Summary

All accuracies a of assigning the additional b jets are summarized in Table 7.6. All DNN trainings are performed on a ≥ 6 jets, ≥ 4 b tagged jets selection. A more inclusive selection such as ≥ 4 jets, ≥ 2 b tagged jets was tested for all DNNs, but did not show

any significant enhancements. Table 7.3 reports the applied network configuration, which achieves the best performance of all the examined configurations. The table also shows the accuracies for the evaluation of the $t\bar{t}$ sample which are not discussed in detail in this thesis. The accuracies in the table demonstrate, as stated at the beginning of this chapter, how similar they are to the accuracies of the $t\bar{t}$+$b\bar{b}$ sample.
The highest accuracy is achieved with the DNN reconstruction method if trained directly on the additional b jets (0.46 for both samples). DNN reconstruction methods that are trained on the $t\bar{t}$ system or parts of it are independent of the modeling of the additional b jets, but perform rather weakly (accuracy of 0.3 and lower). Determining all distances between b jets and selecting the jets whose distance is the smallest constitutes a straightforward but efficient method with moderate accuracy. The observable ΔR_{min} leads to an accuracy of 0.41 for the $t\bar{t}$+$b\bar{b}$ sample and 0.39 for the $t\bar{t}$ sample. Both the DNN reconstruction method with a training directly on the additional b jets and the ΔR_{min} method depend on the modeling of the additional b jets. In Chapter 6, the distribution for the ΔR is examined for different simulation approaches and observed to show the largest differences among the investigated observables. The DNN reconstruction method with a training directly on the additional b jets also uses this information, but based on the sum of the weights of the input variables in the first layer, it is not the most important feature. The DNN evaluates the p_{T} values as particularly important, which show a comparatively smaller modeling dependence in the studies in Chapter 6.

8 Conclusion

In this thesis, the associated production of a pair of top quarks with bottom quarks ($t\bar{t}+b\bar{b}$) is studied in proton-proton collisions at the CERN Large Hadron Collider (LHC). These $t\bar{t}+b\bar{b}$ events contain decisive QCD processes and are of particular relevance for the LHC physics program for several reasons.
On the one hand, the process incorporates two very different energy scales, since the top quark is much heavier than the b quark. This multiscale QCD nature makes the process particularly intriguing and challenging in Monte Carlo event simulations. On the other hand, events with $t\bar{t}+b\bar{b}$ processes represent a large irreducible background in measurements of $t\bar{t}$+H production with H $\rightarrow b\bar{b}$ decays, which allow for a direct probe of the top-Higgs Yukawa coupling. These measurements are an important test of the Standard Model and help to constrain models for physics beyond the Standard Model that predict different coupling strengths. Hence, it is crucial to gain a thorough understanding of the $t\bar{t}+b\bar{b}$ process. To this end, $t\bar{t}+b\bar{b}$ events are examined in the single-lepton decay channel of the $t\bar{t}$ system in this thesis. In two studies, different objectives are addressed.

The first study on generator level focuses on a comparison of different MC generators for $t\bar{t}+b\bar{b}$ events at the CMS and ATLAS experiment. This comparison is conducted within the LHC Higgs Working Group, with the studies on the CMS side being performed as a part of this thesis. The comparison is accomplished in close cooperation with ATLAS, with the compilation of the final comparative distributions being handled by the ATLAS collaboration. Initially, the existing simulation approaches and programs are described at a technical level and the main differences in configurations are pointed out. Based on this, an object and event selection as well as validation observables are defined by which the simulated events are compared. The analysis routine is written in the Rivet framework.

The result of the study is a detailed analysis of the behavior in various observables of simulated events for $t\bar{t}+b\bar{b}$ processes. A total of four simulation approaches and programs are considered for CMS: $t\bar{t}$ Powheg+Pythia8, $t\bar{t}+b\bar{b}$ Powheg+Pythia8, $t\bar{t}+b\bar{b}$ Sherpa and $t\bar{t}+b\bar{b}$ MadGraph5_aMC@NLO. The goal is not only the comparison between several MC generators, but also particularly the comparison of different approaches for modeling additional b jets not associated with the top quark pair. In $t\bar{t}$ simulation approaches, the additional b jets arise from the parton shower, while in $t\bar{t}+b\bar{b}$ simulation approaches the additional b jets are calculated using matrix elements.

It is found that the largest difference between $t\bar{t}$ and $t\bar{t}+b\bar{b}$ simulated events can be observed in the distance ΔR between the closest two b jets. The $t\bar{t}+b\bar{b}$ simulated events feature

considerably smaller values in the ΔR observable compared to the $\mathrm{t\bar{t}}$ simulated events. Beyond that, the jet multiplicity and H_T distributions show distinct differences. Further observables such as p_T or $m\,(\mathrm{bb})$ of the two b jets with the highest p_T or the two closest b jets differ moderately. The three $\mathrm{t\bar{t}}+\mathrm{b\bar{b}}$ simulated event distributions also differ among themselves, but generally show similar behavior with respect to the $\mathrm{t\bar{t}}$ POWHEG+PYTHIA8 simulated events.

The differences between the $\mathrm{t\bar{t}}$ POWHEG+PYTHIA8 simulated events from ATLAS and CMS are found to be small. The $\mathrm{t\bar{t}}+\mathrm{b\bar{b}}$ POWHEG+PYTHIA8 simulated events from ATLAS and CMS show similar trends among themselves and also appear comparable with the $\mathrm{t\bar{t}}+\mathrm{b\bar{b}}$ SHERPA simulated events. Generally, the $\mathrm{t\bar{t}}+\mathrm{b\bar{b}}$ simulated events also reveal similar characteristics compared to $\mathrm{t\bar{t}}$ simulated events as found in the CMS only analysis. It is disclosed that slightly different parameter settings as well as versions of the generators are used at both experiments. The choices of the renormalization and factorization scales are identical for the $\mathrm{t\bar{t}}$ simulated events, but different for all $\mathrm{t\bar{t}}+\mathrm{b\bar{b}}$ simulated events between ATLAS and CMS.

The results of the study constitute the foundation for future discussions on developing a common approach to $\mathrm{t\bar{t}}$ event modeling. This is decisive for modeling the backgrounds and uncertainties in future measurements of $\mathrm{t\bar{t}}+\mathrm{H}$ production with $\mathrm{H} \to \mathrm{b\bar{b}}$ decays.

The second study on reconstruction level focuses on the identification of the additional b jets to identify the origin of the jets in the event. In the final state of a $\mathrm{t\bar{t}}+\mathrm{b\bar{b}}$ event it is unknown which b jets originate from top quark decays and which b jets are not associated with the top quarks. In this study, two methods for the identification of the additional b jets are investigated. An accuracy metric is defined, by which all methods can be compared. For each method, it is determined in how many events the additional b jets are correctly identified relative to all events after a generator level selection on $\mathrm{t\bar{t}}+\mathrm{b\bar{b}}$ signatures in the ≥ 6 jets, ≥ 4 b tagged jets region.

The first method applies the most straightforward approach possible and identifies the observable ΔR_min as the observable by which the additional b jets can be identified most accurately. Assigning the closest two b jets in an event as additional b jets results in an accuracy of approximately 0.4 in the aforementioned phase space.

The second method applies a sophisticated approach via training of deep neural networks. Three different sub-methods are performed in this refined concept: training on the entire $\mathrm{t\bar{t}}$ system, training only on the b jets of the $\mathrm{t\bar{t}}$ system and direct training on the additional b jets. Thus, the first two sub-methods do not directly consider the additional b jets, but only the $\mathrm{t\bar{t}}$ system or parts of it. Based on this, the additional b jets are identified by either the residual highest two p_T jets, the highest two p_T jets which are b tagged, or the two b jets with the highest b tag values in the event. The underlying idea of the two indirect sub-methods is to be independent of the modeling of the additional b jets.
Direct training on the additional b jets achieves an accuracy of 0.46 for the identification of both b jets. The indirect reconstruction methods achieve an accuracy of 0.3 and lower.

To improve the performance of the deep neural networks, hyper parameters, inputs and phase space regions are extensively researched. It is found that the deep neural networks tend to assign b jets with lower p_T values to harder p_T jets in the event. Novel reconstruction methods could eliminate this tendency and increase the accuracy. The results of this study are used to define the strategies of identifying observables in ongoing efforts for differential measurements of the $\mathrm{t\bar{t}}+\mathrm{b\bar{b}}$ process at the CMS experiment. These differential measurements will give important inputs to the theory community for future development of the $\mathrm{t\bar{t}}+\mathrm{b\bar{b}}$ process modeling.

Bibliography

[1] F. Englert and R. Brout. "Broken Symmetry and the Mass of Gauge Vector Mesons". In: *Phys. Rev. Lett.* 13 (9 Aug. 1964), pp. 321–323. DOI: 10.1103/PhysRevLett.13.321.

[2] P. W. Higgs. "Broken Symmetries and the Masses of Gauge Bosons". In: *Phys. Rev. Lett.* 13 (16 Oct. 1964), pp. 508–509. DOI: 10.1103/PhysRevLett.13.508.

[3] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble. "Global Conservation Laws and Massless Particles". In: *Phys. Rev. Lett.* 13 (20 Nov. 1964), pp. 585–587. DOI: 10.1103/PhysRevLett.13.585.

[4] The ATLAS Collaboration. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Physics Letters B* 716.1 (2012), pp. 1–29. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2012.08.020.

[5] The CMS Collaboration. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". In: *Physics Letters B* 716.1 (2012), pp. 30–61. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2012.08.021.

[6] A. M. Sirunyan et al. "Measurement of the cross section for $t\bar{t}$ production with additional jets and b jets in pp collisions at $\sqrt{s}$ =13 TeV". In: *Journal of High Energy Physics* 2020.7 (July 2020). ISSN: 1029-8479. DOI: 10.1007/jhep07(2020)125.

[7] E. Fermi. "Versuch einer Theorie der β-Strahlen. I". In: *Zeitschrift fuer Physik* 88.3-4 (Mar. 1934), pp. 161–177. DOI: 10.1007/BF01351864.

[8] S. Tomonaga. "On a Relativistically Invariant Formulation of the Quantum Theory of Wave Fields". In: *Progress of Theoretical Physics* 1.2 (Aug. 1946), pp. 27–42. ISSN: 0033-068X. DOI: 10.1143/PTP.1.27.

[9] J. Schwinger. "Quantum Electrodynamics. I. A Covariant Formulation". In: *Phys. Rev.* 74 (10 Nov. 1948), pp. 1439–1461. DOI: 10.1103/PhysRev.74.1439.

[10] R. P. Feynman. "The Theory of Positrons". In: *Phys. Rev.* 76 (6 Sept. 1949), pp. 749–759. DOI: 10.1103/PhysRev.76.749.

[11] M. Gell-Mann. *The Eightfold Way: A Theory of strong interaction symmetry*. Tech. rep. Mar. 1961. DOI: 10.2172/4008239.

[12] S. L. Glashow. "Partial-symmetries of weak interactions". In: *Nuclear Physics* 22.4 (1961), pp. 579–588. ISSN: 0029-5582. DOI: 10.1016/0029-5582(61)90469-2.

[13] M. Gell-Mann. "Symmetries of Baryons and Mesons". In: *Phys. Rev.* 125 (3 Feb. 1962), pp. 1067–1084. DOI: 10.1103/PhysRev.125.1067.

[14] J. Goldstone, A. Salam, and S. Weinberg. "Broken Symmetries". In: *Phys. Rev.* 127 (3 Aug. 1962), pp. 965–970. DOI: 10.1103/PhysRev.127.965.

[15] M. Gell-Mann. "A schematic model of baryons and mesons". In: *Physics Letters* 8.3 (1964), pp. 214–215. ISSN: 0031-9163. DOI: 10.1016/S0031-9163(64)92001-3.

[16] A. Salam and J. Ward. "Electromagnetic and weak interactions". In: *Physics Letters* 13.2 (1964), pp. 168–171. ISSN: 0031-9163. DOI: 10.1016/0031-9163(64)90711-5.

[17] S. Weinberg. "A Model of Leptons". In: *Phys. Rev. Lett.* 19 (21 Nov. 1967), pp. 1264–1266. DOI: 10.1103/PhysRevLett.19.1264.

[18] S. L. Glashow, J. Iliopoulos, and L. Maiani. "Weak Interactions with Lepton-Hadron Symmetry". In: *Phys. Rev. D* 2 (7 Oct. 1970), pp. 1285–1292. DOI: 10.1103/PhysRevD.2.1285.

[19] M. Kobayashi and T. Maskawa. "CP-Violation in the Renormalizable Theory of Weak Interaction". In: *Progress of Theoretical Physics* 49.2 (Feb. 1973), pp. 652–657. ISSN: 0033-068X. DOI: 10.1143/PTP.49.652.

[20] C. S. Wu et al. "Experimental Test of Parity Conservation in Beta Decay". In: *Phys. Rev.* 105 (4 Feb. 1957), pp. 1413–1415. DOI: 10.1103/PhysRev.105.1413.

[21] M. Goldhaber, L. Grodzins, and A. W. Sunyar. "Helicity of Neutrinos". In: *Phys. Rev.* 109 (3 Feb. 1958), pp. 1015–1017. DOI: 10.1103/PhysRev.109.1015.

[22] J. E. Augustin et al. "Discovery of a Narrow Resonance in e^+e^- Annihilation". In: *Phys. Rev. Lett.* 33 (23 Dec. 1974), pp. 1406–1408. DOI: 10.1103/PhysRevLett.33.1406.

[23] J. J. Aubert et al. "Experimental Observation of a Heavy Particle J". In: *Phys. Rev. Lett.* 33 (23 Dec. 1974), pp. 1404–1406. DOI: 10.1103/PhysRevLett.33.1404.

[24] S. W. Herb et al. "Observation of a Dimuon Resonance at 9.5 GeV in 400 GeV-Proton-Nucleus Collisions". In: *Phys. Rev. Lett.* 39 (5 Aug. 1977), pp. 252–255. DOI: 10.1103/PhysRevLett.39.252.

[25] M. Banner et al. "Observation of single isolated electrons of high transverse momentum in events with missing transverse energy at the CERN pp collider". In: *Physics Letters B* 122.5 (1983), pp. 476–485. ISSN: 0370-2693. DOI: 10.1016/0370-2693(83)91605-2.

[26] The UA1 Collaboration. "Experimental observation of isolated large transverse energy electrons with associated missing energy at $\sqrt{s} = 540$ GeV". In: *Physics Letters B* 122.1 (1983), pp. 103–116. ISSN: 0370-2693. DOI: 10.1016/0370-2693(83)91177-2.

[27] The UA1 Collaboration. "Experimental observation of lepton pairs of invariant mass around 95 GeV/c^2 at the CERN SPS collider". In: *Physics Letters B* 126.5 (1983), pp. 398–410. ISSN: 0370-2693. DOI: 10.1016/0370-2693(83)90188-0.

[28] The D0 Collaboration. "Observation of the Top Quark". In: *Phys. Rev. Lett.* 74 (14 Apr. 1995), pp. 2632–2637. DOI: 10.1103/PhysRevLett.74.2632.

[29] The CDF Collaboration. "Observation of Top Quark Production in $\overline{p}p$ Collisions with the Collider Detector at Fermilab". In: *Phys. Rev. Lett.* 74 (14 Apr. 1995), pp. 2626–2631. DOI: 10.1103/PhysRevLett.74.2626.

[30] The NA48 Collaboration. "A new measurement of direct CP violation in two pion decays of the neutral kaon". In: *Physics Letters B* 465.1 (1999), pp. 335–348. ISSN: 0370-2693. DOI: 10.1016/S0370-2693(99)01030-8.

[31] R. Wolf. *The Higgs Boson Discovery at the Large Hadron Collider*. Vol. 264. Springer, 2015. ISBN: 9783319185125, 9783319185118. DOI: 10.1007/978-3-319-18512-5.

[32] J. van der Linden. "Limit on $\mathrm{t\bar{t}} + Z$ production in the $Z \to \mathrm{b\bar{b}}$ channel at the CMS experiment". Master thesis. Karlsruhe Institute of Technology (KIT), 2019.

[33] E. Noether. "Invariante Variationsprobleme". ger. In: *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse* 1918 (1918), pp. 235–257. URL: http://eudml.org/doc/59024.

[34] H. Yukawa. "On the Interaction of Elementary Particles. I". In: *Progress of Theoretical Physics Supplement* 1 (Jan. 1955), pp. 1–10. ISSN: 0375-9687. DOI: 10.1143/PTPS.1.1.

[35] P. A. Zyla et al. "Review of Particle Physics". In: *Progress of Theoretical and Experimental Physics* 2020.8 (Aug. 2020). 083C01. ISSN: 2050-3911. DOI: 10.1093/ptep/ptaa104.

[36] M. Aker et al. "Improved Upper Limit on the Neutrino Mass from a Direct Kinematic Method by KATRIN". In: *Physical Review Letters* 123.22 (Nov. 2019). ISSN: 1079-7114. DOI: 10.1103/physrevlett.123.221802.

[37] P. A. Dirac. "Quantum theory of emission and absorption of radiation". In: *Proc. Roy. Soc. Lond. A* 114 (1927), p. 243. DOI: 10.1098/rspa.1927.0039.

[38] R. P. Feynman. "Space-Time Approach to Quantum Electrodynamics". In: *Phys. Rev.* 76 (6 Sept. 1949), pp. 769–789. DOI: 10.1103/PhysRev.76.769.

[39] J. D. Bjorken and E. A. Paschos. "Inelastic Electron-Proton and γ-Proton Scattering and the Structure of the Nucleon". In: *Phys. Rev.* 185 (5 Sept. 1969), pp. 1975–1982. DOI: 10.1103/PhysRev.185.1975.

[40] The NNPDF Collaboration. *Parton distributions from high-precision collider data*. 2017. arXiv: 1706.00428 [hep-ph].

[41] G. Altarelli and G. Parisi. "Asymptotic freedom in parton language". In: *Nuclear Physics B* 126.2 (1977), pp. 298–318. ISSN: 0550-3213. DOI: 10.1016/0550-3213(77)90384-4.

[42] E. Mobs. "The CERN accelerator complex - 2019. Complexe des accélérateurs du CERN - 2019". General Photo. July 2019. URL: https://cds.cern.ch/record/2684277.

[43] J. T. Boyd. *LHC Run-2 and Future Prospects*. 2020. arXiv: 2001.04370 [hep-ex].

[44] L. Evans and P. Bryant. "LHC Machine". In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08001–S08001. DOI: 10.1088/1748-0221/3/08/s08001.

[45] The ALICE Collaboration. "The ALICE experiment at the CERN LHC". In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08002–S08002. DOI: 10.1088/1748-0221/3/08/s08002.

[46] The ATLAS Collaboration. "The ATLAS Experiment at the CERN Large Hadron Collider". In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08003–S08003. DOI: 10.1088/1748-0221/3/08/s08003.

[47] The CMS Collaboration. "The CMS experiment at the CERN LHC". In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08004–S08004. DOI: 10.1088/1748-0221/3/08/s08004.

[48] The LHCb Collaboration. "The LHCb Detector at the LHC". In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08005–S08005. DOI: 10.1088/1748-0221/3/08/s08005.

[49] S. R. Davis. "Interactive Slice of the CMS detector". Aug. 2016. URL: https://cds.cern.ch/record/2205172.

[50] R. Gluckstern. "Uncertainties in track momentum and direction, due to multiple scattering and measurement errors". In: *Nuclear Instruments and Methods* 24 (1963), pp. 381–389. ISSN: 0029-554X. DOI: 10.1016/0029-554X(63)90347-1.

[51] The CMS Collaboration. "The CMS trigger system". In: *Journal of Instrumentation* 12.01 (Jan. 2017), P01020–P01020. DOI: 10.1088/1748-0221/12/01/p01020.

[52] P. Billoir. "Progressive track recognition with a Kalman-like fitting procedure". In: *Computer Physics Communications* 57.1 (1989), pp. 390–394. ISSN: 0010-4655. DOI: 10.1016/0010-4655(89)90249-X.

[53] The CMS Collaboration. "Description and performance of track and primary-vertex reconstruction with the CMS tracker". In: *Journal of Instrumentation* 9.10 (Oct. 2014), P10009–P10009. DOI: 10.1088/1748-0221/9/10/p10009.

[54] A. M. Sirunyan et al. "Particle-flow reconstruction and global event description with the CMS detector". In: *JINST* 12 (2017), P10003. DOI: 10.1088/1748-0221/12/10/P10003.

[55] M. Cacciari, G. P. Salam, and G. Soyez. "The anti-kt jet clustering algorithm". In: *Journal of High Energy Physics* 2008.04 (Apr. 2008), pp. 063–063. DOI: 10.1088/1126-6708/2008/04/063.

[56] The CMS Collaboration. "Performance of the DeepJet b tagging algorithm using 41.9/fb of data from proton-proton collisions at 13 TeV with Phase 1 CMS detector". Nov. 2018. URL: https://cds.cern.ch/record/2646773.

[57] TWiki. *Heavy flavour tagging for 13 TeV data in 2018 and 10_2_X MC*. https://twiki.cern.ch/twiki/bin/viewauth/CMS/BtagRecommendation102X r20 (internal documentation). 2020.

[58] The CMS Collaboration. "Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s} = 13$ TeV using the CMS detector". In: *Journal of Instrumentation* 14.07 (July 2019), P07004–P07004. DOI: 10.1088/1748-0221/14/07/p07004.

[59] The ATLAS Collaboration. "Search for the standard model Higgs boson produced in association with top quarks and decaying into a $b\bar{b}$ pair in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: *Phys. Rev. D* 97 (7 Apr. 2018), p. 072016. DOI: 10.1103/PhysRevD.97.072016.

[60] D. de Florian et al. "Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector". In: *arXiv e-prints*, arXiv:1610.07922 (Oct. 2016). eprint: 1610.07922 (hep-ph).

[61] The CMS Collaboration. *Measurement of* $\mathrm{t\bar{t}H}$ *production in the* $\mathrm{H} \to \mathrm{b\bar{b}}$ *decay channel in* $41.5\,\mathrm{fb}^{-1}$ *of proton-proton collision data at* $\sqrt{s} = 13\,\mathrm{TeV}$. Tech. rep. CMS-PAS-HIG-18-030. Geneva: CERN, 2019. URL: http://cds.cern.ch/record/2675023.

[62] The CMS Collaboration. "Measurements of $\mathrm{t\bar{t}}$ cross sections in association with b jets and inclusive jets and their ratio using dilepton final states in pp collisions at $\sqrt{s} = 13\,\mathrm{TeV}$". In: *Physics Letters B* 776 (2018), pp. 355–378. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2017.11.043.

[63] T. Ježo et al. "New NLOPS predictions for $t\bar{t} + b$ -jet production at the LHC". In: *The European Physical Journal C* 78.6 (June 2018). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-018-5956-0.

[64] B. Andersson et al. "Parton fragmentation and string dynamics". In: *Physics Reports* 97.2 (1983), pp. 31–145. ISSN: 0370-1573. DOI: 10.1016/0370-1573(83)90080-7.

[65] B. Webber. "A QCD model for jet fragmentation including soft gluon interference". In: *Nuclear Physics B* 238.3 (1984), pp. 492–528. ISSN: 0550-3213. DOI: 10.1016/0550-3213(84)90333-X.

[66] A. Buckley et al. "General-purpose event generators for LHC physics". In: *Physics Reports* 504.5 (July 2011), pp. 145–233. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2011.03.005.

[67] The ATLAS Collaboration. *Proposal for truth particle observable definitions in physics measurements*. Tech. rep. ATL-PHYS-PUB-2015-013. Geneva: CERN, June 2015. URL: https://cds.cern.ch/record/2022743.

[68] M. Cacciari, G. P. Salam, and G. Soyez. "The catchment area of jets". In: *Journal of High Energy Physics* 2008.04 (Apr. 2008), pp. 005–005. DOI: 10.1088/1126-6708/2008/04/005.

[69] TWiki. *Hadron based origin identification of heavy flavour jets at generator level*. https://twiki.cern.ch/twiki/bin/view/CMSPublic/GenHFHadronMatcher r30. 2017.

[70] TWiki. *LHC Higgs Working Group*. https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHWG r384. 2020.

[71] P. Nason. "A new method for combining NLO QCD with shower Monte Carlo algorithms". In: *JHEP* 11 (2004), p. 040. DOI: 10.1088/1126-6708/2004/11/040.

[72] S. Frixione, P. Nason, and C. Oleari. "Matching NLO QCD computations with parton shower simulations: the POWHEG method". In: *JHEP* 11 (2007), p. 070. DOI: 10.1088/1126-6708/2007/11/070.

[73] S. Alioli et al. "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX". In: *JHEP* 06 (2010), p. 043. DOI: 10.1007/JHEP06(2010)043.

[74] J. M. Campbell et al. "Top-pair production and decay at NLO matched with parton showers". In: *Journal of High Energy Physics* 2015.4 (Apr. 2015). ISSN: 1029-8479. DOI: 10.1007/jhep04(2015)114.

[75] T. Sjöstrand et al. "An Introduction to PYTHIA 8.2". In: *Comput. Phys. Commun.* 191 (2015), p. 159. DOI: 10.1016/j.cpc.2015.01.024.

[76] A. M. Sirunyan et al. "Extraction and validation of a new set of CMS pythia8 tunes from underlying-event measurements". In: *The European Physical Journal C* 80.1 (Jan. 2020). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-019-7499-4.

[77] A. M. Sirunyan et al. "Measurement of jet substructure observables in $t\bar{t}$ events from proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$". In: *Phys. Rev. D* 98 (9 Nov. 2018), p. 092014. DOI: 10.1103/PhysRevD.98.092014.

[78] The CMS Collaboration. "Search for $\mathrm{t\bar{t}H}$ production in the $\mathrm{H} \to \mathrm{b\bar{b}}$ decay channel with leptonic $\mathrm{t\bar{t}}$ decays in proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$". In: *Journal of High Energy Physics* 2019.3 (Mar. 2019). ISSN: 1029-8479. DOI: 10.1007/jhep03(2019)026.

[79] F. Maltoni, G. Ridolfi, and M. Ubiali. "b-initiated processes at the LHC: a reappraisal". In: *Journal of High Energy Physics* 2012.7 (July 2012). ISSN: 1029-8479. DOI: 10.1007/jhep07(2012)022.

[80] S. Frixione and M. L. Mangano. "Heavy-quark jets in hadronic collisions". In: *Nuclear Physics B* 483.1 (1997), pp. 321–338. ISSN: 0550-3213. DOI: 10.1016/S0550-3213(96)00577-9.

[81] T. Jezo. "NLO matching for $t\bar{t}b\bar{b}$ production with massive b-quarks". In: *PoS* DIS2018 (2018), p. 089. DOI: 10.22323/1.316.0089.

[82] T. Ježo and P. Nason. "On the treatment of resonances in next-to-leading order calculations matched to a parton shower". In: *Journal of High Energy Physics* 2015.12 (Dec. 2015), pp. 1–47. ISSN: 1029-8479. DOI: 10.1007/jhep12(2015)065.

[83] F. Cascioli et al. "NLO matching for $t\bar{t}b\bar{b}$ production with massive b-quarks". In: *Phys. Lett.* B734 (2014), pp. 210–214. DOI: 10.1016/j.physletb.2014.05.040. eprint: 1309.5912 (hep-ph).

[84] R. D. Ball et al. "Parton distributions for the LHC run II". In: *Journal of High Energy Physics* 2015.4 (Apr. 2015). ISSN: 1029-8479. DOI: 10.1007/jhep04(2015)040.

[85] S. Schumann and F. Krauss. "A Parton shower algorithm based on Catani-Seymour dipole factorisation". In: *JHEP* 03 (2008), p. 038. DOI: 10.1088/1126-6708/2008/03/038.

[86] J. Alwall et al. "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations". In: *Journal of High Energy Physics* 2014.7 (July 2014). ISSN: 1029-8479. DOI: 10.1007/jhep07(2014)079.

[87] ATLAS Collaboration. *ATLAS Pythia 8 tunes to 7 TeV data*. ATL-PHYS-PUB-2014-021. 2014. URL: https://cds.cern.ch/record/1966419.

[88] S. Mrenna and P. Skands. "Automated parton-shower variations in pythia 8". In: *Phys. Rev. D* 94 (7 Oct. 2016), p. 074005. DOI: 10.1103/PhysRevD.94.074005.

[89] P. Skands, S. Carrazza, and J. Rojo. "Tuning PYTHIA 8.1: the Monash 2013 tune". In: *The European Physical Journal C* 74.8 (Aug. 2014). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-014-3024-y.

[90] A. Buckley et al. *Rivet user manual*. 2013. arXiv: 1003.0694 [hep-ph].

[91] C. Bierlich et al. "Robust Independent Validation of Experiment and Theory: Rivet version 3". In: *SciPost Physics* 8.2 (Feb. 2020). ISSN: 2542-4653. DOI: 10.21468/scipostphys.8.2.026.

[92] A. Knue and E. Pfeffer. *A $t\bar{t}$+$b\bar{b}$ Rivet routine*. https://github.com/emanuelpf/rivet3-analyses/. 2020.

[93] R. Rojas. *Neural Networks - A Systematic Introduction*. Springer-Verlag, 1996. URL: http://page.mi.fu-berlin.de/rojas/neural/.

[94] F. Chollet et al. *Keras*. https://keras.io. 2015.

[95] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. "Learning representations by back-propagating errors". In: *Nature* 323 (1986), pp. 533–536. DOI: 10.1038/323533a0.

[96] C. Nwankpa et al. *Activation Functions: Comparison of trends in Practice and Research for Deep Learning*. 2018. arXiv: 1811.03378 [cs.LG].

[97] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN: 0262018020.

[98] Martıín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[99] M. D. Zeiler. *ADADELTA: An Adaptive Learning Rate Method*. 2012. arXiv: 1212.5701 [cs.LG].

[100] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. URL: %5Curl%7Bhttp://www.deeplearningbook.org%7D.

[101] N. Srivastava et al. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In: *Journal of Machine Learning Research* 15.56 (2014), pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

Appendix

A CMS MC generator comparison

In the following all distributions of the validation observable (see Table 6.3) of the ‘‘1 lepton, $\geq$ 4 jets, 3 b jets’’ category for the $t\bar{t}$ POWHEG+PYTHIA8, $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8, $t\bar{t}$+$b\bar{b}$ SHERPA and $t\bar{t}$+$b\bar{b}$ MG5AMC(NLO) simulation can be found as discussed in chapter 6. This is followed by category ‘‘1 lepton, $\geq$ 4 jets, $\geq$ 4 b jets’’.

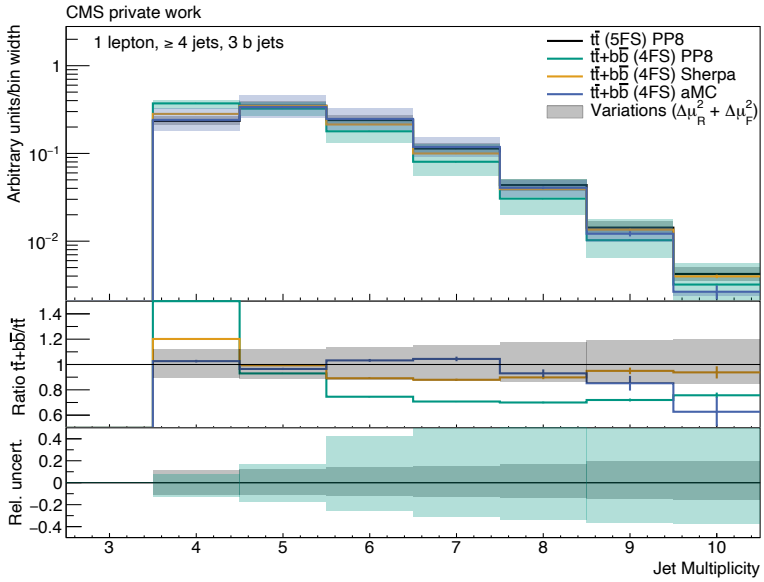


Figure A.1: Jet multiplicity of the $t\bar{t}$ Powheg+Pythia8 (black), $t\bar{t}$+$b\bar{b}$ Powheg+Pythia8 (green), $t\bar{t}$+$b\bar{b}$ Sherpa (orange) and $t\bar{t}$+$b\bar{b}$ MG5aMC-(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_R and μ_F are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, 3 b jets".

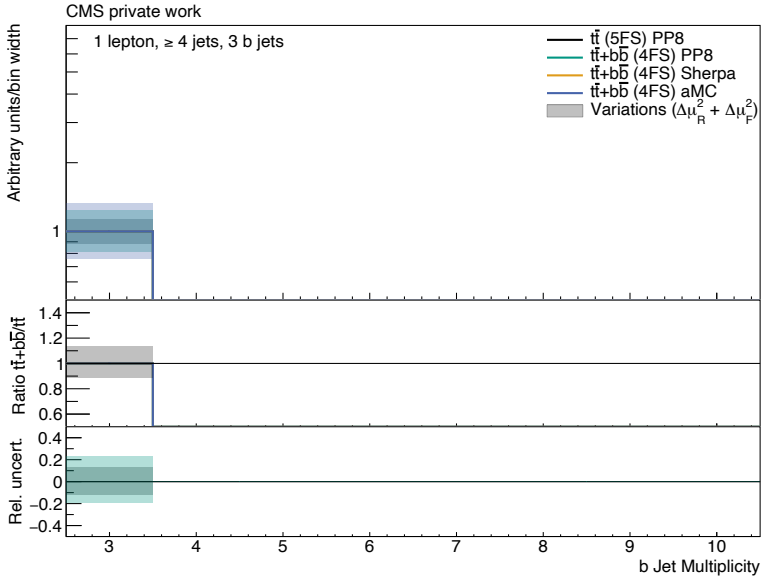


Figure A.2: b jet multiplicity of the $t\bar{t}$ Powheg+Pythia8 (black), $t\bar{t}$+$b\bar{b}$ Powheg+Pythia8 (green), $t\bar{t}$+$b\bar{b}$ Sherpa (orange) and $t\bar{t}$+$b\bar{b}$ MG5aMC-(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_R and μ_F are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, 3 b jets".

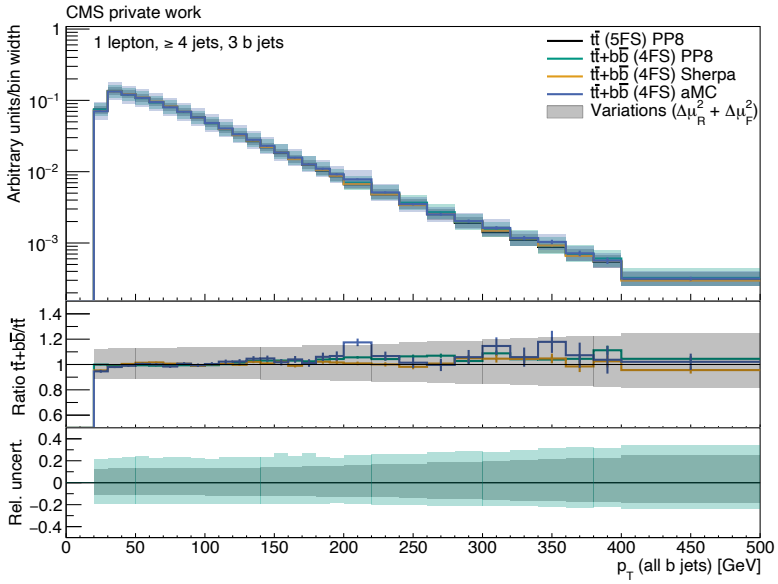


Figure A.3: p_{T} (all b jets) of the $\mathrm{t\bar{t}}$ POWHEG+PYTHIA8 (black), $\mathrm{t\bar{t}+b\bar{b}}$ POWHEG+PYTHIA8 (green), $\mathrm{t\bar{t}+b\bar{b}}$ SHERPA (orange) and $\mathrm{t\bar{t}+b\bar{b}}$ MG5AMC(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_{R} and μ_{F} are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, 3 b jets".

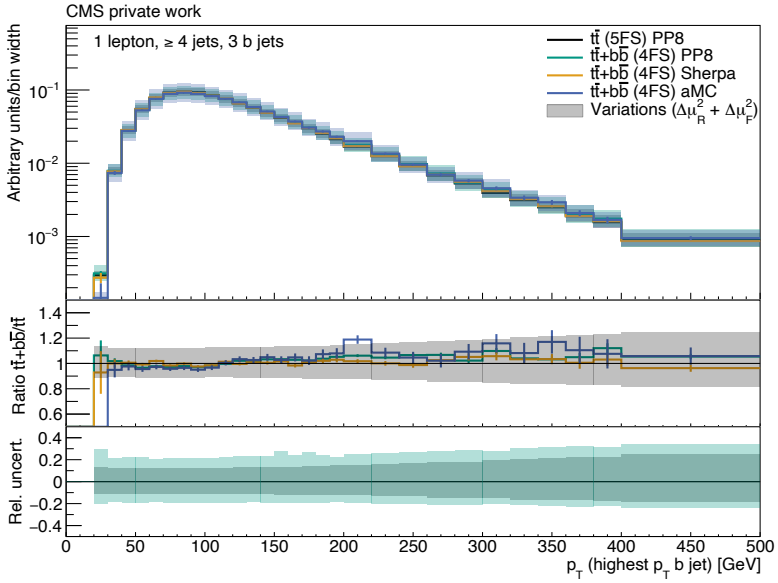


Figure A.4: Leading b jet p_{T} of the $\mathrm{t\bar{t}}$ POWHEG+PYTHIA8 (black), $\mathrm{t\bar{t}+b\bar{b}}$ POWHEG+PYTHIA8 (green), $\mathrm{t\bar{t}+b\bar{b}}$ SHERPA (orange) and $\mathrm{t\bar{t}+b\bar{b}}$ MG5AMC-(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_{R} and μ_{F} are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, 3 b jets".

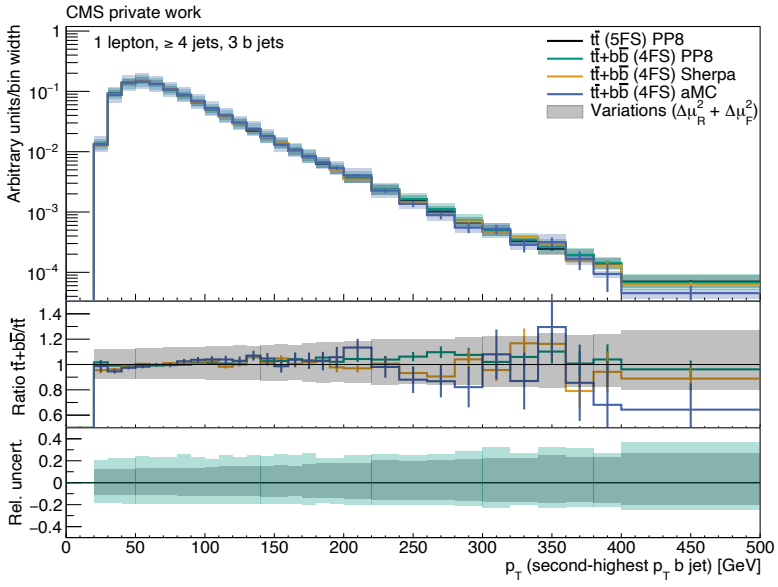


Figure A.5: Sub-leading b jet p_{T} of the $\mathrm{t\bar{t}}$ POWHEG+PYTHIA8 (black), $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ POWHEG+PYTHIA8 (green), $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ SHERPA (orange) and $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ MG5AMC-(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_{R} and μ_{F} are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, 3 b jets".

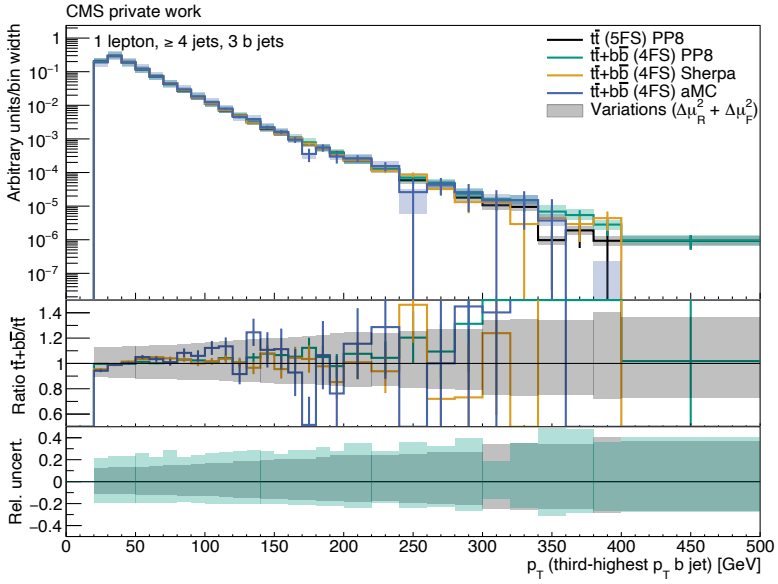


Figure A.6: Third b jet p_{T} of the $\mathrm{t\bar{t}}$ POWHEG+PYTHIA8 (black), $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ POWHEG+PYTHIA8 (green), $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ SHERPA (orange) and $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ MG5AMC(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_{R} and μ_{F} are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, 3 b jets".

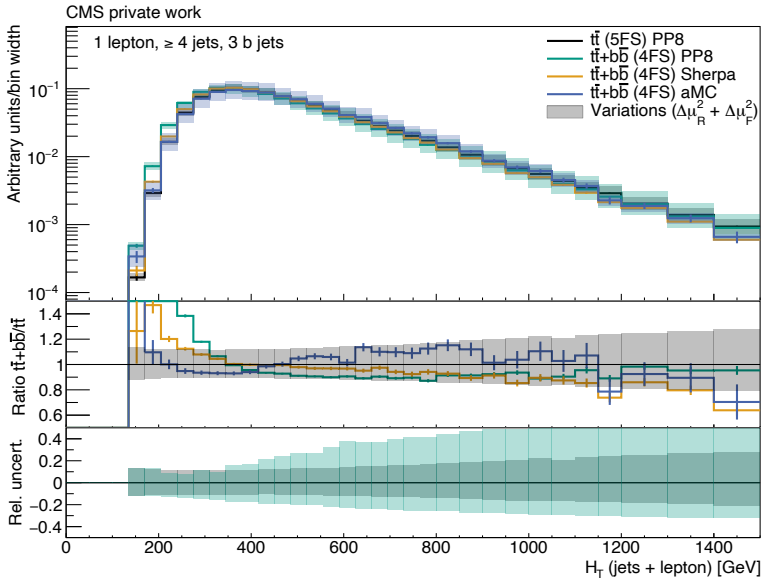


Figure A.7: H_T (jets+lepton) of the $t\bar{t}$ POWHEG+PYTHIA8 (black), $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8 (green), $t\bar{t}$+$b\bar{b}$ SHERPA (orange) and $t\bar{t}$+$b\bar{b}$ MG5AMC-(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_R and μ_F are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, 3 b jets".

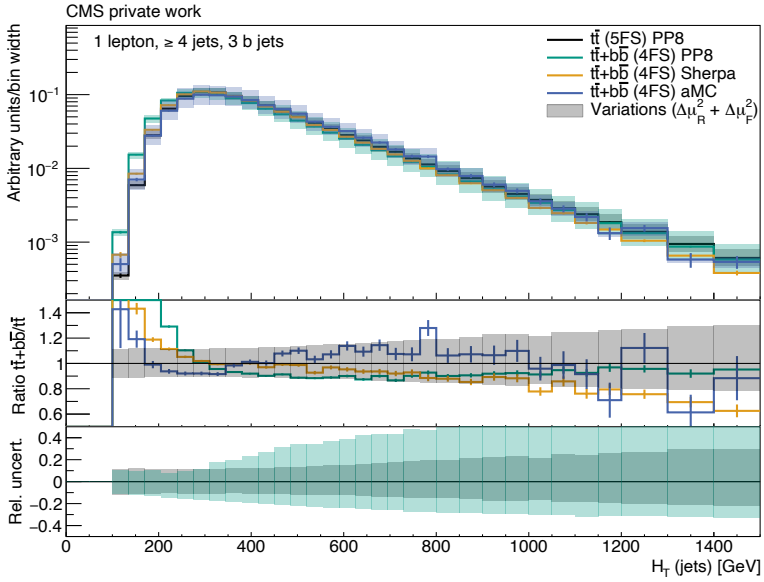


Figure A.8: H_T (jets) of the $t\bar{t}$ POWHEG+PYTHIA8 (black), $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8 (green), $t\bar{t}$+$b\bar{b}$ SHERPA (orange) and $t\bar{t}$+$b\bar{b}$ MG5AMC(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_R and μ_F are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, 3 b jets".

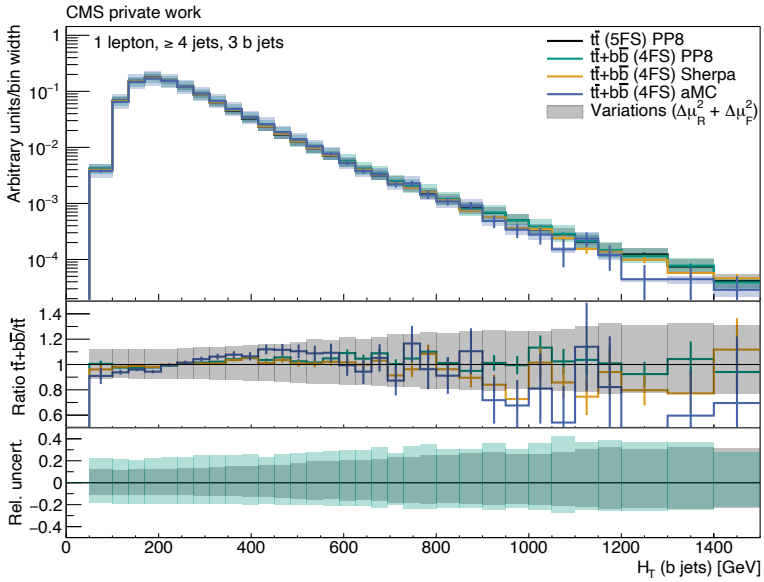


Figure A.9: H_{T} (b jets) of the $\mathrm{t\bar{t}}$ POWHEG+PYTHIA8 (black), $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ POWHEG+PYTHIA8 (green), $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ SHERPA (orange) and $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ MG5AMC(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_{R} and μ_{F} are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, 3 b jets".

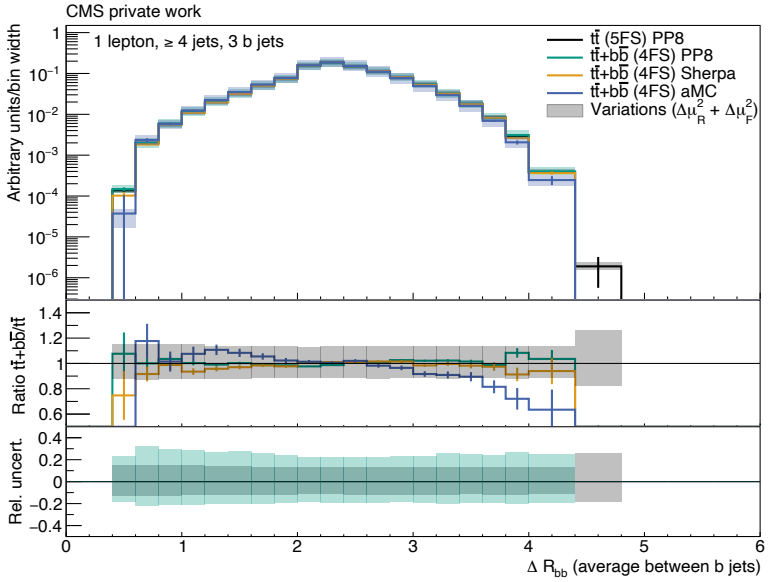


Figure A.10: $\Delta R\,(\mathrm{bb})$ (average) of the $\mathrm{t\bar{t}}$ POWHEG+PYTHIA8 (black), $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ POWHEG+PYTHIA8 (green), $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ SHERPA (orange) and $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ MG5AMC-(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_{R} and μ_{F} are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, 3 b jets".

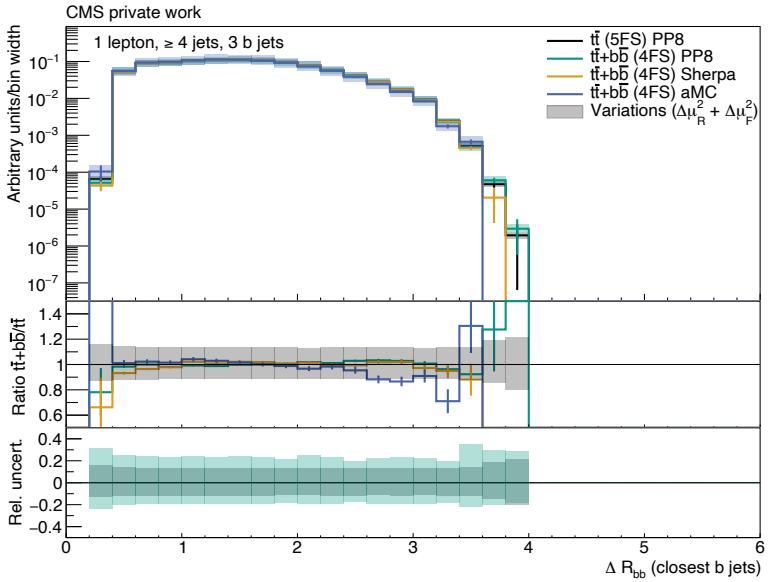


Figure A.11: ΔR(bb) (closest) of the $t\bar{t}$ Powheg+Pythia8 (black), $t\bar{t}$+$b\bar{b}$ Powheg+Pythia8 (green), $t\bar{t}$+$b\bar{b}$ Sherpa (orange) and $t\bar{t}$+$b\bar{b}$ MG5aMC-(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_R and μ_F are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, 3 b jets".

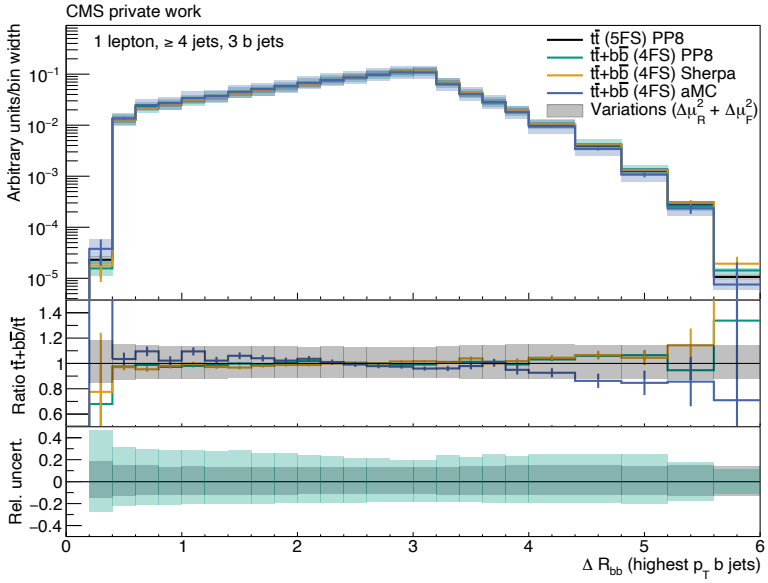


Figure A.12: ΔR(bb) (leading) of the $t\bar{t}$ Powheg+Pythia8 (black), $t\bar{t}$+$b\bar{b}$ Powheg+Pythia8 (green), $t\bar{t}$+$b\bar{b}$ Sherpa (orange) and $t\bar{t}$+$b\bar{b}$ MG5aMC-(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_R and μ_F are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, 3 b jets".

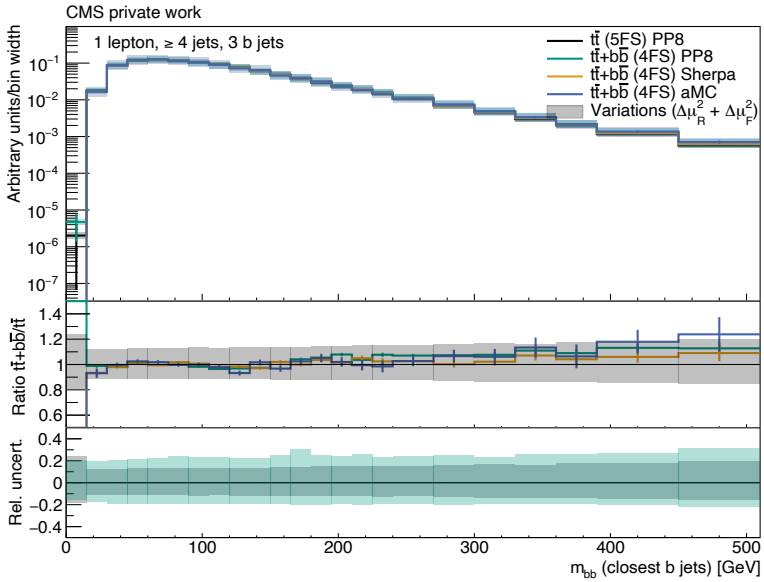


Figure A.13: $m\,(\mathrm{bb})$ (closest) of the $\mathrm{t\bar{t}}$ Powheg+Pythia8 (black), $\mathrm{t\bar{t}}+\mathrm{b\bar{b}}$ Powheg+Pythia8 (green), $\mathrm{t\bar{t}}+\mathrm{b\bar{b}}$ Sherpa (orange) and $\mathrm{t\bar{t}}+\mathrm{b\bar{b}}$ MG5aMC-(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_R and μ_F are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, 3 b jets".

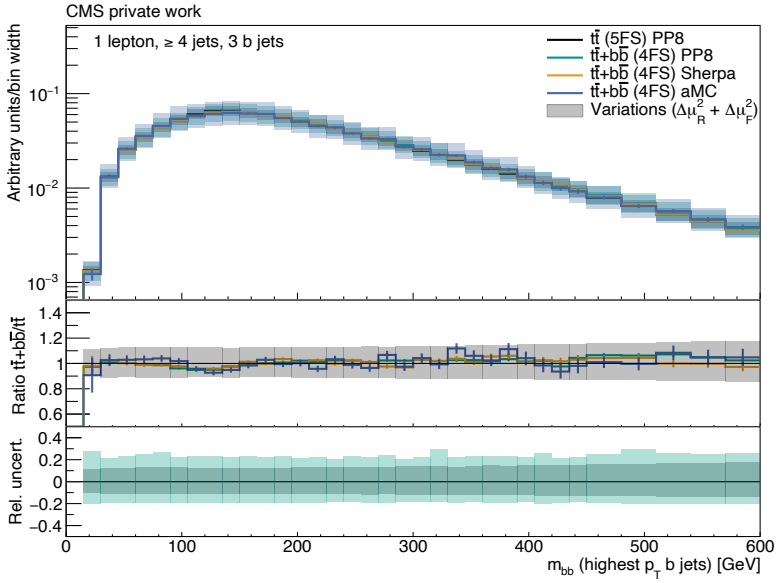


Figure A.14: $m\,(\mathrm{bb})$ (leading) of the $\mathrm{t\bar{t}}$ Powheg+Pythia8 (black), $\mathrm{t\bar{t}}+\mathrm{b\bar{b}}$ Powheg+Pythia8 (green), $\mathrm{t\bar{t}}+\mathrm{b\bar{b}}$ Sherpa (orange) and $\mathrm{t\bar{t}}+\mathrm{b\bar{b}}$ MG5aMC-(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_R and μ_F are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, 3 b jets".

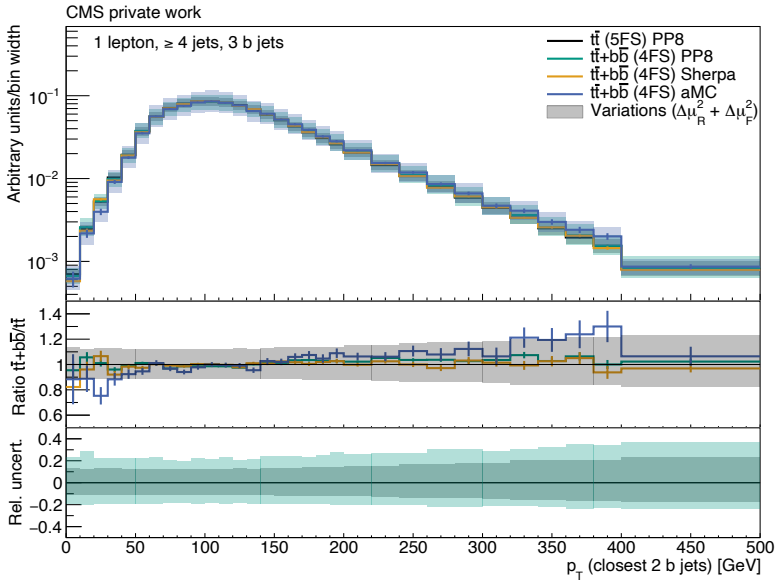


Figure A.15: p_{T} (bb) (closest) of the $\mathrm{t\bar{t}}$ Powheg+Pythia8 (black), $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ Powheg+Pythia8 (green), $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ Sherpa (orange) and $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ MG5aMC-(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_{R} and μ_{F} are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, 3 b jets".

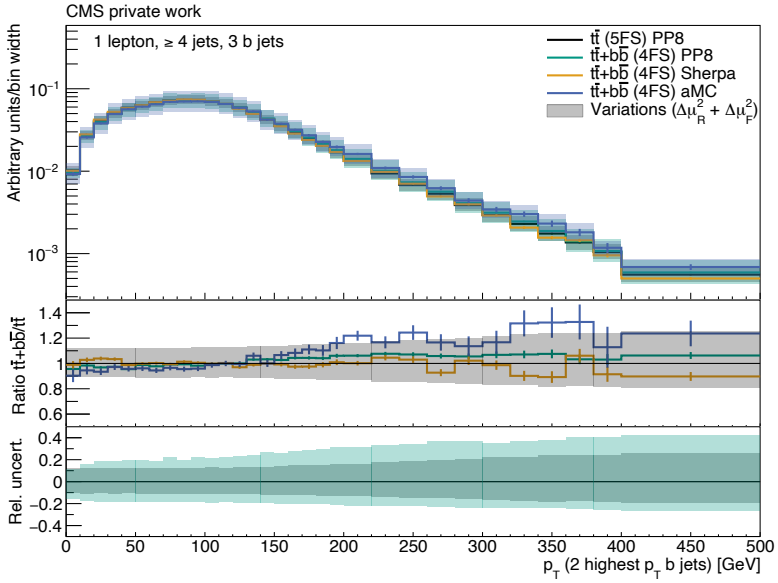


Figure A.16: p_{T} (bb) (leading) of the $\mathrm{t\bar{t}}$ Powheg+Pythia8 (black), $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ Powheg+Pythia8 (green), $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ Sherpa (orange) and $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ MG5aMC-(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_{R} and μ_{F} are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, 3 b jets".

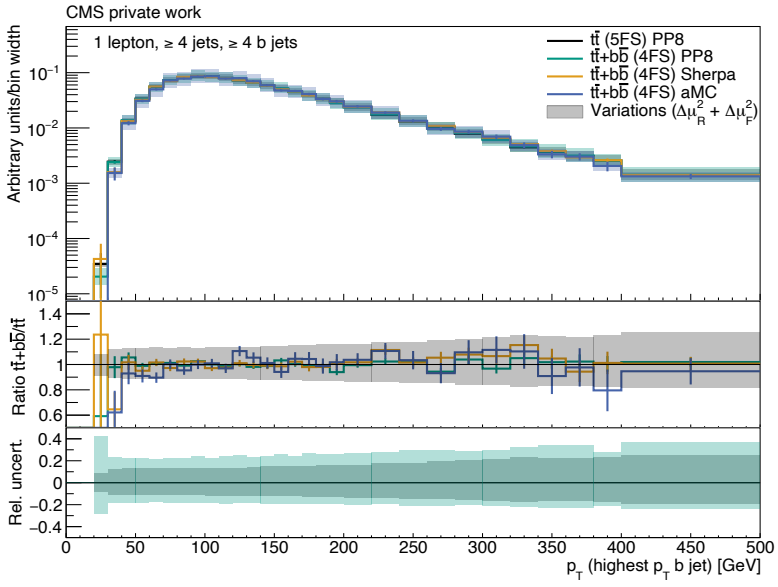


Figure A.17: Leading b jet p_{T} of the $\mathrm{t\bar{t}}$ Powheg+Pythia8 (black), $\mathrm{t\bar{t}+b\bar{b}}$ Powheg+Pythia8 (green), $\mathrm{t\bar{t}+b\bar{b}}$ Sherpa (orange) and $\mathrm{t\bar{t}+b\bar{b}}$ MG5aMC-(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_{R} and μ_{F} are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, ≥ 4 jets, ≥ 4 b jets".

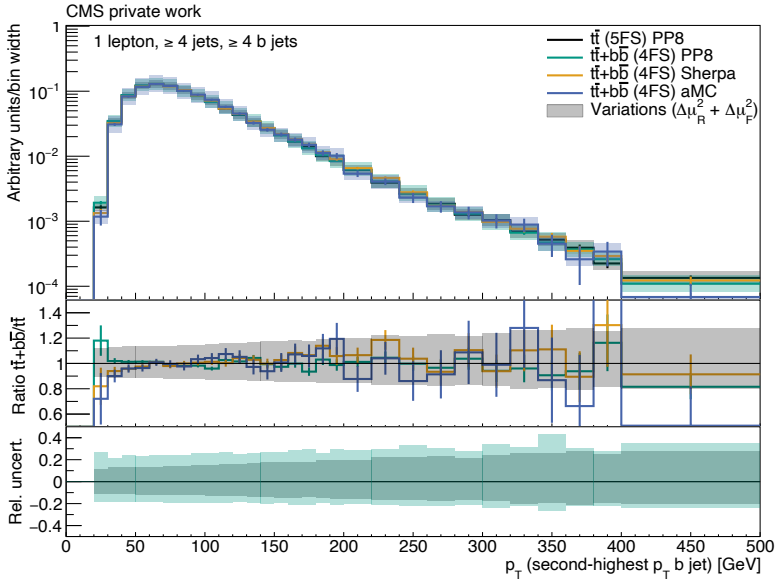


Figure A.18: Sub-leading b jet p_{T} of the $\mathrm{t\bar{t}}$ Powheg+Pythia8 (black), $\mathrm{t\bar{t}+b\bar{b}}$ Powheg+Pythia8 (green), $\mathrm{t\bar{t}+b\bar{b}}$ Sherpa (orange) and $\mathrm{t\bar{t}+b\bar{b}}$ MG5aMC-(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_{R} and μ_{F} are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, ≥ 4 jets, ≥ 4 b jets".

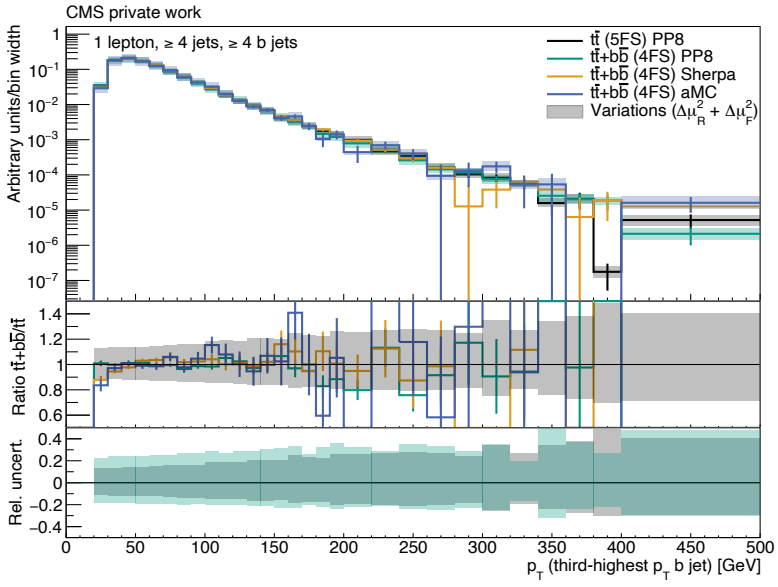


Figure A.19: Third b jet p_{T} of the $\mathrm{t\bar{t}}$ POWHEG+PYTHIA8 (black), $\mathrm{t\bar{t}+b\bar{b}}$ POWHEG+PYTHIA8 (green), $\mathrm{t\bar{t}+b\bar{b}}$ SHERPA (orange) and $\mathrm{t\bar{t}+b\bar{b}}$ MG5AMC-(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_{R} and μ_{F} are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, $\geq$ 4 b jets".

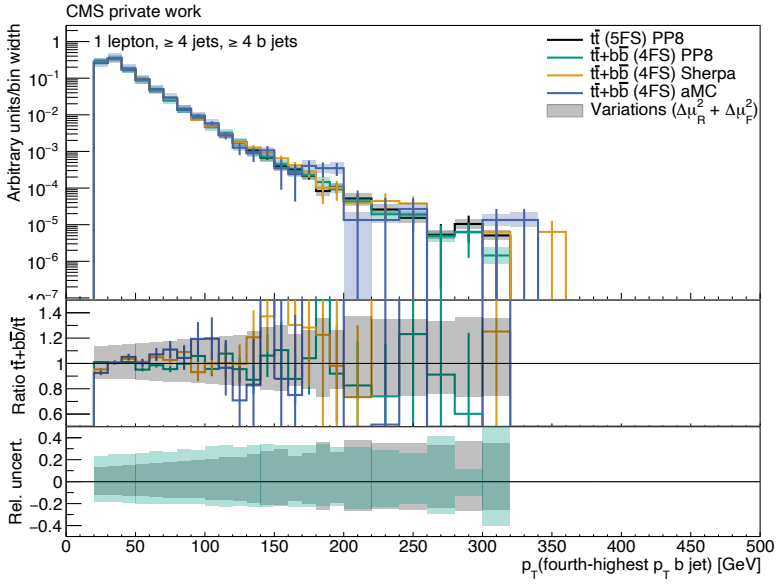


Figure A.20: Fourth b jet p_{T} of the $\mathrm{t\bar{t}}$ POWHEG+PYTHIA8 (black), $\mathrm{t\bar{t}+b\bar{b}}$ POWHEG+PYTHIA8 (green), $\mathrm{t\bar{t}+b\bar{b}}$ SHERPA (orange) and $\mathrm{t\bar{t}+b\bar{b}}$ MG5AMC-(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_{R} and μ_{F} are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, $\geq$ 4 b jets".

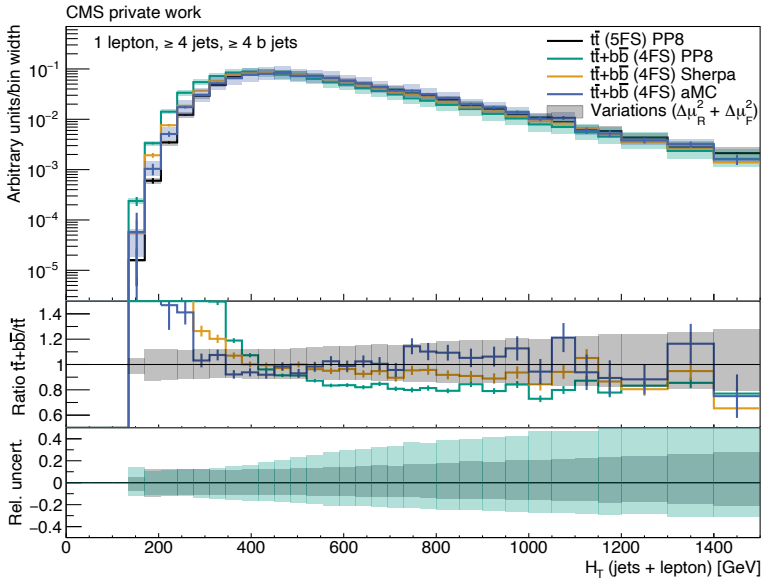


Figure A.21: H_T (jets+lepton) of the $\mathrm{t\bar{t}}$ POWHEG+PYTHIA8 (black), $\mathrm{t\bar{t}+b\bar{b}}$ POWHEG+PYTHIA8 (green), $\mathrm{t\bar{t}+b\bar{b}}$ SHERPA (orange) and $\mathrm{t\bar{t}+b\bar{b}}$ MG5AMC-(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_R and μ_F are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, $\geq$ 4 b jets".

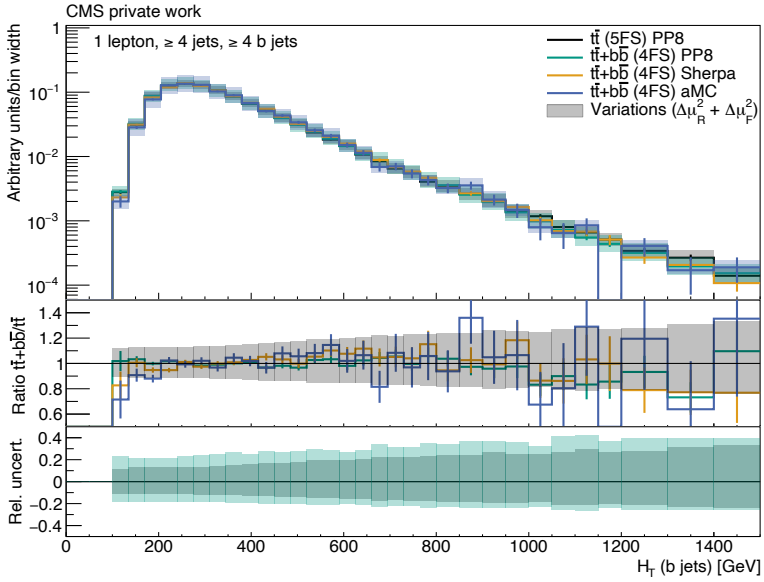


Figure A.22: H_T (b jets) of the $\mathrm{t\bar{t}}$ POWHEG+PYTHIA8 (black), $\mathrm{t\bar{t}+b\bar{b}}$ POWHEG+PYTHIA8 (green), $\mathrm{t\bar{t}+b\bar{b}}$ SHERPA (orange) and $\mathrm{t\bar{t}+b\bar{b}}$ MG5AMC(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_R and μ_F are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, $\geq$ 4 b jets".

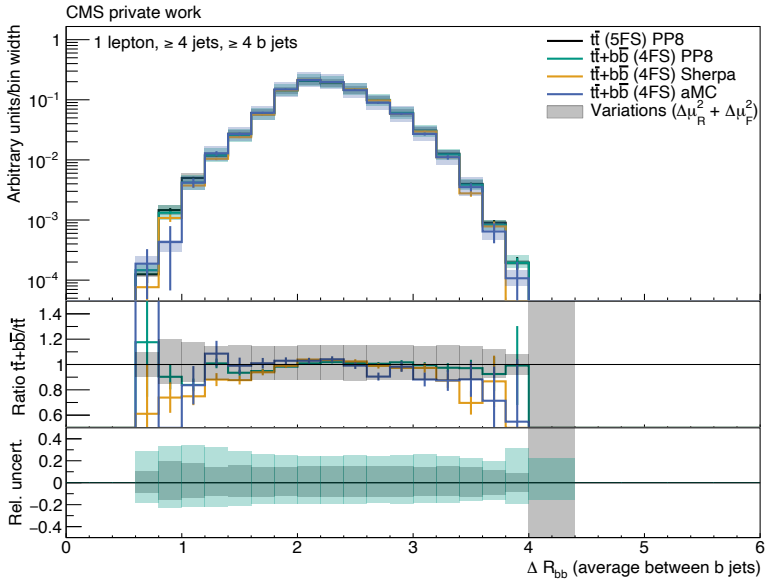


Figure A.23: ΔR(bb) (average) of the $t\bar{t}$ POWHEG+PYTHIA8 (black), $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8 (green), $t\bar{t}$+$b\bar{b}$ SHERPA (orange) and $t\bar{t}$+$b\bar{b}$ MG5AMC-(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_R and μ_F are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, $\geq$ 4 b jets".

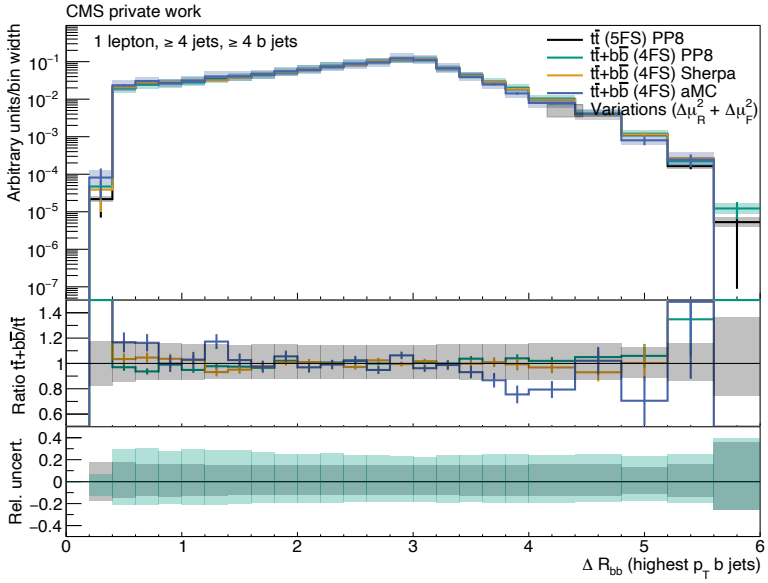


Figure A.24: ΔR(bb) (leading) of the $t\bar{t}$ POWHEG+PYTHIA8 (black), $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8 (green), $t\bar{t}$+$b\bar{b}$ SHERPA (orange) and $t\bar{t}$+$b\bar{b}$ MG5AMC-(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_R and μ_F are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, $\geq$ 4 b jets".

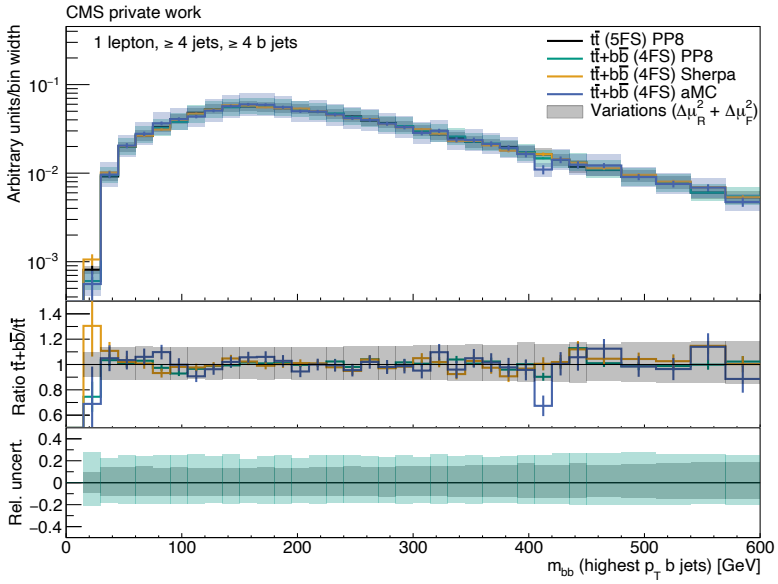


Figure A.25: m (bb) (leading) of the $t\bar{t}$ POWHEG+PYTHIA8 (black), $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8 (green), $t\bar{t}$+$b\bar{b}$ SHERPA (orange) and $t\bar{t}$+$b\bar{b}$ MG5AMC-(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_R and μ_F are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, $\geq$ 4 b jets".

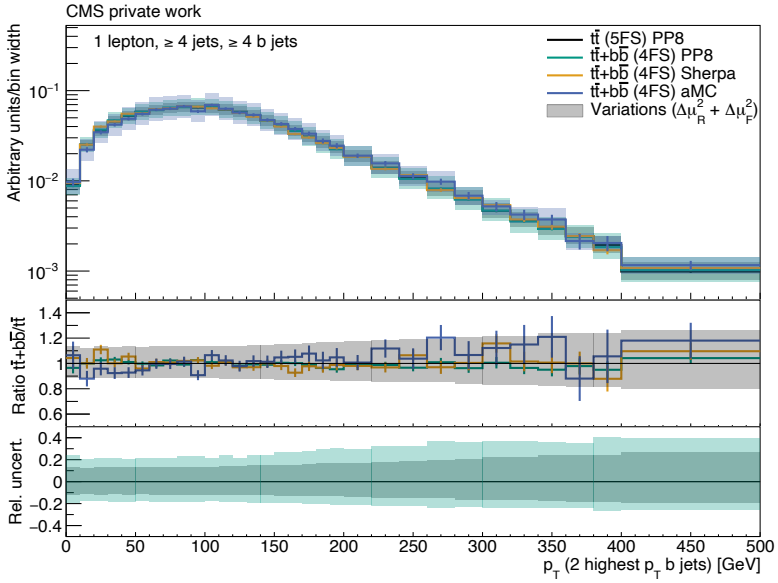


Figure A.26: p_T (bb) (leading) of the $t\bar{t}$ POWHEG+PYTHIA8 (black), $t\bar{t}$+$b\bar{b}$ POWHEG+PYTHIA8 (green), $t\bar{t}$+$b\bar{b}$ SHERPA (orange) and $t\bar{t}$+$b\bar{b}$ MG5AMC-(NLO) (blue) simulated events. The statistical uncertainties are shown as bars. The quadratically summed variations of μ_R and μ_F are visualized as bands. All distributions are normalized to an integral value of 1. All events pass the selection "1 lepton, $\geq$ 4 jets, $\geq$ 4 b jets".

B Input variables

The input variables for the DNN training in Chapter 7 are shown below. A list of all input variables is given in Table 7.4.

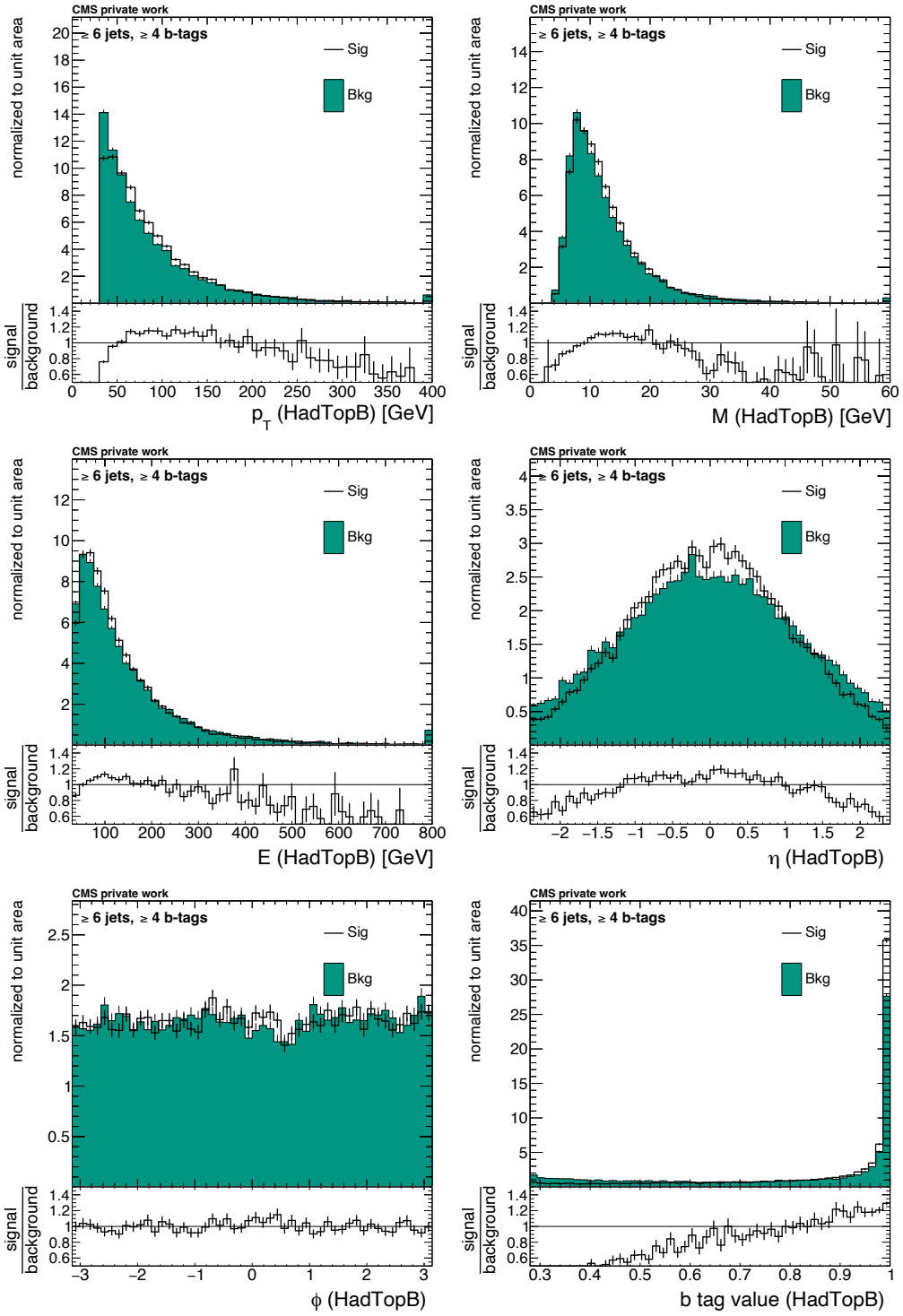


Figure B.27: Input variables of HadTopB for DNN training. The solid line (sig) shows the distribution of the given variable for the correct jet assignment. The histogram (bkg) shows a wrong jet assignment.

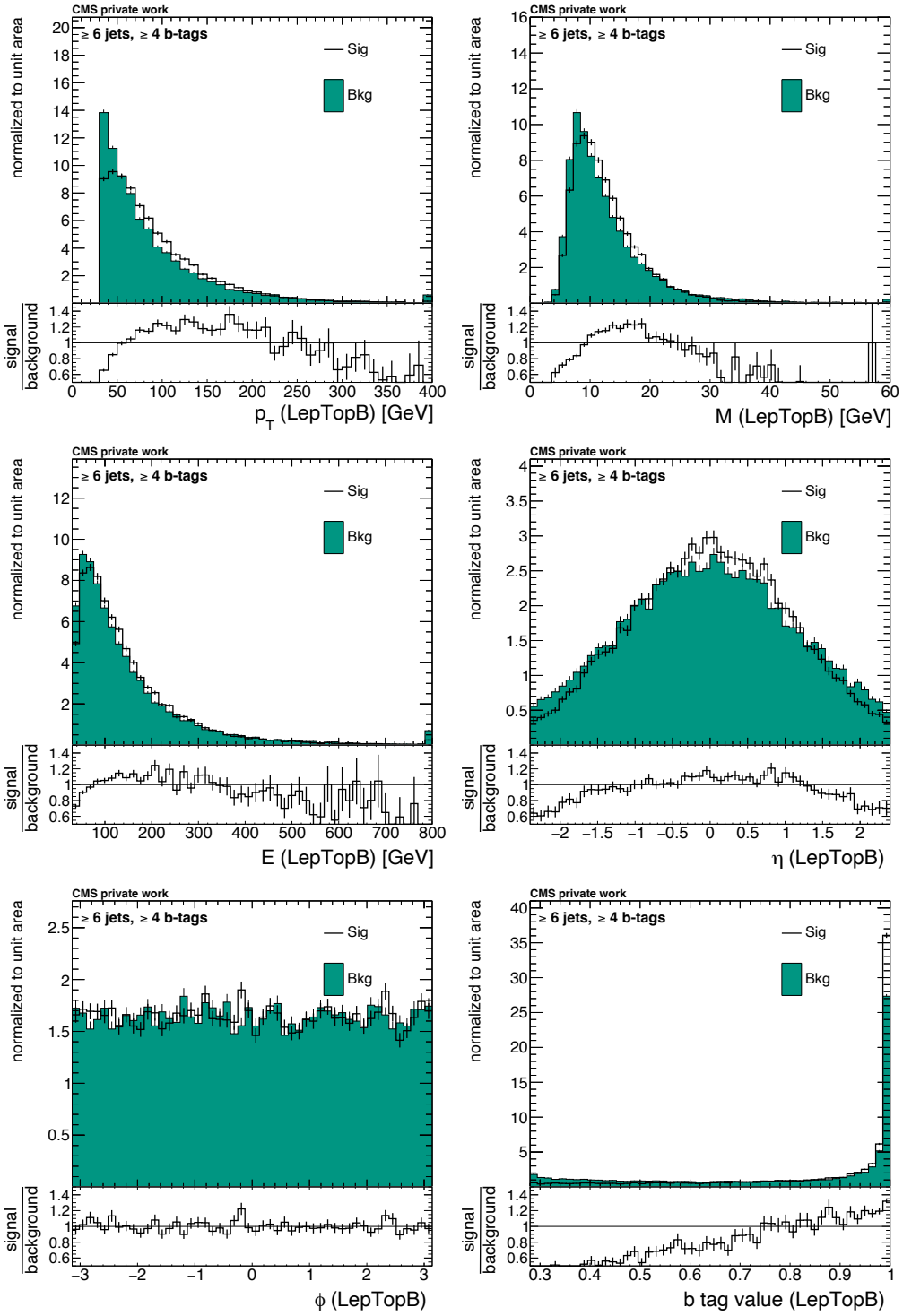


Figure B.28: Input variables of LepTopB for DNN training. The solid line (sig) shows the distribution of the given variable for the correct jet assignment. The histogram (bkg) shows a wrong jet assignment.

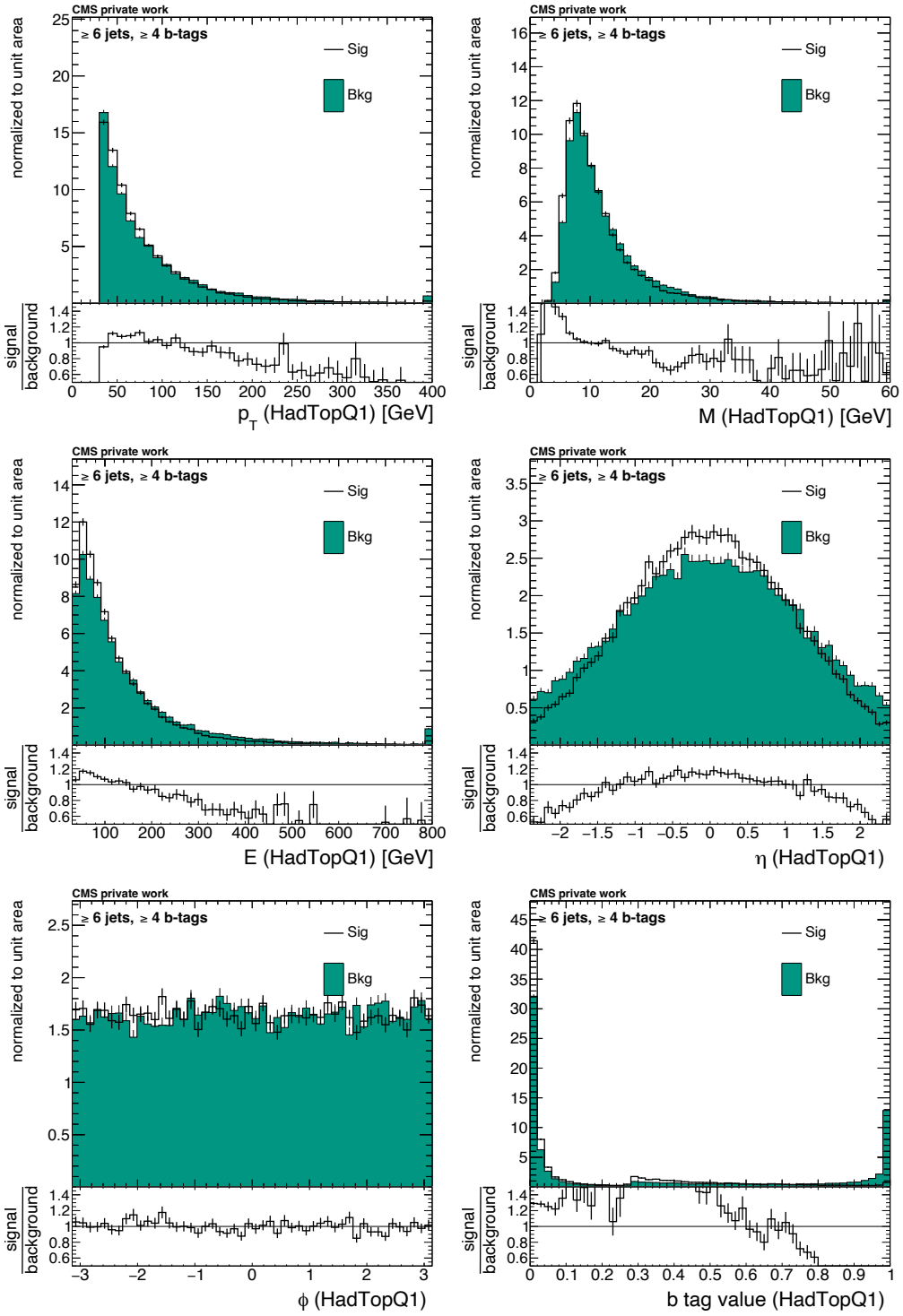


Figure B.29: Input variables of HadTopQ1 for DNN training. The solid line (sig) shows the distribution of the given variable for the correct jet assignment. The histogram (bkg) shows a wrong jet assignment.

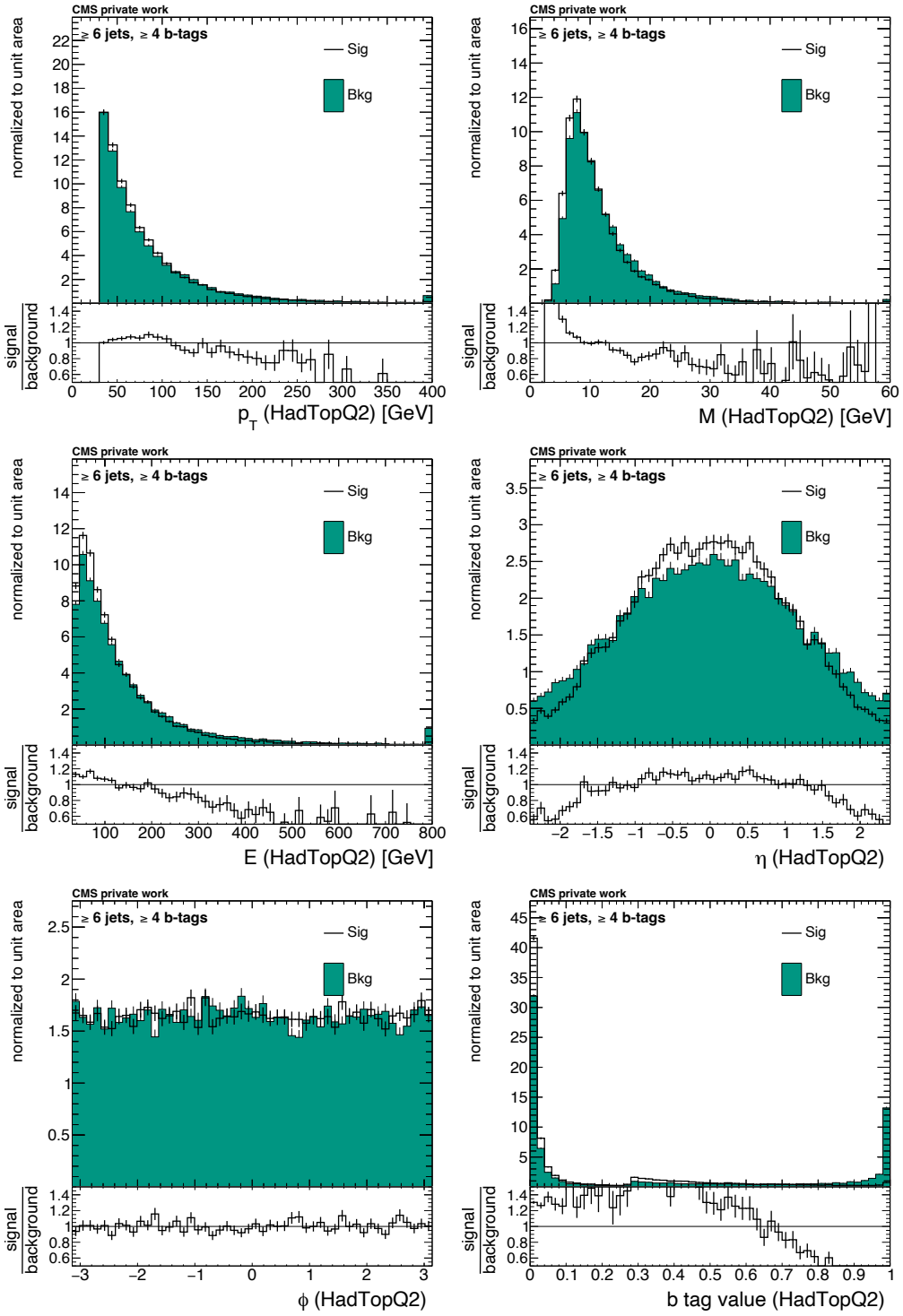


Figure B.30: Input variables of HadTopQ2 for DNN training. The solid line (sig) shows the distribution of the given variable for the correct jet assignment. The histogram (bkg) shows a wrong jet assignment.

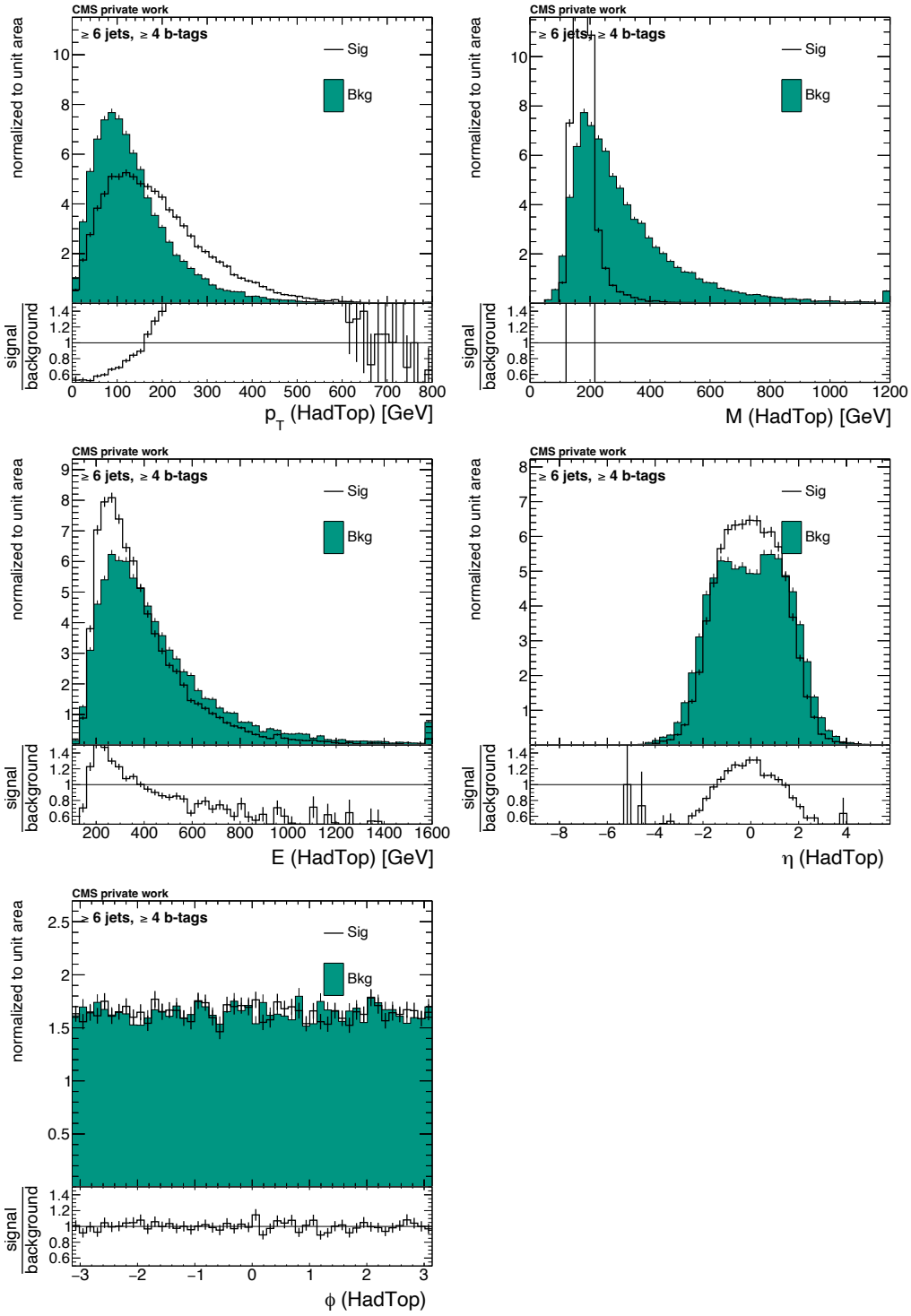


Figure B.31: Input variables of HadTop for DNN training. The solid line (sig) shows the distribution of the given variable for the correct jet assignment. The histogram (bkg) shows a wrong jet assignment.

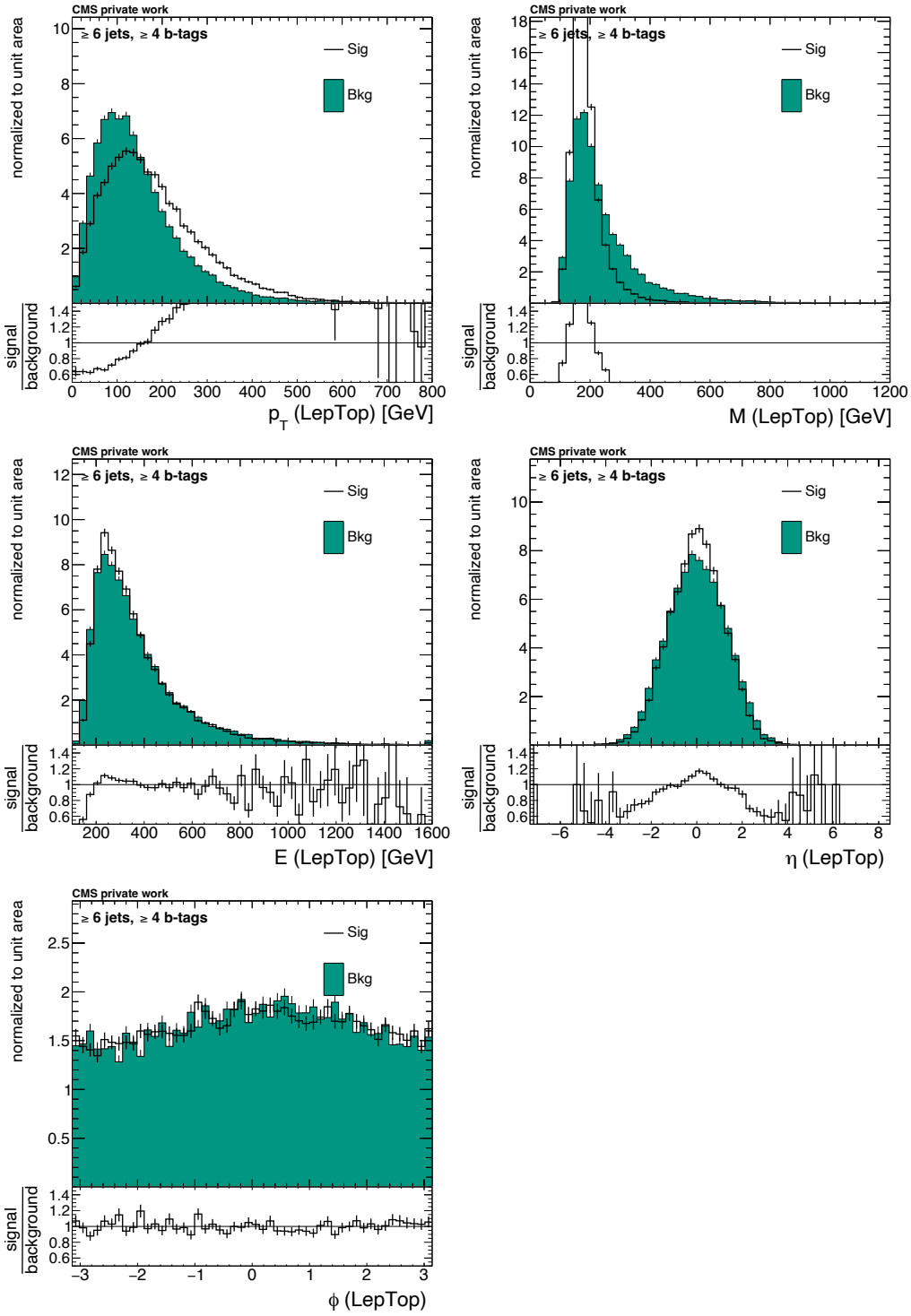


Figure B.32: Input variables of LepTop for DNN training. The solid line (sig) shows the distribution of the given variable for the correct jet assignment. The histogram (bkg) shows a wrong jet assignment.

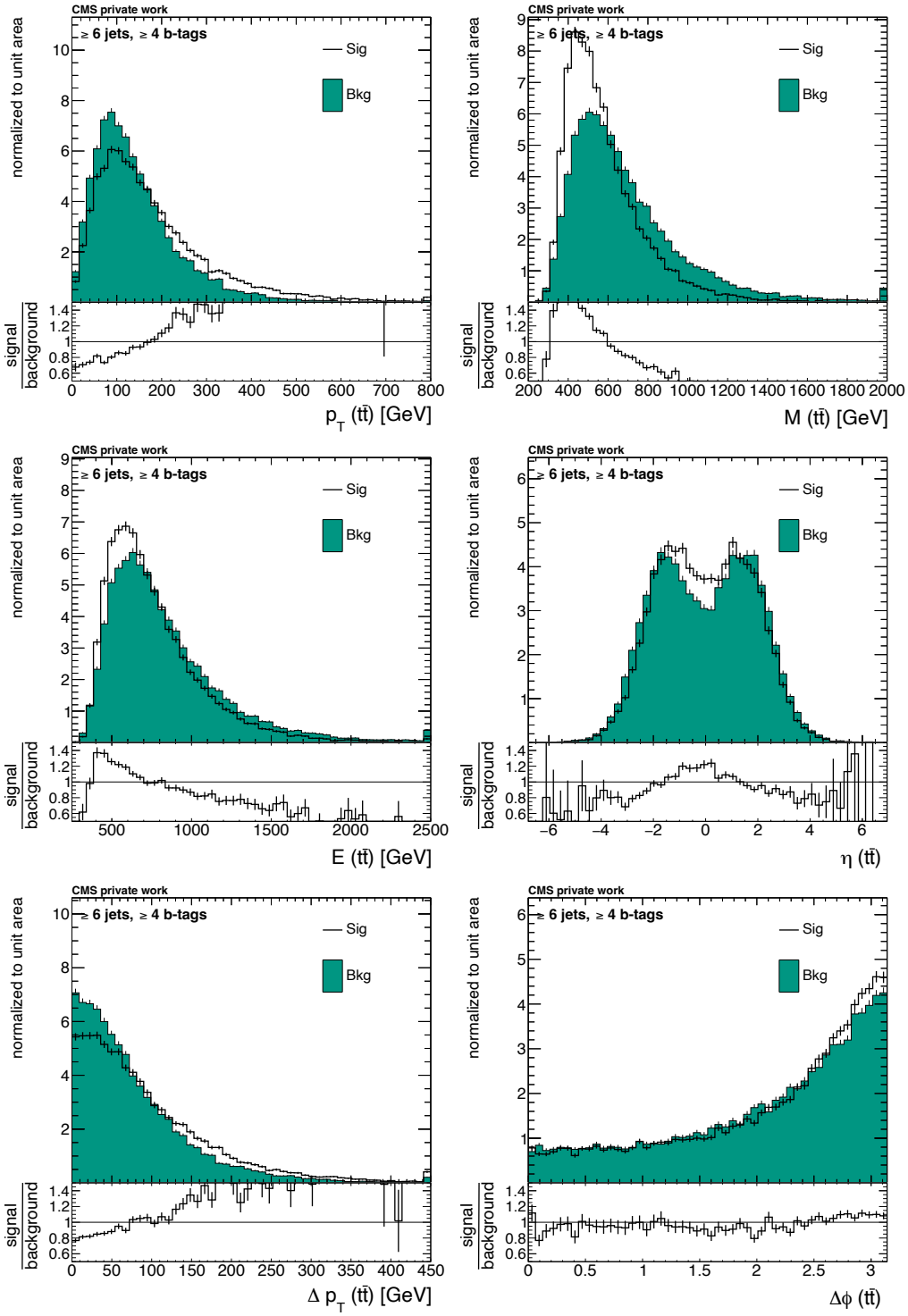


Figure B.33: Input variables of $t\bar{t}$ for DNN training. The solid line (sig) shows the distribution of the given variable for the correct jet assignment. The histogram (bkg) shows a wrong jet assignment.

List of Tables

List of Figures

Acknowledgments

First of all, I would like to thank Prof. Dr. Ulrich Husemann for the extensive support and individual mentoring during the master thesis. His inspiring enthusiasm shaped my entire path through the master's program from lectures to the thesis and set my focus on experimental particle physics.

Furthermore, I would particularly like to thank Prof. Dr. Thomas Müller for being my second reviewer and providing me with the opportunity to write my master thesis at the Institute of Experimental Particle Physics.

I am deeply grateful to Dr. Matthias Schröder and Jan van der Linden. Without their seamless and enormous support, their expertise and their meticulous comments, this thesis would not have been possible.

Special thanks also to all postdoctoral researchers and Ph.D. students who were always helpful with physics and technical questions, as well as looking at the work from a variety of angles: Dr. Karim El Morabit, Dr. Michael Waßmer, Philip Keicher, Sebastian Wieland and Nikita Shadskiy.

The comparison studies with the ATLAS experiment within the LHC Higgs Working Group would not have materialized without the intensive and collaborative work with Dr. Judith Katzy and Dr. Andrea Knue. My sincere thanks go to both of them.

To close, I am deeply grateful to my family for supporting my desire to study physics since day one. I appreciate the inspirational thoughts both inside and outside of physics by my friends.