Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

Data-driven selection of top quark pairs in multi-jet events at CMS

Master Thesis

by

David Spataro

(born: 07.07.1988)

angefertig im **Institut für Experimentalphysik**

vorgelegt der **Fakultät für Mathematik, Informatik und Naturwissenschaften**

1. Gutachter: Prof. Dr. Peter Schleper
2. Gutachter: Prof. Dr. Gregor Kasieczka

Hamburg, August 20, 2020

Spataro, David
Data-driven selection of top quark pairs
in multi-jet events at CMS
Master Thesis, Universität Hamburg, 2020.

Zusammenfassung

In dieser Thesis werden mittels eines neuronalen Netzes Top-Quark-Antiquark-Paare selektiert, die nach ihrem Zerfall in einem voll-hadronischen Endzustand aus sechs Jets im Detektor vorliegen. Die verwendeten Messdaten stammen vom CMS Experiment aus dem 2016 Run II des LHC mit einer Schwerpunktsenergie von $\sqrt{s}$ = 13 TeV. Es wird eine spezielle Trainingsmethode, Classification without Labels (CWoLa), verwendet und das Netz auf Daten trainiert, ohne die wahre Klassenzugehörigkeit der einzelnen Ereignisse zu kennen. Es wird gezeigt, dass eine Verbessung der Anzahl an richtig ausgewählten Ereignissen um mehr als das Zweifache möglich ist, bei gleichbleibenden Verhältnis von $t\bar{t}$-Paaren zu QCD basierten Multijet Ereignissen. Ein kurzer Ausblick des Einflusses auf die systematischen Fehler der Top-Quark-Massenmessung wird gegeben.

Abstract

This thesis uses a neural network to select top-quark-antiquark pairs, which after their decay into a fully hadronic final state are characterized by six jets in the detector. The measurement data used are from the 2016 Run II of the CMS experiment at the LHC with a center of mass energy of $\sqrt{s}$ = 13 TeV. A special training method, called Classification without Labels (CWoLa), is used and the network is trained on data without knowing the true class labels of the individual events. The number of selected events can be improved in comparison to previous selections by more than twofold with a constant ratio of $t\bar{t}$ pairs to QCD-based multijet events. A short outlook on the influence on the systematic errors of the top quark mass measurement is given.

Acknowledgements

I would like to thank Prof. Dr. Schleper for inviting me into his group and giving me the opportunity to work on such a fascinating project and Prof. Dr. Gregor Kasieczka for agreeing to be the second advisor for my thesis.

Dr. Hartmut Stadie was the supervisor of my thesis. I would like to thank him for his patience and all his effort. I had many questions regarding computation, statistics or physics and felt always welcomed to ask when appearing in front of his office.

I would also like to thank Dr. Johannes Lange, Mr. Christoph Garbers and Mr. Torben Lange who helped me to overcome technical problems, answered my various questions concerning physics, Linux and ROOT, what probably made room 110 my favourite one of the building.

I am very grateful to my soon-to-be wife, my family and friends for their support. I would never have accomplished this otherwise.

Contents

1. Introduction

The top quark plays a key role in understanding the Standard Model of Particle Physics. As the heaviest particle in the Standard Model its mass is the most important parameter. It also influences the Higgs boson mass through loop corrections. In previous analyses the most precise measurements are obtained through direct mass reconstruction, e.g. $t\bar{t} \to W^+bW^-\bar{b} \to (q\bar{q}'b)(q''\bar{q}'''\bar{b})$, a six jet decay topology [42]. The data from the 2016 Run II of the CMS experiment at the LHC is used for this thesis and consists of not only $t\bar{t}$ but also multi-jet events which fake this decay topology. Therefore it is necessary to differentiate between these cases in order to be able to carry out a top quark mass extraction. Systematic effects are limiting the precision of the top quark mass measurement, such as event selection efficiencies, which are obtained from simulation.

Here for the first time a fully data driven machine learning approach is used for event selection following classification without labels (CWoLa [30]) to improve mass measurements. This is a special training method, which can be instrumented for classification problems if the true class labels of the used data are not known, but a separation into samples with different signal fractions is possible. Here this is realised by using b-tagging information (1 or 2 b-tags) for creating unbalanced signal fractions.

First a neural network configuration is presented and trained in full supervision mode on MC simulated events. The result is compared to a previous selection. CWoLa is tested on simulated events to study advantages and disadvantages of the method itself. Thereafter the fully data driven approach is run. Eventually the influence on a background estimation is investigated and an outlook on the impact on systematic uncertainties is given.

2. Theoretical overview

2.1. The Standard Model

The Standard Model of Particle Physics (SM) is a theory of elementary particles and their interaction through three of the four fundamental forces (not including gravitational force): *weak interaction*, *electromagnetic interaction* and *strong interaction*.

The elementary particles are assigned as fermions (spin = $\frac{1}{2}$), gauge vector bosons (spin = 1) or as the scalar Higgs boson (spin = 0). The fermions are divided into six quarks and leptons, which are grouped into three generations, sometimes referred to as "flavour":

- *up-type quarks* ($Q = +\frac{2}{3}$): up (u), charm (c), top (t)
- *down-type quarks* ($Q = -\frac{1}{3}$): down (d), strange (s), bottom (b)
- *leptons* ($Q = -1$): electron (e), myon (μ), tau (τ)
- *neutrinos* ($Q = 0$): electron- (ν_e), myon- (ν_μ), tau- (ν_τ) neutrino

with Q as the electric charge. A graphical overview is given in fig. 1. Each of the quarks carries a colour charge, namely *red* (r), *green* (g), *blue* (b), whereas the corresponding antiquarks have the anticolours *antired* ($\bar{r}$), *antigreen* ($\bar{g}$), *antiblue* ($\bar{b}$). While leptons only underlay electromagnetic and weak interaction, quarks take part in strong interaction. Quarks are not appearing in isolation and can therefore not be observed directly. They only appear in multi-quark states bound by the strong force and are known as hadrons. This phenomenon is called confinement [35, 17].

The Higgs mechanism is giving mass to other particles and the gauge vector bosons mediate the forces between them:

- 8 vector gluons: mediating the strong interaction between colour charged particles
- 4 vector bosons: Z^0, $W^\pm$ for particles with different flavour and photons(γ) for charged particles, together mediating unified electroweak interactions
- 1 scalar boson : the Higgs boson, introduced by the Higgs mechanism, which is giving mass to all other particles via interacting with the Higgs field

All observed matter in the universe is made of these tiny building blocks. For example, protons are bound states of three quarks, consisting of two up-quarks and a down-quark: p(uud). Neutrons are composited of two down-quarks and one up-quark: n(ddu) [31].

Although the SM is one of the most outstanding theories of particle physics, many unanswered questions remain. Matching the few theoretical ingredients, such as the *Dirac equation*, the *Quantum Field Theory*, *local gauge principles* and the *Higgs mechanism*, experimental tests have verified the reliability on the SM up to energies at the electroweak scale. The SM also needs 26 parameters to fit the measurements [17, 35], which are grouped into:

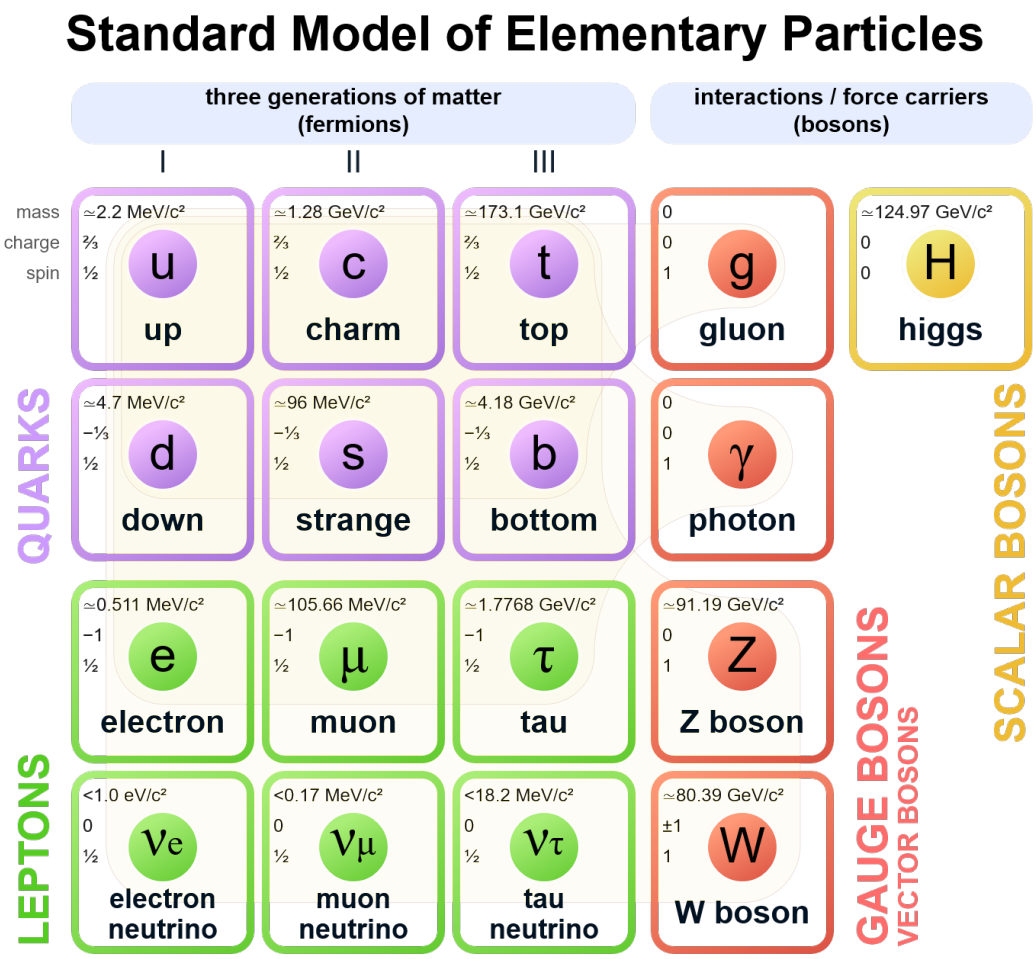


Figure 1: Particle content of the Standard Model [50].

- 12 Yukawa couplings of the Higgs field: m_{ν_1}, m_{ν_2}, m_{ν_3}, m_e, m_μ, m_τ, m_d, m_s, m_b, m_u, m_c, m_t
- 3 coupling constants: g, g_s and g'
- 2 parameters describing the Higgs potential: μ^2, λ
- 8 mixing angles of the PMNS[1] and CKM[2] matrices: θ_1, θ_2, θ_3, δ, γ, A, ρ, η
- 1 strong CP phase from the Lagrangian of QCD: $\theta_{CP} \simeq 0$

The construction of the SM Lagrangian can be summed up by defining guiding principals along with considering the SM being built within the local Quantum field theory [31]:

- a *generalized correspondence* to existing theories, in particular Quantum Mechanics, the Fermi model, QED etc.
- *minimality*, avoiding unnecessary fields, objects or interactions

[1] Pontecorvo–Maki–Nakagawa–Sakata matrix
[2] Cabibbo-Kobayashi-Maskawa-Matrix

- *unitarity*, e.g a condition for cross sections, transformations of fields and probabilities being limited from above by unity
- *renormalizability*: needed at quantum level for derivation of finite predictions for observable quantities
- *gauge principle* for introduction of interactions
- *symmetry*, the main guiding principle. The SM posseses various kinds of symmetries such as CPT, Lorentz symmetry and the already mentioned gauge symmetries $SU(3)_C \times SU(2)_L \times U(1)_Y$

These principles can be described with the Lagrange formalism, where the structures of the interactions of the fields are written as Lagrangian density $\mathscr{L}$ and the dynamics by the Euler-Lagrange equations [35]:

$$\frac{\partial \mathscr{L}}{\partial \varphi} = \partial_\mu \left(\frac{\partial \mathscr{L}}{\partial(\partial_\mu \varphi)} \right) \tag{1}$$

For the gauge fields of $SU(3)_C$, $SU(2)_L$ and $U(1)_Y$ the Lagrangian term is denoted as follows:

$$\mathscr{L}_g = -\frac{1}{4} F^{b\mu\nu} F^b_{\mu\nu} - \frac{1}{4} W^{a\mu\nu} W^a_{\mu\nu} - \frac{1}{4} B^{\mu\nu} B_{\mu\nu} \tag{2}$$

with b = 1,...,8 for $SU(3)_C$ and a = 1,..,3 for $SU(2)_L$. The next term adds all fermion fields. Left handed fermions are grouped into $SU(2)$ doublets and right handed fermions into $SU(2)$ singulets:

$$L_L = \begin{pmatrix} \nu_e \\ e \end{pmatrix}, \begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix}, \begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}, l_r = e_R, \mu_R, \tau_R \tag{3}$$

$$Q^c_L = \begin{pmatrix} u^c \\ d'^c \end{pmatrix}, \begin{pmatrix} c^c \\ s'^c \end{pmatrix}, \begin{pmatrix} t^c \\ b'^c \end{pmatrix}, q^c_R = u^c_R, d'^c, c^c, s'^c, t^c, b'^c \tag{4}$$

with

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = V_{CKM} \begin{pmatrix} d \\ s \\ b \end{pmatrix} \tag{5}$$

This leads to:

$$\begin{aligned} \mathscr{L}_f = & \sum_{L=e,\mu,\tau} \bar{L}_L i\gamma^\mu (\partial_\mu - i\frac{g'}{2} Y B_\mu - igT^a W^a) L_L \\ & + \sum_{l=e,\mu,\tau} \bar{l}_R i\gamma^\mu (\partial_\mu - i\frac{g'}{2} Y B_\mu) l_R \\ & + \sum_{Q=u,c,t} \sum_{c,c'} \bar{Q}^c_L i\gamma^\mu (\partial_\mu - i\frac{g'}{2} Y B_\mu - igT^a W^a_\mu - ig_s T^b_{s_{c,c'}} G^b_\mu) Q^{c'}_L \\ & + \sum_{q=u,d,c,s,t,b} \bar{q}^c_R i\gamma^\mu (\partial_\mu - i\frac{g'}{2} Y B_\mu - ig_s T^a_{s_{cc'}} G^a) q^{c'}_R . \end{aligned} \tag{6}$$

with $T_s^a = \frac{\gamma^a}{2}$, a = 1,...,8 and $[T^a, T_s^b] = [T_s^a, \mathrm{Y}] = 0$. For the Higgs mechanism the scalar Lagrange density is:

$$\mathscr{L}_H = (D_\mu \phi)^\dagger D^\mu \phi - \mu^2 \phi^\dagger \phi - \lambda (\phi^\dagger \phi)^2 \tag{7}$$

$$D_\mu \phi = (\partial_\mu - i\frac{g'}{2} Y B_\mu - igT^a W_\mu^a)\phi \tag{8}$$

From the Yukawa terms the fermion masses are received:

$$\mathscr{L}_Y = -\sum_{l=\mu,e,\tau} G_L \bar{L}_L \phi l_R - \sum_{q=u,c,t} G_q \bar{Q}_L \tilde{\phi} q_R - \sum_{p=d,s,b} G_p \bar{Q}_L \phi p'_R + h.k. \tag{9}$$

The complete Lagrangian density is the sum of the four densities above:

$$\mathscr{L}_{SM} = \mathscr{L}_g + \mathscr{L}_f + \mathscr{L}_H + \mathscr{L}_Y \tag{10}$$

More details about the derivation are given in [35]. A short form of this equation (and therefore slightly changed notation) is possible in the following way [56]:

$$\begin{aligned} \mathscr{L}_{SM} = & -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} \\ & + i\bar{\psi} \not{D} \psi + h.c \\ & + \psi_i y_{ij} \psi_j \phi + h.c \\ & + |D_\mu \phi|^2 - V(\phi). \end{aligned} \tag{11}$$

The first line describes the interaction particles, the second one takes care of the interactions between matter particles, while the third line deals with the mass for matter particles and antiparticles (h.c). Eventually the last line characterizes the mass for interaction particles and the Higgs self-interactions. More detailed information about the derivation can be found in [31, 17].

There are many extension of the SM, for example, supersymmetry (SUSY), large-scale extra dimensions or string theory, which are trying to cover the topics the SM is not able to explain, for example Dark Matter, Matter-Anti-Matter Asymmetry, gravity or neutrino masses.

2.2. The top quark

In the SM, the top quark is the heaviest particle with a mass of about m_t=172.26±0.07 (stat+JSF) ±0.61 (syst) GeV according to measurements at the LHC [39]. The measurement was performed with a integrated luminosity of 35.9 fb^{-1} and uses the $t\bar{t}$ all-jets final state in combination with the lepton+jets channel. The data were collected with the CMS detector from proton-proton collisions at $\sqrt{s}$ = 13 TeV. Due to its high mass it is the only quark decaying into a real W boson and a b quark. Top quarks decay before they hadronize because of their short lifetime of around $5 \cdot 10^{-25}$s. This is shorter than the timescale on which strong interaction of QCD has an effect [11]. This timescale is estimated as $\tau_{had} \approx \frac{1}{\Lambda_{QCD}} = 3 \cdot 10^{-24}$ [33].

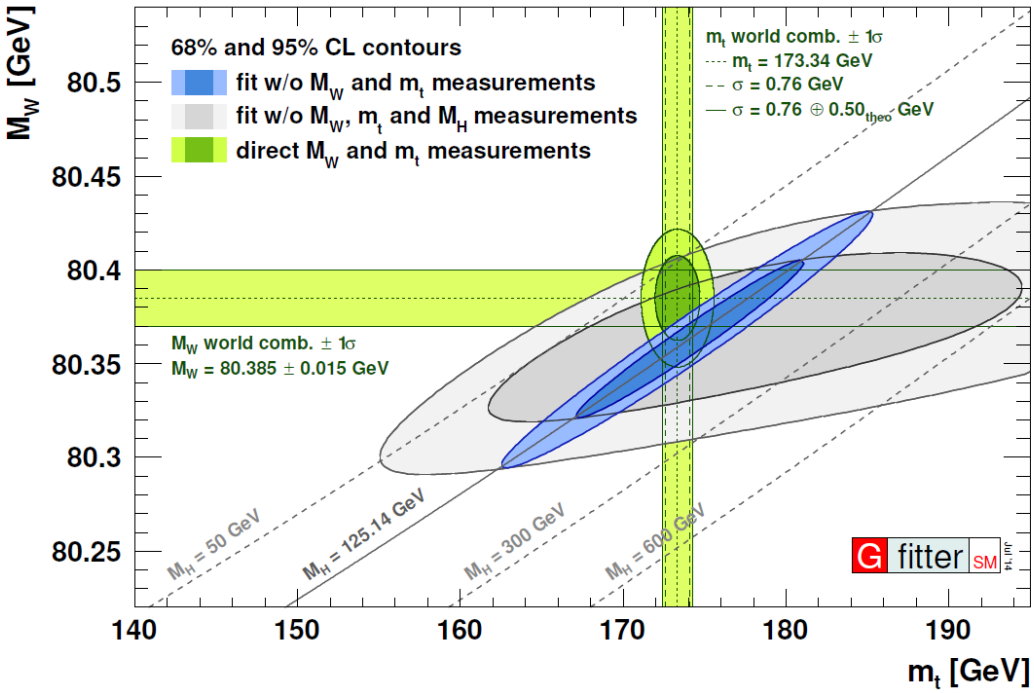


Figure 2: 68% and 95% confidence level contours for electroweak fits including and excluding the Higgs. These contours are obtained from scanning m_t vs M_W, adding a theoretical uncertainty of 0.5 GeV for the direct top mass measurement [19].

The top quark is the only quark with a Yukawa coupling to the Higgs boson near unity. Therefore it contributes strongly to loop corrections of the Higgs boson, which depends on the mass of other particles of the SM. The Higgs boson mass m_H has a quadratic dependence on the top quark mass m_t and a logarithmic one on the mass of the W boson M_W. These two masses were used to constrain the Higgs mass before it was discovered [33]. An example of this constraint is given in fig. 2 from the Gfitter group. These properties make the top quark playing a special role in the SM. A better understanding of its properties and a more accurate measurement of its mass will lead to improved information on fundamental interactions at the electroweak symmetry-breaking scale and beyond SM physics [37].

2.3. Production and decay of top quark pairs

At LHC the most common productions of top quarks happens via gluon-gluon fusion or quark-antiquark annihilation. In this process, mediated through the strong interaction, top quarks are produced along with their antipartner. This is shown at leading order in fig. 4. The gluon-gluon fusion process is dominant with a ratio of about 90% at $\sqrt{s}$ = 13 TeV. The data used for this thesis consist of about 30 million $t\bar{t}$ pairs produced at each collision point. Creating single top quarks is also possible, but this process has a production rate significantly smaller than the one of $t\bar{t}$ [33].

A summary of the total production cross-section for different center-of-mass energies can be found in [40]. For different masses the formula from [31] is used and parameters are determined

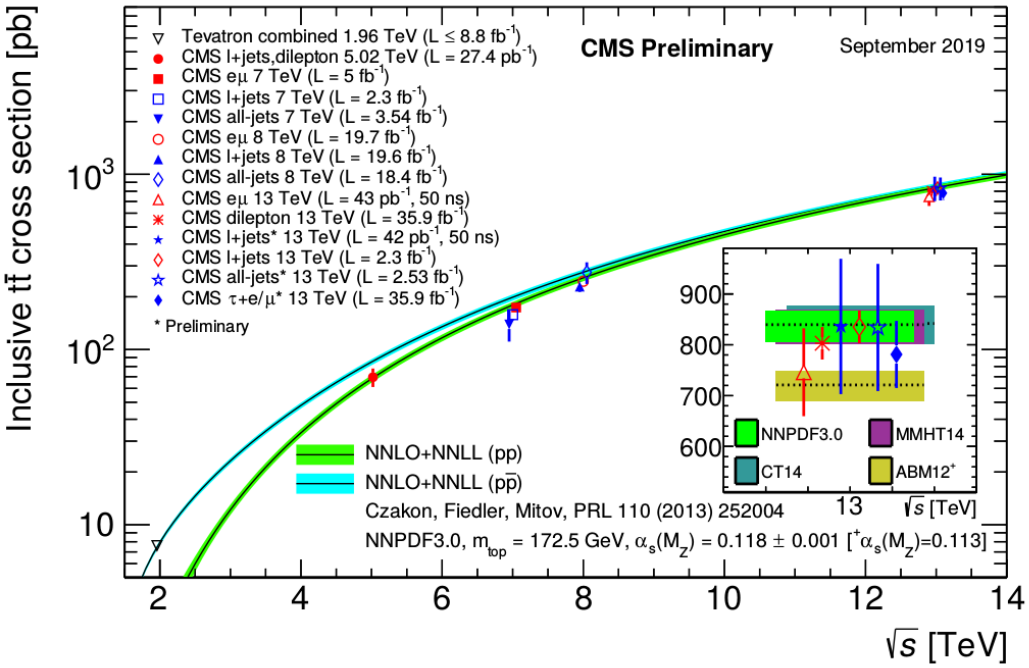


Figure 3: Top quark pair cross section summary of CMS measurements in comparison to theoretical calculations at NNLO+NNLL accuracy. The Tevatron measurements are also shown [41].

by CMS and ATLAS. The reference mass is set as m_{ref} = 172.5 GeV:

$$\sigma(m_t) = \left(\frac{m_{ref}}{m_t}\right)^4 \sigma(m_{ref}) \left[1 + a_1 \left(\frac{m_t - m_{ref}}{m_{ref}}\right) + a_2 \left(\frac{m_t - m_{ref}}{m_{ref}}\right)^2\right] \tag{12}$$

For $t\bar{t}$ events at $\sqrt{s}$ = 13 TeV the cross section is 831.76pb [37]. A summary of the measurements of the total $t\bar{t}$ cross section measured by the CMS collaboration is shown in fig. 3.

The top quark decays according to the SM almost exclusively to a b quark and a W boson. Hence the final state topology depends on whether the W bosons decay into quarks or leptons and neutrinos. There are three options [37]:

- **All-jets**: $t\bar{t} \to W^+bW^-\bar{b} \to (q\bar{q}'b)(q''\bar{q}'''\bar{b})$ (45.7%)
- **Lepton+jets**: $t\bar{t} \to W^+bW^-\bar{b} \to (q\bar{q}'b)(l\nu_l\bar{b})$ (43.8%)
- **Dilepton**: $t\bar{t} \to W^+bW^-\bar{b} \to (\bar{l}\nu_l\bar{b})(l'\nu'_l\bar{b})$ (10.5%)

In this thesis only data of the all-jets decay are used. Jets are narrow cones of hadrons and other particles. They are produced by hadronization of quarks and gluons. Obeying confinement, several particles carrying colour charges form colourless objects. These colourless objects then form observable jet structures because the colour charges carrying fragments tend to travel in the same direction. The signature of all-jet events are six jets, two originating from two b-quarks, the four remaining jets result from the quarks coming from the decays of the W bosons. This event topology is displayed in fig. 5. The appearance of additional jets due to gluon radiation is possible [8, 1].

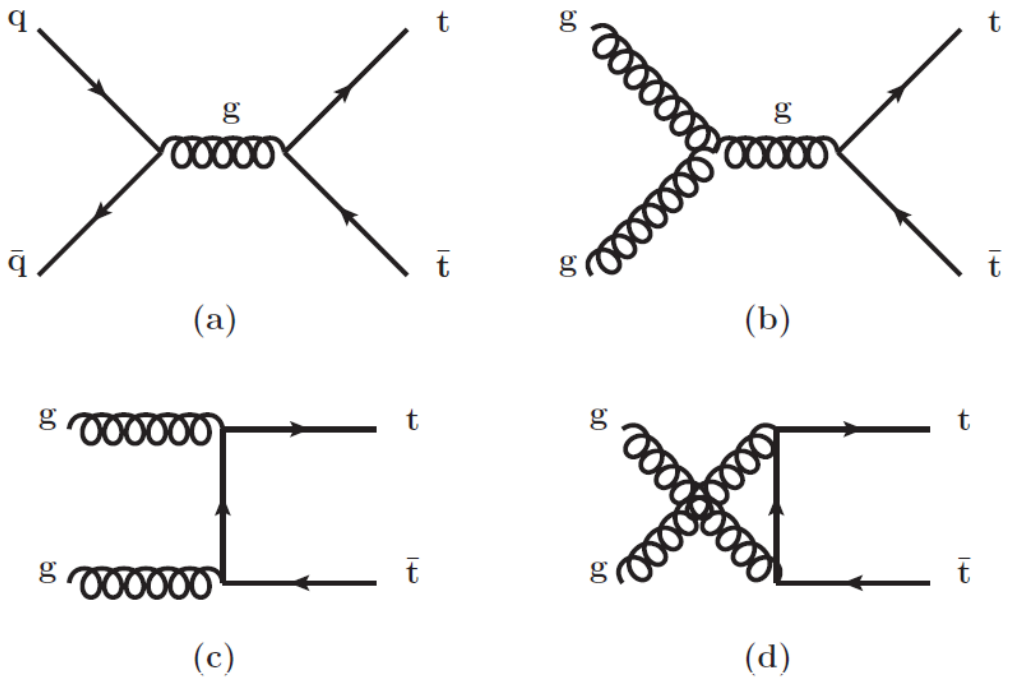


Figure 4: Feynman diagram of the $t\bar{t}$ creation process via $q\bar{q}$ annihilation (a) and g-g fusion in s-channel (b), t-channel (c) and u-channel (d) [33].

A challenging task is to filter these all-jets events from the underlying QCD-multijet events. QCD-multijet events often have a similar signature in the detector could be mistreated as $t\bar{t}$-all-jets events. The particle-flow algorithm [20] combines the information of all subdetectors in the CMS detector at LHC and reconstructs each particle individually. Combining following processes with additional quarks or gluon fusion might lead to detect a similar topology to the $t\bar{t}$ decay [8]:

- $gg \rightarrow gg$
- $gg \rightarrow q\bar{q}$
- $qg \rightarrow qg$
- $qq \rightarrow qq$
- $q\bar{q} \rightarrow q\bar{q}$
- $q\bar{q} \rightarrow gg$

A graphical representation of QCD-multijets events is given in fig. 6.

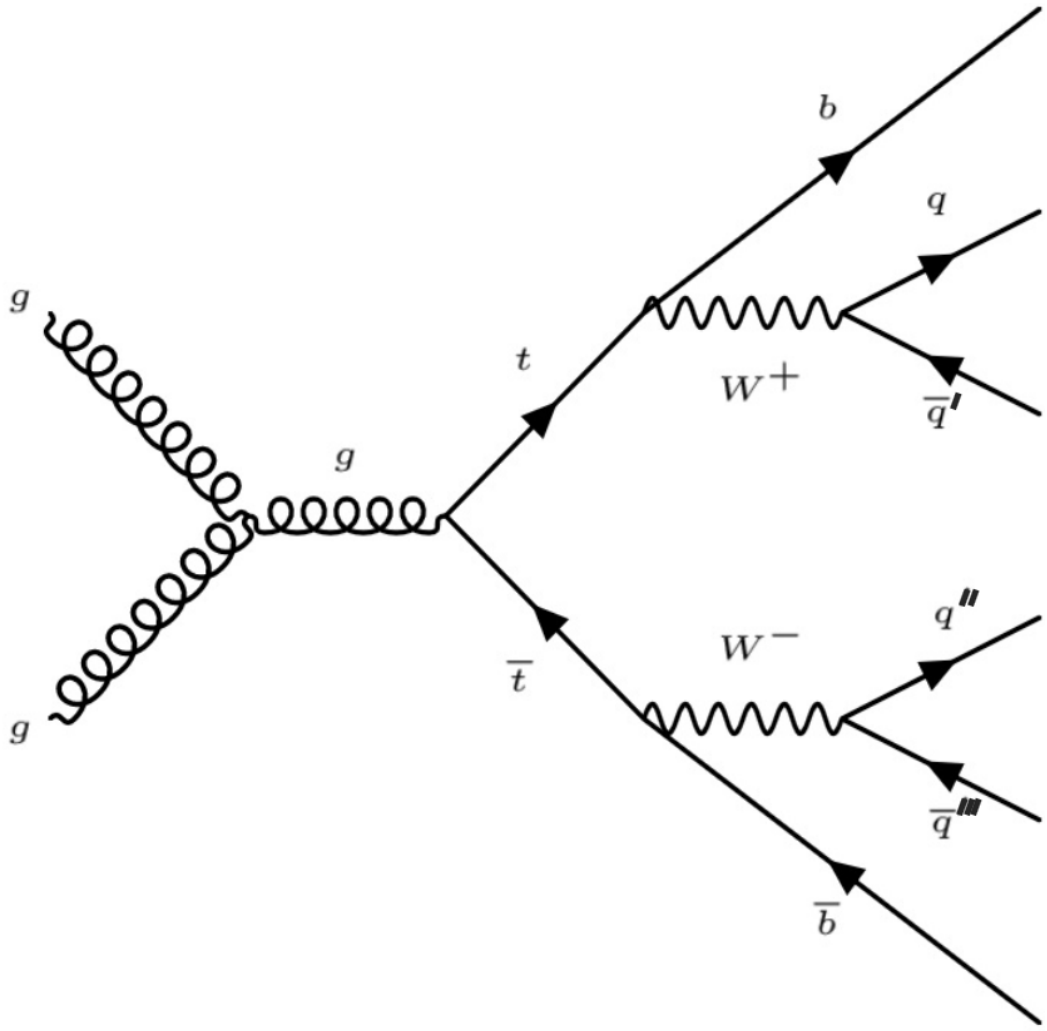


Figure 5: Full hadronic decay of a $t\bar{t}$ pair into two W bosons and two b quarks. The W bosons then decay to quarks themselves.

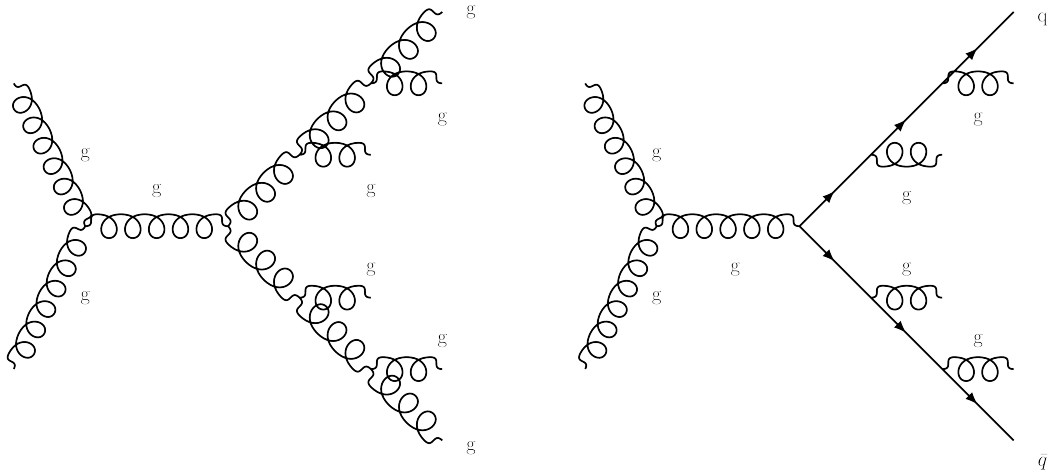


Figure 6: Possible QCD-multijet background with six jets and a similar decay topology to a $t\bar{t}$-all-jet event decay.

3. Experimental Setup

The data used in this thesis are recorded with the CMS[3] detector which is operated at the LHC[4]. The LHC is part of the CERN[5] accelerator complex. It was constructed in the former LEP tunnel near Geneva, Switzerland. CMS is located near the village of Cessy, 100 metres underground [12].

3.1. Large Hadron Collider (LHC)

The LHC is the most powerful particle accelerator and with a length ob 27km also the longest one. It started in 2008 and is the latest upgrade of the CERN accelerator complex, designed to be used for either protons or lead. At one of the four interaction points the CMS detectoris located. ATLAS[6] serves the same purpose in terms of measuring quantities but uses different technologies. While LHCb[7] is used to to analyze b hadrons, ALICE[8] is built to study collisions of heavy ion nuclei [44].

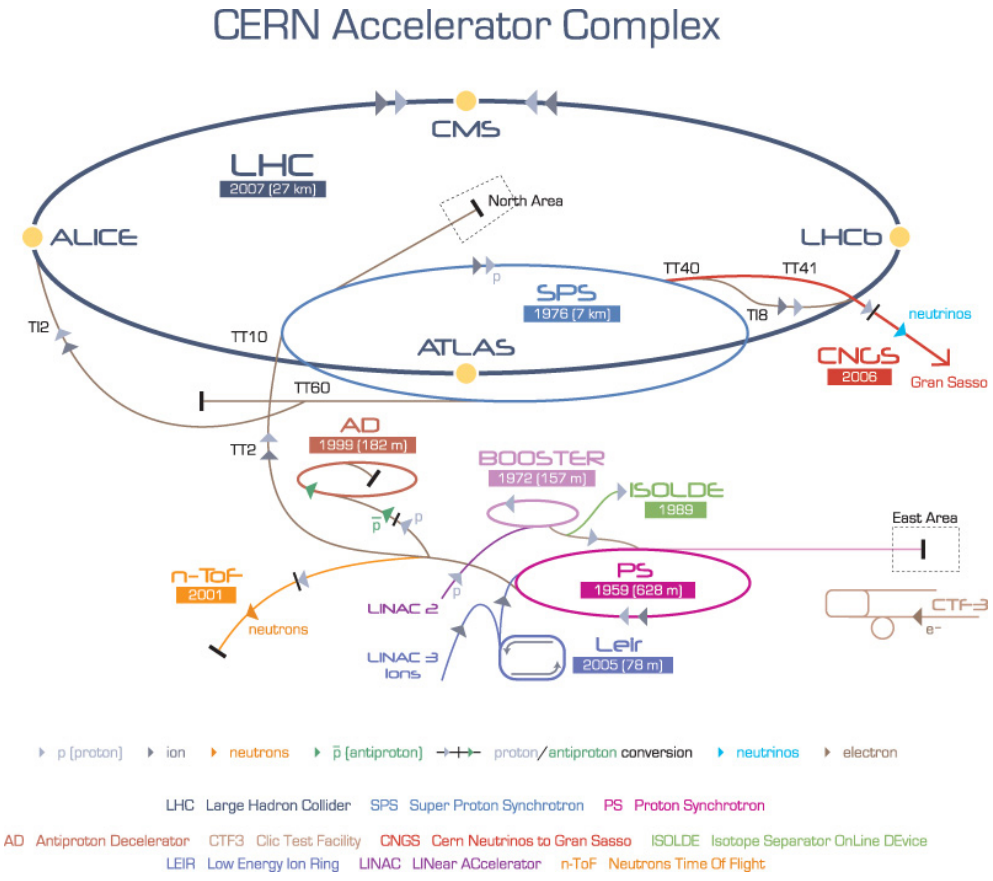


Figure 7: Overview of the CERN accelerator complex [44].

[3] **C**ompact **M**uon **S**olenid
[4] **L**arge **H**adron **C**olider
[5] **C**onseil **e**uropéen pour la **r**echerche **n**ucélaire
[6] **A** **T**oroidal **L**HC **A**pparatu**S**
[7] **L**arge **H**adron **C**olider **b**eauty
[8] **A** **L**arge **I**on **C**ollider **E**xperiment

As source of protons a bottle of hydrogen gas is used. The protons then enter the accelerator chain. In Linac 2 they are brought to an energy of 50 MeV. Afterwards the beam is injected into the Proton Synchrotron Booster (PSB), followed by the Proton Synchrotron (PS), where the energy of the accelerated protons rises to 25 GeV. Another step before injecting the beam to LHC is the Super Proton Synchrotron (SPS) which accelerates the protons to 450 GeV. The beam is then divided into two beam pipes, which circulate in opposed directions inside the LHC. When they reach their final energy of 6.5 TeV, they collide at the collision points at a center off mass energy of $\sqrt{s} = 13\ TeV$. The whole accelerator complex is shown in fig. 7 [44].

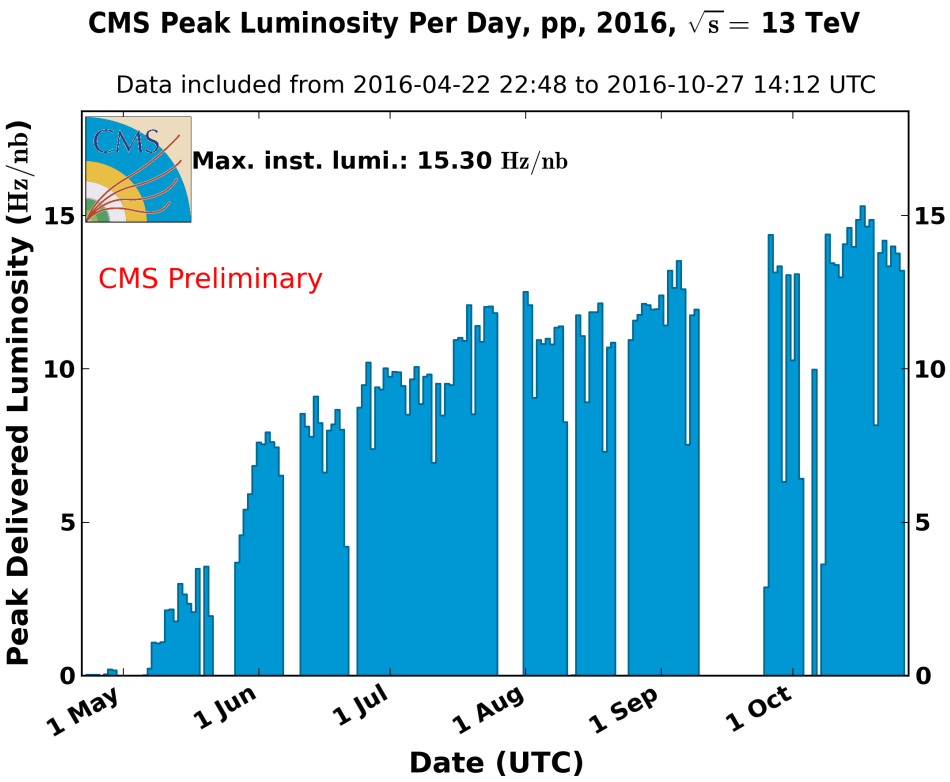


Figure 8: Peak luminosity on a day-by-day basis in 2016 [55]

LHC was designed for a luminosity of $10^{34} cm^{-2} s^{-1}$ at $\sqrt{s} = 14$ TeV but delivered a peak luminosity of 15.3 Hz/nb in the 2016 Run, as shown in fig. 8. The number of protons in each bunch is $N_b \approx 10^{11}$ which collide every 25ns with another bunch. The luminosity for round beams is calculated as

$$L = n_b \cdot \frac{N_b^2 f_{rev} k_B}{4\pi\beta^* \varepsilon_{xy}} \times F, \tag{13}$$

where n_b is the number of bunches per beam, 2808 for Run 2016, f_{rev} the revolution frequency, $\varepsilon_{xy} = \varepsilon_n / (\gamma_{rel} \beta_{rel})$ and F a geometric reduction factor, sometimes referred to as hourglass factor [26]. The integrated luminosity is then received by

$$\mathscr{L} = \int L\, dt, \tag{14}$$

and the number of produced events

$$N_{events} = \mathscr{L} \cdot \sigma, \tag{15}$$

is obtained with σ as cross section for a specific process.

3.2. CMS detector

The construction of the detector is displayed in fig. 9. A slice of it is displayed in fig. 10. The heart of the detector is a 4 Tesla superconducting solenoid. It is about 13m long and has a diameter of 6m.

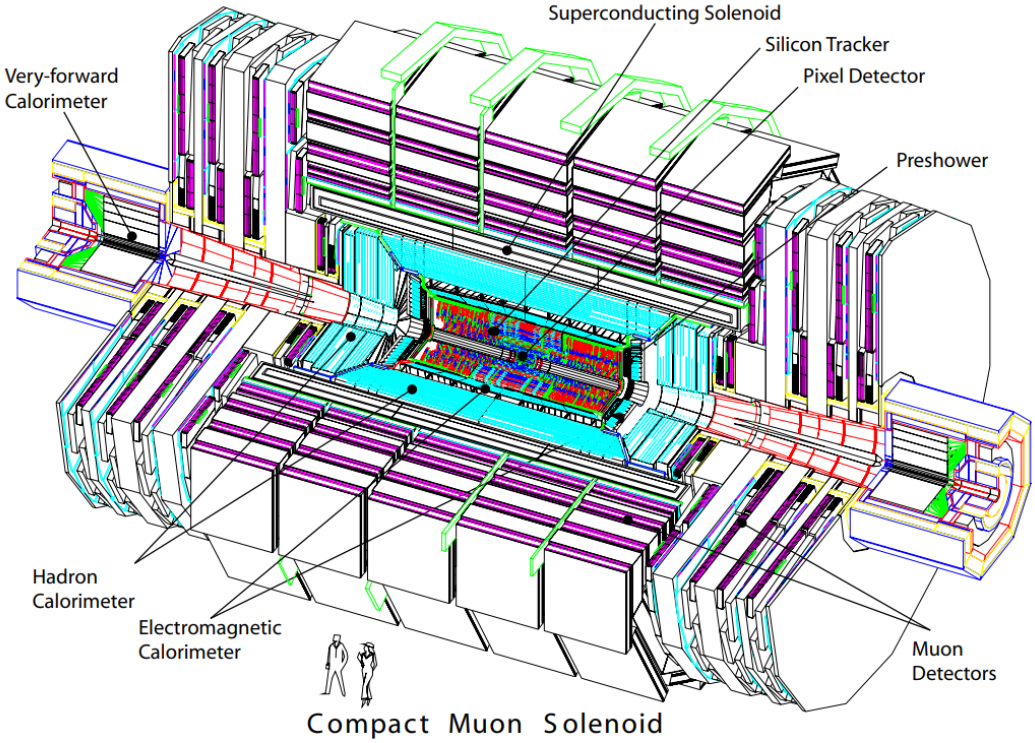


Figure 9: Overview of the CMS detector [12]

The coordinate system is centered at the nominal collision point. The z-axis is parallel to the beam, whereas the y-axis points upwards and the x-axis points in the direction towards the center of the LHC. The azimuthal angle $\phi \in [-\pi, \pi]$ is measured from the x-axis into the x-y-plane. The polar angle θ is measured from the z-axis. Pseudorapidity is then defined as η = -ln(tan($\frac{\theta}{2}$)). The distance in the $\eta - \phi$ plane is calculated as $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$ [12].

The detector consists of four subsystems. The innermost part is a silicon based tracking system. The tracking volume is given by a 2.6m diameter and 5.8m length cylinder, used for reconstruction of trajectories stemming from charged particles. Primary and secondary vertices, for example from the decay of b-hadrons are being identified. Three layers of silicon pixel detectors along with ten layers of silicon strip detectors are used there [12].

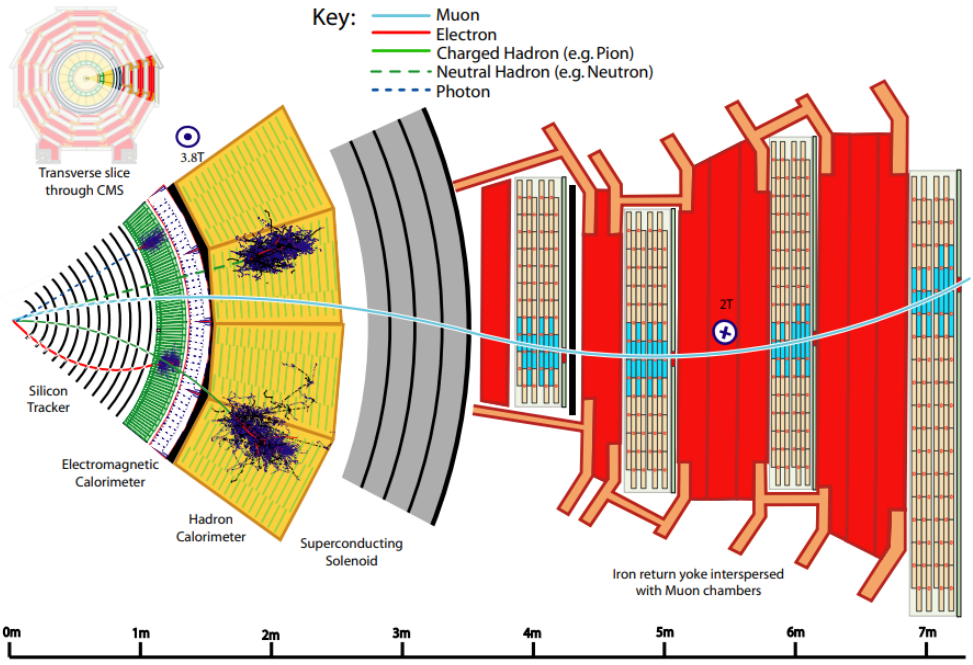


Figure 10: Slice of the CMS detector [28]

Lead tungstate ($PbWO_4$) crystals are utilized for the electromagnetic calorimeter (ECAL). They cover up a region up to $|\eta| < 3.0$. Scintillation light is detected by vacuum photo-triodes (VPTs) and silicon avalanche photo diodes (APDs). For rejecting π^0, a preshower system before the ECAL is installed [12]. The energy resolution of the ECAL is determined through stochastic effects, noise and a constant factor. The parameters have been determined in a beam test [7]. The resulting relative energy resolution is with E measured in GeV:

$$\left(\frac{E}{\sigma}\right)^2 = \left(\frac{S}{\sqrt{E}}\right)^2 + \left(\frac{N}{E}\right)^2 + (C)^2 \tag{16}$$

with S = 2.8% $\text{GeV}^{\frac{1}{2}}$, N = 12% GeV and C = 0.30%.

A brass/scintillator sampling hadron calorimeter (HCAL) surrounds the ECAL, covering up to $|\eta| < 3.0$. Wavelength-shifting fibres are embedded in the scintillator tiles to convert scintillation light. It is then channeled to photo-detectors via fibres and detected by hybrid photo-diodes (HPDs). Additionally a tail catcher is employed in the barrel region to ensure hadronic showers are sampled with about eleven hadronic interaction lengths. An iron/quartz/fibre calorimeter is used to provide a coverage of $|\eta|$ up to 5.0. The emitted Cerenkov light is detected by photo-multipliers [12].

After the HCAL the superconducting solenoid is located, followed by the iron return yoke interspersed with myon chambers, which gives the detector its name. The superconducting magnet reaches a 4-Tesla field in a free bore of 12.5m length and 6m diameter. The stored energy is about 2.5 GJ at full current. The iron yoke weights around 10 000 tons, comprising five wheels and two endcaps [12].

Outside of the solenoid, the myon detection system is placed. It consists of four stations forming concentric cylinders around the beam line. While the outer cylinder has 70 drift chambers, the inner ones have 60 each. The drift chambers may be used as tracking detectors for the barrel myon system due to the low expected rate [12].

Storing and processing all the data obtained from collision process is not possible. To reduce this unhandleable amount of data a trigger system is employed. The trigger system consists of a Level-1 (L1) trigger, made of programmable electronics and is therefore a hardware trigger. The second part is a High Level Trigger (HLT), a software system. The beam collisions occur every 25ns, respectively at 40MHz. The L1 trigger only has an output rate of 100kHz, but is operated at 30kHz. The HLT trigger then performs calculations on the remaining data. There are different HLT paths, each designed for different final states [12].

4. Neural Networks

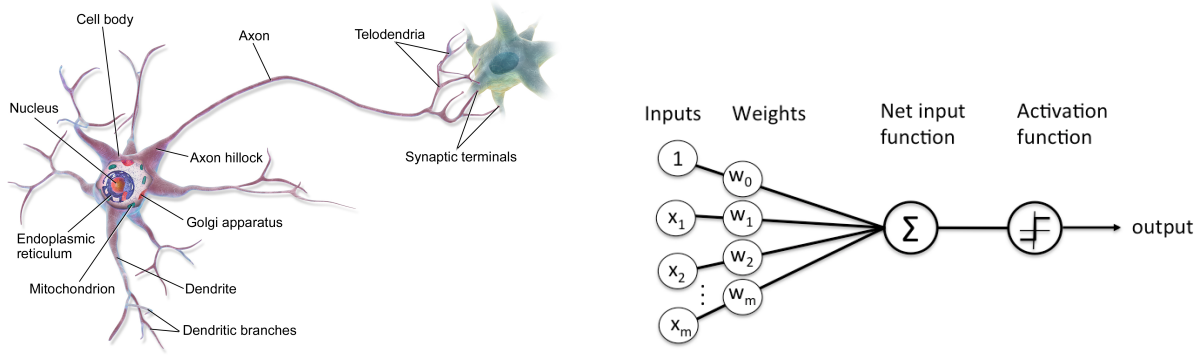
4.1. Introduction

Artificial Neural Networks (ANNs or NNs) have a broad range of applications in our modern society, like pattern or image recognition, speech analysis, classification and learning from observation. Today the usage of NNs is possible due to the increasing computational power in the last decades. Complex processes can be modeled, evaluated and even predicted beforehand, i.e. early alert mechanisms. NNs are algorithms imitating processes of the human brain to learn patterns. Data fed to the NN has to be transformed into numerical values, be it pictures, sound or text [51, 38]. For this thesis NNs are employed for a classification problem, deciding if an event belongs to one of two different classes. Therefore a rather simple approach with only fully connected dense layers is used. There are also a lot of more sophisticated NNs types like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which are not discussed here.

4.2. Operating Neural Networks

Architecture

NNs are build of different layers, each containing a certain number of nodes, which may differ from layer to layer. The smallest building block of a neural network is a node, resembling a neuron of the human brain. fig. 11 shows a neuron compared to a node in a NN.



(a) (b)

Figure 11: (a): neuron from the human body [43] (b): node of a layer of a NN [51].

In fact, nodes are not exactly the equivalent of neurons, but imitate their structure of information transfer in a simple mathematical way. A neuron in the human body takes more than one input, in particular electric signals, and combines them to a single output information and forwards it to other neurons via synapses. This multiple-input single-output functionality is modeled as activation function in a NN. Our brain e.g. has the ability to increase or decrease the strength of certain synapses. This is essential for learning and is transferred to the Neural Network by means of weights for each incoming information. Hence adjusting weights enables

NNs to learn new patterns [38]. An example for a small NN is given in fig. 12. An input layer with three nodes, a hidden layer with four nodes and an output layer with two nodes is shown. Arrows mark the contribution of an information to a node of the next layer. It is sometimes purposeful to use different activation functions for some layers, the output layer in particular.

Deep Neural Networks consist of more than one hidden layer. The number of adjustable parameters (weights) rises sharply with an increasing number of hidden layers. The NN used for the analysis has an input layer with 32 nodes, followed by four hidden layers with 32, 64, 32 and 16 nodes respectively, and an output layer with 2 nodes. Between each of the layers batch normalization layers[9] are put in. The total amount of trainable parameters is 6546. This information is provided by keras itself when calling NN.summary(), replacing NN by the actual employed model[48].

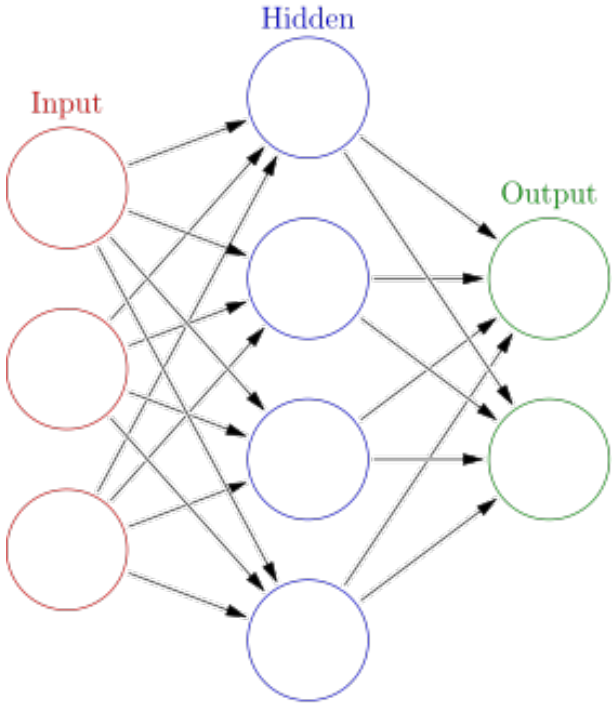


Figure 12: An example of a small NN with three layers [47].

Activation functions define the output of a node, given input. Choosing the best activation function for a problem is important, since step functions will work well when building logical functions like XOR or AND, but won't be of use when using backpropagation algorithm for example, because the derivative of the function is needed. The most common activation functions are summarized in fig. 13. For the analysis SELU is used for all NNs. It is defined as follows:

$$f(x) = \begin{cases} \text{scale} \cdot x & \text{for x} > 0 \\ \text{scale} \cdot \alpha \cdot (e^x - 1) & \text{for x} < 0 \end{cases} \tag{17}$$

[9]Batch normalization layers are used for speeding up learning. This is possible due to making a covariate shift. A detailed explanation can be found here [25].

with $\alpha = 1.67326324$ and scale = 1.05070098 predefined by keras API [54, 48]. Both values are chosen regarding mean and variance of inputs between two layers. This is strongly connected with initialisation of weights. Hence for weight initialisation a Lecun Normal distribution is chosen in this thesis for the used NN.

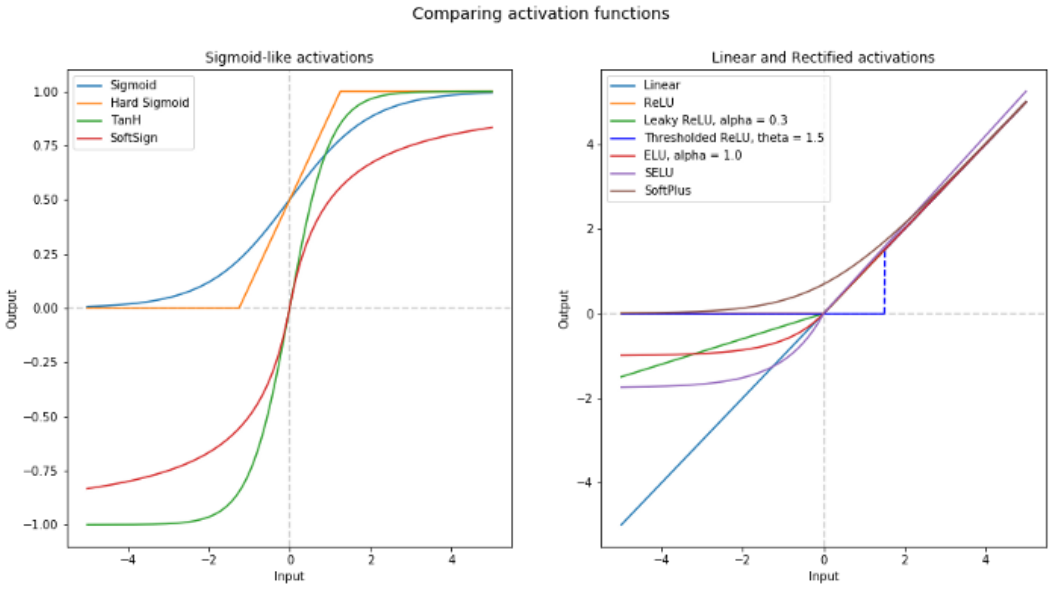


Figure 13: The most common activation functions [49].

Training procedure

In the training process of a NN, a quantity, called loss function, is minimized. There are several predefined loss functions, which serve the most general regression or classification problems. It is possible to define own loss functions, but this would be an advanced approach for a special or way more complicated problem. For this thesis sparse categorical crossentropy as loss function is used, which is an integer based version of the categorical crossentropy, where targets are already labeled as integers. This function is taken when there are two or more labeled classes with an integer representation per class [54, 48]. The categorical crossentropy is calculated as follows [52]:

$$LOSS = -\sum_{i=1}^{output\ size} y_i \cdot log(\hat{y}_i), \tag{18}$$

with y_i as target value, $\hat{y}_i$ as the i-th scalar value in the model output and output size as the number of scalar values in the model output. The minimization process can be performed by various different algorithms. In the analysis part ADAM[10] is employed for the optimization.

Usually a NN is trained on only parts of the whole data set, while an independent sample is used for validation. After each epoch the weights are adjusted to minimize the loss function. To avoid overtraining, which is learning the training sample itself, but not the underlying pat-

[10]No acronym, but derived from 'adaptive moment estimation' [23].

tern, the loss value calculated on the validation sample is observed. The minimization is done exclusively on the training data, but the weights of the model are saved for the further analysis if validation loss is at a minimum.

K-fold cross-validation

When using a NN, the underlying data are split into three sets, training, test and validation. The test set is the part of the data, where the actual test of the model has to be performed. Training and validation data are therefore 'wasted', which leads to lower statistics on the test sample. To reduce this wasting of data, k-fold-cross validation may be used. The whole data set is divided into only two parts, a training/validation sample and a test sample and the training/validation sample is split into k parts. Starting with leaving out the k-th part and training on combined parts 1 to (k-1), this procedure is done k times with leaving out a different one of the k training/validation parts at a time and using it for validation [53]. This is displayed in fig. 14. With this method, sets of hyperparameters can be evaluated, e.g. a mean accuracy score over all iterations with low variance induces a configuration, which may fit the given task.

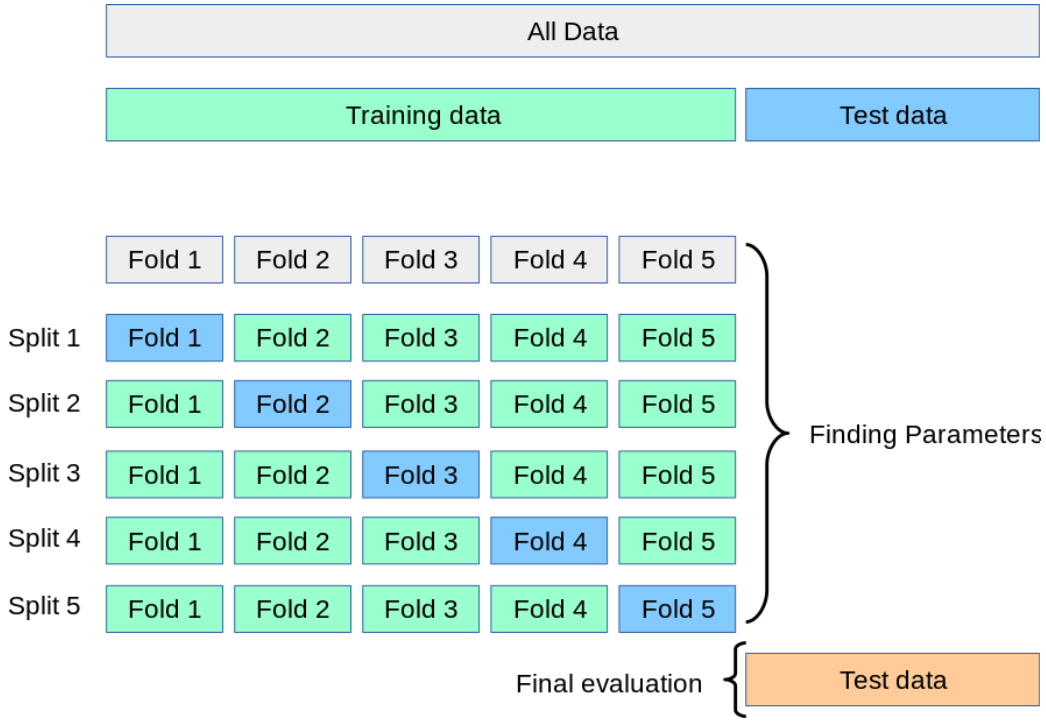


Figure 14: An example of how k-fold cross validation is used for finding parameters [53].

ROC curve and AUC value

A ROC curve (receiver operating characteristic curve) shows the performance of classification models. This graphical representation displays the performance for all threshold values at once. Only two parameters are used, the True Positive Rate (TPR) and the False Positive Rate (FPR)

[45]:

$$TPR = \frac{TP}{TP+FN} \tag{19a}$$

$$FPR = \frac{FP}{FP+TN} \tag{19b}$$

The AUC value (area under curve value) is the area under the ROC curve. In general, a higher AUC value means a better performance of the NN on the classification. But one has to consider the actual used threshold value, since curvatures may differ and a slightly worse AUC value might lead to a better classification at a specific threshold, though. An example of a ROC curve and AUC value is displayed in fig. 15.

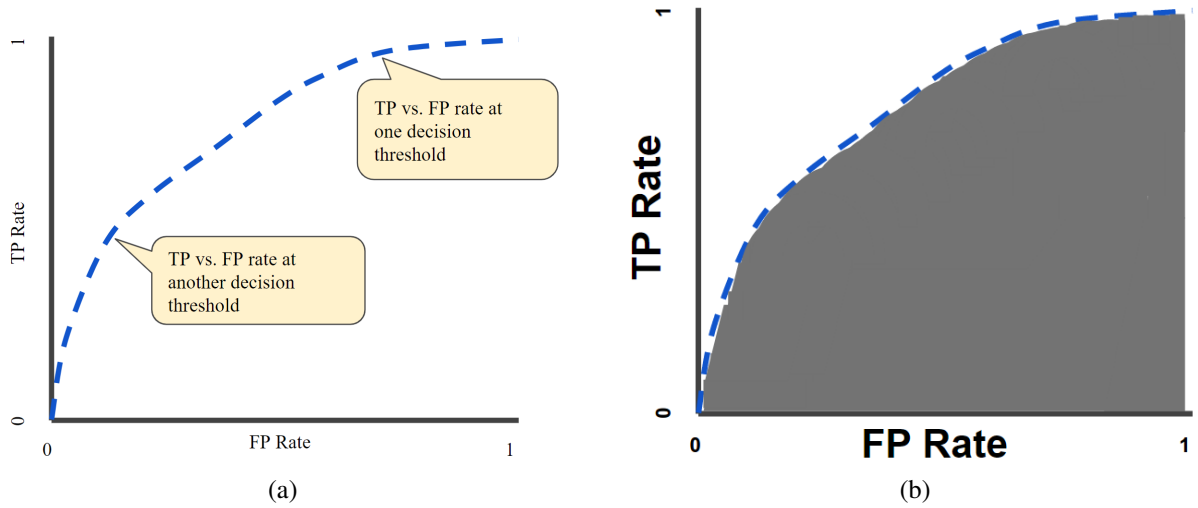


Figure 15: (a) shows the calculation of a ROC curve via TPR and FPR while (b) displays the AUC value [45].

5. Analysis

5.1. Data samples

Experimental data samples

For this analysis the 2016 CMS data set (Run B-H) with an integrated luminosity of 35.9fb^{-1} at $\sqrt{s} = 13$ TeV is used. The rounded single luminosities for each run are shown in tab. 1.

2016B	5.79 fb^{-1}
2016C	2.57 fb^{-1}
2016D	4.25 fb^{-1}
2016E	4.01 fb^{-1}
2016F	3.10 fb^{-1}
2016G	7.54 fb^{-1}
2016H	8.61 fb^{-1}
2016 B-H	35.9 fb^{-1}

Table 1: Run eras of the 2016 CMS data taking. The processed integrated luminositys for the periods (B-H) add up to 35.9 fb^{-1} [42].

Simulated samples

Monte Carlo (MC) simulated events are used to investigate and verify the results. PYTHIA 8.219 [9] for parton shower and hadronization with POWHEG v2 [3, 6, 13, 21] at next-to-leading order (NLO) perturbative QCD was used to simulate $\bar{t}t$ events. The tune CUETP8M2T4 is used [27, 24]. For the parton distribution functions (PDFs) NNPDF3.0 NLO [22] is taken with $\alpha_s = 0.118$ as strong coupling constant. To simulate the response of the CMS detector GEANT4 is used [2].

A sample with generated mass m_t = 172.5 GeV is taken as default. The sample is normalized to next-to-next-to-leading order (NNLO) cross section of $\sigma_{t\bar{t}}$ = 831.76 pb. This is calculated with the TOP++ program [15]. Additional samples with different generated masses are re-scaled using eq. 12. In tab. 2 samples with different generated masses, corresponding cross sections and generated events are listed. QCD multijet samples are generated in different bins of generator-level H_T (see eq. 20). Leading order MADGRAPH with MLM matching scheme [18, 5] is used for the matrix element generation. For fragmentation and hadronization PYTHIA 8 with CUETP8M1 tune is utilized. The generated numbers, cross sections and H_T bins are listed in tab. 3.

m_t^{gen} [GeV]	$\sigma_{t\bar{t}}$ [pb]	N_{gen}
166.5	983.72	19380254
169.5	903.82	29369560
171.5	855.01	19578812
172.5	831.76	77229341
173.5	809.24	19419050
175.5	766.30	59384660
178.5	706.75	16377176

Table 2: Generated top masses, cross sections and event numbers are displayed. The cross sections are calculated with eq. 12 from the default sample with m_t = 172.5 GeV [42].

H_T [GeV]	$\sigma_{t\bar{t}}$ [pb]	N_{gen}
100 - 200	27990000	80684349
200 - 300	1712000	57580393
300 - 500	347700	54537903
500 - 700	32100	62271343
700 - 1000	6831	45412780
1000 - 1500	1207	15127293
1500 - 2000	119.9	11826702
2000 - ∞	25.24	6039005

Table 3: Simulated QCD multijet events, cross sections and generated event numbers per bin [42].

5.2. Event selection

The event selection is based on the analyses in [42, 39]. A more detailed explanation of the data set used along with corrections concerning trigger efficiency are made in [42]. Hence a short overview is given here.

Jet clustering is done by Particle Flow (PF) candidates, using the anti-k_t algorithm with a distance parameter of 0.4 [14, 10]. It is ensured that only charged hadrons from the primary proton-proton vertex are considered for the jet clustering. Only jets with $p_T > 30$GeV and $|\eta| < 2.4$ are used. Additionally jets have to fulfill the standard loose jet criteria, listed in tab. 4.

Neutral Hadron Fraction	< 0.99
Neutral EM Fraction	< 0.99
Number of Constituents	> 1
Charged Hadron Fraction	> 0
Charged Multiplicity	> 0
Charged EM Fraction	< 0.99

Table 4: Standard loose jet identification criteria [40].

The scalar sum of all jet transverse momenta describes the hadronic activity of an event:

$$H_T = \sum_{jets} p_T \tag{20}$$

High Level Trigger path HLT_PFHT450_SixJet40_b-tagCSV_p056 is used to record signal events. A minimum of six PF jets, each with $p_T > 40$ GeV, along with $|\eta| < 2.6$ and $H_T > 450$ GeV is required. Furthermore at least on b-tagged jet is needed. For MC simulated samples an HLT emulation is used instead. To maintain a good agreement between data and simulation an efficiency correction is applied to simulation. Efficiency differences resulting from functions of kinematic parameters have to be corrected. A total normalization difference would not affect the results of the top mass extraction. Jets are ordered in p_T. The hadronic activity and the transverse momentum of the sixth leading jet (p_T^{jet6}) are used to correct the weighting of MC simulated events to match CMS data. The derivation of this correction is given in detail in [42]. For each event the dimensionless correction factor (CF) is then obtained through:

$$CF = 0.986614 \cdot \frac{\left(e^{(-0.0235098\cdot(H_T/\mathrm{GeV}-454.875)}+1\right)\cdot\left(e^{-0.246628\cdot(p_T^{jet6}/\mathrm{GeV}-39.3759)}+1\right)}{\left(e^{-0.0289016\cdot(H_T/\mathrm{GeV}-466.40)}+1\right)\cdot\left(e^{-0.27647\cdot(p_T^{jet6}/\mathrm{GeV}-40.736)}+1\right)}$$

b-tagging

The event topology of $t\bar{t}$ decays, described in ch. 2.3, implies the almost exclusive decay into two b-quarks and a $W^{\pm}$ pair. Jets originating from b quarks can be identified by finding secondary vertices, produced by their decay. Fulfilling $|\eta| < 2.4$ ensures the jets to be reconstructed within tracker coverage.

For b-tagging the Combined Secondary Vertex algorithm (CSVv2) [16, 36] with the tight working point (WP) is used. It combines information about displaced tracks and secondary vertices, associated with jets. An example is given in fig. 16. The efficiency is approximately 50% with a mistake probability of approximately 0.1% at a discriminator value of 0.9535. The number of jets with a discriminator value greater than 0.9535 is referred to as number of b-tags of an event in the following. Information about displaced tracks (dashed blue line) and a displaced secondary vertex (red dot) are combined to rate a jet with a value between 0 and 1. In previous analyses two b-tags per event are required [42, 32]. In simulated pure $t\bar{t}$ decays many events will be lost due to only 50% efficiency at this tight WP. This analysis pursues the approach of loosening this criteria to only one required b-tag per event as a first selection step. Further developed b-tagging techniques with major changes to the efficiency will therefore have a great impact on the received results.

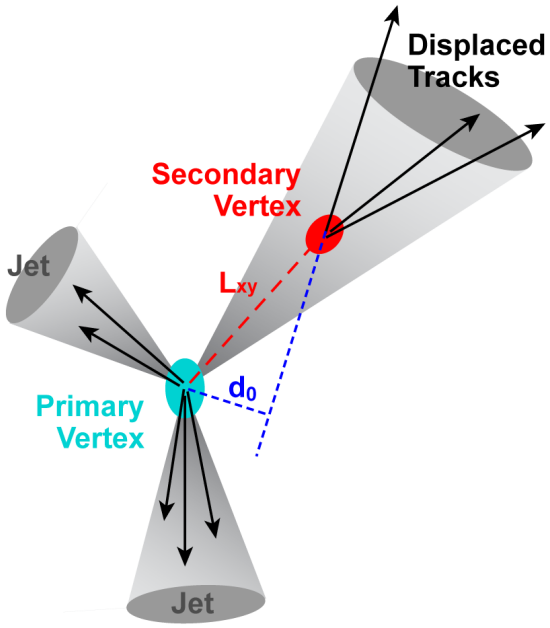


Figure 16: Example of an event with two vertices. The secondary vertex is e.g. produced by the decay of a b quark [46].

Kinematic fit

Regarding the compatibility of selected events with the $t\bar{t}$ decay hypothesis, a kinematic fit is applied to data via the KinFit [4] package of the CMS software. The kinematic fit is applied to the six leading (highest p_T) jets for each possible permutation. It fits the three-momenta (p_T, η, ϕ) and is constraint by $m_{W^+} = m_{W^-} = 80.4$ GeV and $m_t = m_{\bar{t}}$. The value

$$\chi^2 = \sum_{j \in jets} \left[\frac{(p_{T_j}^{reco} - p_{T_j}^{fit})^2}{\sigma_{p_{T_j}}^2} + \frac{(\eta_j^{reco} - \eta_j^{fit})^2}{\sigma_{\eta_j}^2} + \frac{(\phi_j^{reco} - \phi_j^{fit})^2}{\sigma_{\phi_j}^2} \right] \tag{21}$$

is minimized. Quantities with the label 'reco' are from originally reconstructed jets, values labeled with 'fit' are varried to minimize the χ^2 value. The sigma values in the denominator refer to the resolutions of the momenta which are obtained from simulation. The p value, now called P_{GoF}, is obtained via

$$P_{GoF} = 1 - \mathrm{erf}\left(\sqrt{\frac{\chi^2}{2}}\right) + \sqrt{\frac{2\chi^2}{\pi}} e^{-\frac{\chi^2}{2}} \tag{22}$$

with erf as the gaussian error function. Only permutations with $P_{GoF} > 0$ are used for the analysis. The values of P_{GoF} are not used for the selection step, but as an input feature for the NN. A more detailed description on the kinematic fit can be found in [42, 16] where it plays a key part in the analysis.

Jet combinatorics

Each event consists of up to 90 permutations. This is the maximum number of possible different assignments of the six (p_T)-leading jets. Ordering 6 items in different ways would, mathematically, lead to 6! (=720) permutations but changing e.g the assignment of the two decay products of one of the W bosons would not lead to a different decay kinematics of the event. Permutations with high P_{GoF} values are more likely to be arranged correctly and more likely to be $t\bar{t}$-decays than those with a low value. Using at least six permutations reduces the possibility to overlook a permutation which actually is the correct candidate of the permutations for the $t\bar{t}$-decay hypothesis. Tab. 5 shows the average number of the first six permutations per event passing the selection for the NN used later. At the first three selection steps, not all events have six permutations which fulfill $P_{GoF} > 0$. As an example , this is why in tab. 5 the average number of passing permutations is six only on average ($\approx$5.987) for the first selection step.

Preselection

The preselection is split into an event-based and permutation based selection. A sketch is shown in fig. 17. Event based refers to criteria, which are the same for all permutation. A passing event needs at least one b-tag, $H_T > 450$ GeV and $p_T^{jet6} > 40$ GeV. These criteria are required due to

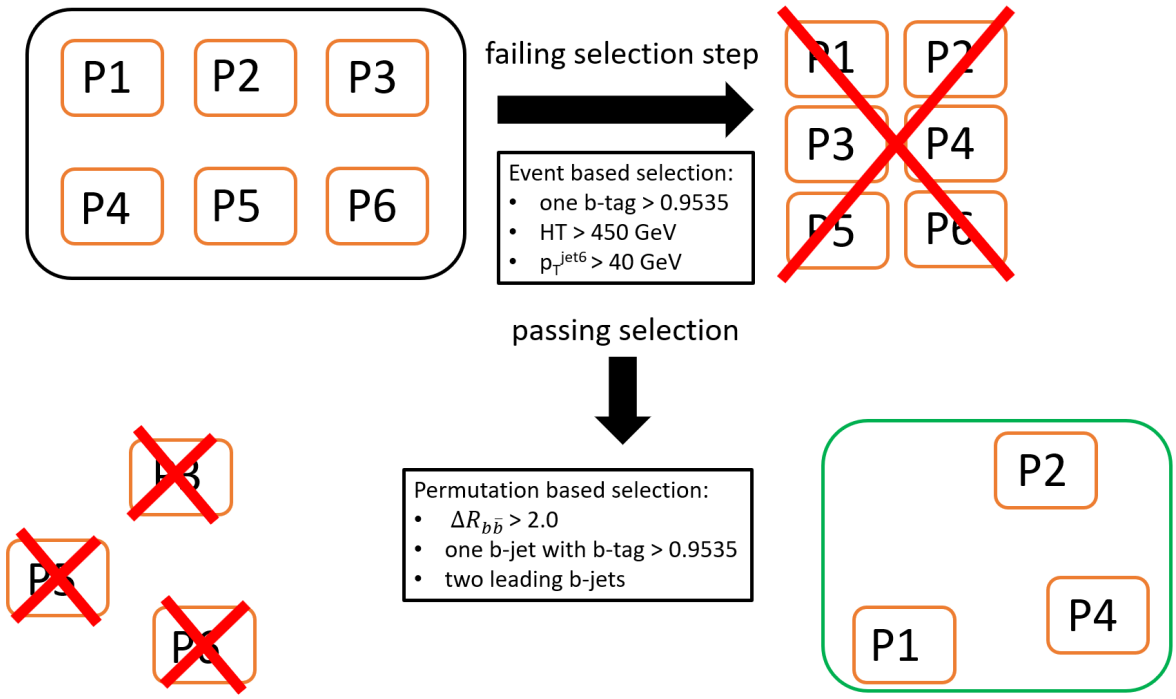


Figure 17: Selection steps for the NN input. If any event based selection fails, all permutations are disregarded. If an event passes the three first selection steps and e.g. permutations P1, P2 and P4 fulfill all the criteria of the permutation based selection, they are used for the analysis.

trigger conditions. Detailed information about this topic is given in [42]. If an event passes the criteria of the event based selection, the permutation based selections follows. Requiring at least one b-tag is quite a lose criteria compared to [42, 16] for example, regarding b-tagging efficiency. Suppose a $t\bar{t}$ decay with two b-jets and an uncorrelated b-tagging efficiency of 50%, the probability of assigning both b-jets correctly is only 25%, but assigning at least one of the two b-jets correctly is 75%. Requiring two b-tags will lead to cut away lots of actual $t\bar{t}$ events.

The selection step $\Delta R_{b\bar{b}} > 2.0$ aims at preparing the samples for a later carried out background estimation. This has to be done due to splitting of gluons into two b quarks, hence faking parts of the $t\bar{t}$ decay topology, as mentioned in ch. 2.3. In [42] it is mentioned that this correlated production of two jets cannot be reproduced by the background estimation method used in this analysis.

The last two selection steps are made to take into account the decay topology itself. Jets could be assigned as b-jets but not be b-tagged. In contrast, jets faking b-jets could have a b-tag value greater than 0.9535. As a consequence at least one jet assigned as b-jet has to be b-tagged. Supplementary, both b-jets are to have a greater b-tag value than the four remaining jets (two leading b-jets).

The selection steps are applied to CMS data, MC simulated QCD and $t\bar{t}$ events. Relying on $\frac{\#\ signal}{\#\ (signal+background)}$ as purity is not purposeful due to using more than one permutation per selection step in parts of the analysis As shown in tab.5, simulated events show a difference to data of about 10%. MC simulated events were also re-weighted to match CMS data taking and corrected for a trigger efficiency. These weights have a range from approx 10^{-4} up to

data set	selection step	passing perm.	∅ per event
MC $t\bar{t}$	one b-tag value > 0.9535	5 543 876	6.0
	H_T > 450 GeV	5 428 555	6.0
	$p_{T[5]}$ > 40 GeV	4 929 814	6.0
	$\Delta R_{b\bar{b}}$ > 2.0	3 295 309	4.1
	one b-jet with b-tag value > 0.9535	1 157 689	2.2
	two leading b-jets	227 340	1.4
MC QCD	one b-tag value > 0.9535	42 916 396	6.0
	H_T > 450 GeV	41 547 511	6.0
	$p_{T[5]}$ > 40 GeV	35 594 862	6.0
	$\Delta R_{b\bar{b}}$ > 2.0	25 412 107	4.4
	one b-jet with b-tag value > 0.9535	8 571 730	2.5
	two leading b-jets	1 583 251	1.7
CMS Run B-H	one b-tag value > 0.9535	53 190 207	6.0
	H_T > 450 GeV	51 267 918	6.0
	$p_{T[5]}$ > 40 GeV	42 577 096	6.0
	$\Delta R_{b\bar{b}}$ > 2.0	29 736 731	4.3
	one b-jet with b-tag value > 0.9535	10 148 825	2.2
	two leading b-jets	1 893 539	1.5

Table 5: Cut flow of MC Simulated QCD and $t\bar{t}$-events, compared to data. The last column shows the average number of passing permutations per event. The number of passing MC simulated permutations is weighted to match the integrated luminosity of CMS Run B-H of 35.9fb^{-1}.

about 400 before adjusting to the integrated luminosity of data. This adds another factor of 35.9 to the weights, because the simulated weights were normalized to sum up to 1 fb^{-1} and the integrated luminosity for CMS Run B-H adds up to 35.9 fb^{-1}. Therefore, when using multiple permutations per event, purity will only be used as a forecast on remaining signal fractions in the later used data samples. After applying the NN to the samples, only one permutation per event is allowed to pass. Then the purity will be used to compare results to previous analyses.

5.3. NN construction and feature selection

The construction of the NN is inspired by [30], where two layers with 30 nodes connected to a 2 node layer output is used for 5 different input features. Since 12 input features are used here, the NN is chosen larger and deeper. Exclusively all layers of the networks are fully connected dense layers. Between every two layers , except before the output layer, a batch normalisation layer (BNL) is inserted to increase the performance of the NN. The Hyperparameters were manually optimized regarding stability of the training process and efficiency of the then obtained classifier. AUC values of obtained ROC curves are used for comparing the NNs. Implementation is done with keras [48] and TensorFlow backend [54]. The hyperparameters are displayed in tab. 6.

For the training procedure, the input samples are re-weighted by a constant factor to match the equivalent of 10^6 weighted events. This is necessary due to before observed stability problems

Nodes of layers	32, BNL, 32, BNL, 64, BNL, 32, BNL 16, 2
Activation function	Softmax (output layer) and SELU (other layers)
Kernel/bias initializer	Lecun normal
Batch size	256
Loss	Sparse categorical crossentropy
Metric	Accuracy

Table 6: Hyperparameters for all Neural Networks used in this thesis.

with high MC event weights. A batch size of 256 is used to avoid weighting issues affecting stability of the training. Batch size defines the number of samples which will be propagated through the NN before the models internal parameters are updated. If using a larger batch size, events with high weights will be compensated by other events in the batch and the loss function will not be driven away. The convergence of the training procedure is ensured by observing the loss value. For consistency reason, reweighting of this kind is maintained for training on CMS data, although all weights have the value 1. To ensure a good generalization k-fold cross validation with k=10 is used . Thus all events are randomly split into 10 partitions, each adding up to 1/10 of the events with respect to their adjusted weights.

The input features are defined in tab. 7. The choice of the parameters for this classification problem are motivated through:

- Representing each event kinematically through the transverse momenta p_T. Additionally the reconstructed masses of the decay products stemming from the W boson decays and the fitted top mass is used. In ch.5.5.2 it will be shown, that using this feature will enable the NN to learn about the underlying mass distribution of a sample, but not learn a specific mass value. Therefore using it as feature is not a problem for a top mass extraction. The angle $\Delta R_{b\bar{b}}$ is used to have more information about the spatial spread. It also plays an important role for the background estimation.

- Parameters must not work as b-tagging indicators. Therefore further information of detailed properties related to b-tagging have to be withhold. This is essential for the CWoLa approach when the data is split by the number of b-tags an event possesses.

- Not only maximising the AUC value of ROC-Curves has to be considered but also the signal purity at specific working points.

5.4. Performance of the NN on MC simulated events

In this chapter the NN is trained and evaluated on MC simulated events. As a first step to compare results training is done with pure samples of QCD and $t\bar{t}$. In fig. 18 the results of the training are shown. The result is not only used as a benchmark test for the CWoLa approach in ch. 5.5 but also for deciding if a a selection with a NN is able to improve the previous selection from [42, 34].

Parameter	ROOT tree address
p_T of the six leading jets	top.recoB1.Pt top.recoB2.Pt top.recoW1Prod1.Pt top.recoW1Prod2.Pt top.recoW2Prod1.Pt top.recoW2Prod2.Pt
p_T of the sixth jet ($jet-level$)	jet.jet[5].Pt
$\Delta R_{b\bar{b}}$ with $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$	top.fitB1.DeltaR(top.fitB2)
$m_{W_{1,2}}^{reco}$	top.recoW1.M top.recoW2.M
P_{GoF}	top.fitProb
m_t^{fit}	top.fitTop1.M

Table 7: Input features used for the $t\bar{t}$ vs. QCD classification.

The result of the previous analysis is shown in tab. 8. If a higher purity at a similar efficiency or the same purity at an increased efficiency is achieved, the NN is considered to be performing better. In tab. 8 as a first selection step two b-tags are required with only considering the permutation with the highest P_{GoF} value per event. At the final step a purity of 75.2% at 10799 data events is achieved.

Since multiple permutations per event are allowed in the selection beforehand, the NN is also used for deciding which permutation is taken for the further selection, so only the one with the highest prediction of the classifier is used in the further analysis. Hence a well defined usage of purity is possible. MC simulated $t\bar{t}$ events hold information about the true combination type, a value which tells if jet-matching is done correctly if we assume a $t\bar{t}$ decay. Applying the obtained classifier from fig. 18 at a cut value of 0.9 on the two leading b-jets selection results in a purity of 75.4% at 20938 data events. The fraction of correctly permutated $t\bar{t}$ events is about 61%, which is also an increment of 10 percentage points to previous analyses [42, 34]. For comparison, purity and correctly permutated fraction had to be calculated on the whole data set of simulated events, which is somehow violating good machine learning practice, because parts of the data were used for training, also. In the first place, these results are just meant to show, that an improvement of previous selections is possible through using this particular NN configuration along with the described input features.

Selection step	data	$t\bar{t}$	purity
2 b-tags	1392670	237126	17.0%
$H_T > 450$ GeV	1342630	232498	17.3%
$p_T^{jet6} > 40$ GeV	1130714	212697	18.8%
$\Delta R_{b\bar{b}} > 2.0$	359456	103882	28.9%
$P_{GoF} > 0.1$	10799	8126	75.2%

Table 8: Cut flow from [42]. This is used as a benchmark for this analysis.

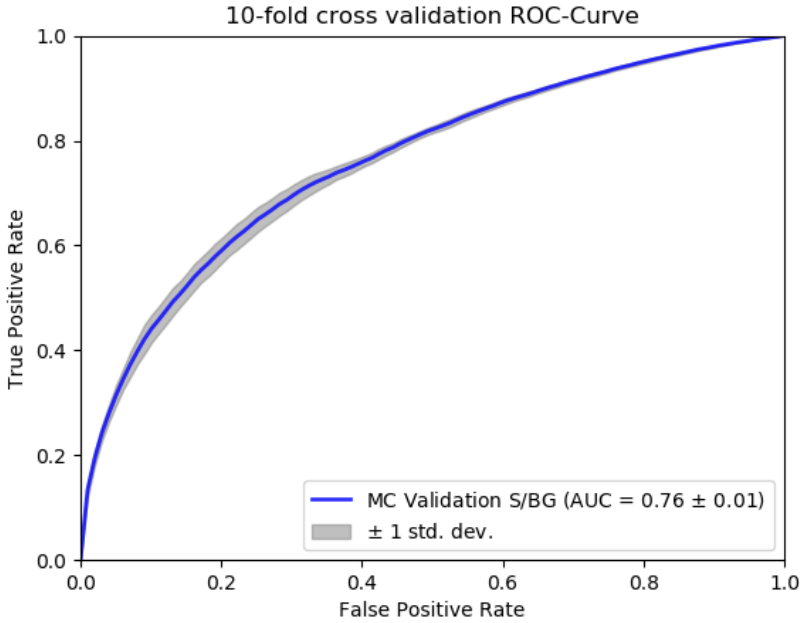
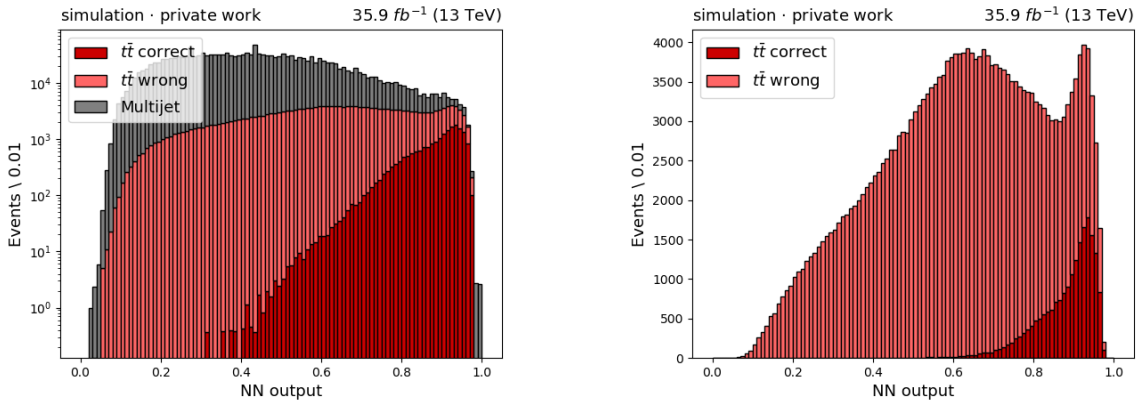


Figure 18: ROC-Curve of the $t\bar{t}$ vs. QCD classification. The grey area is 1σ uncertainty on the ROC-Curve and AUC value, received from 10-fold cross validation. The calculation of the ROC-Curve takes into account all permutations per event, in average 1.4-1.7 permutations, according to tab. 5.



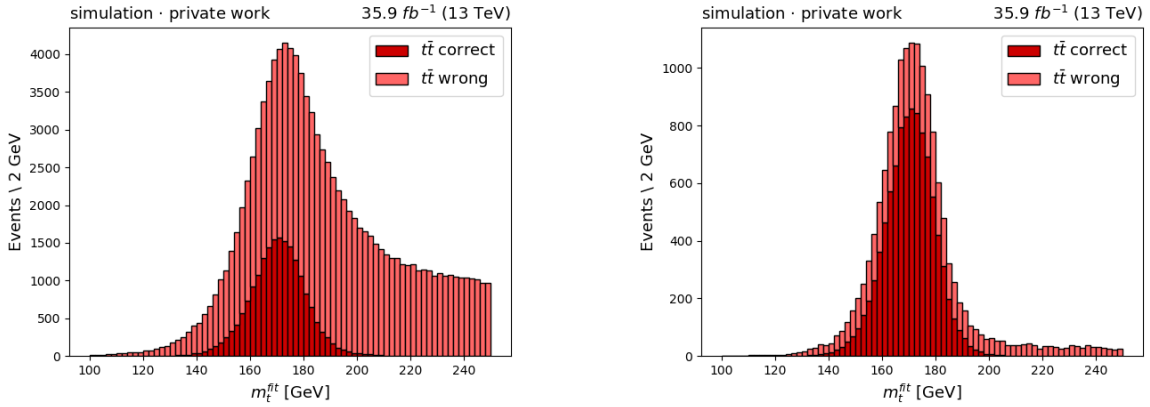
(a) (b)

Figure 19: (a) shows the stacked NN output distribution of simulated samples. (b) shows the NN output distribution for only the $t\bar{t}$ sample, divided into correct and wrong permutations. The model rates correct permutations higher than wrong permutations on average.

Fig. 19 shows the stacked NN output distribution for simulated samples. Only here all permutations are taken into account. Some 'spikes' are appearing, for example at a value of around 0.45. They occur if multiple permutations of an event with a high weight pass the preselection criteria shown in tab. 5. But these spikes are at low prediction values and will be sorted out by applying a cut value on the NN selection, so they won't affect the further analysis.

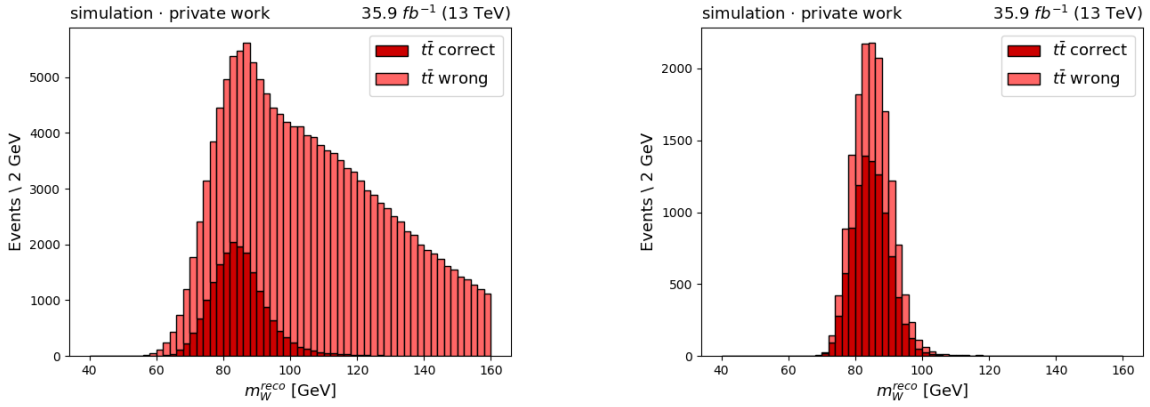
The $t\bar{t}$ prediction distribution has two peaks. It looks like they're originating from two different distributions. This is due to a subdivision into correct and wrong permutations which is

shown in fig. 20. The NN also learns to prefer correct permutations over wrong ones, which is an important information. When training and evaluation is done on data, where no information about the permutation type is provided, a similar outcome is assumed.



(a) (b)

Figure 20: Distributions of m_t^{fit}: (a) shows the selection after the preselection cut flow (b) shows the distribution after applying an additional cut at 0.9 on the NN classifier. Only the permutation with the highest prediction per event is used.



(a) (b)

Figure 21: Distributions of m_t^{fit}: (a) shows the selection after requiring two leading b-jets and (b) shows the distribution after applying a cut value of 0.9. Only the permutations with the highest prediction per event is used.

Fig. 20 and fig. 21 show the resulting m_t^{fit} and on m_W^{reco} for $t\bar{t}$ simulation. For m_W^{reco} the average of the masses of both w boson masses is taken. The kinematic fit is constraint by a W boson mass of 80.4 GeV. A highly pure m_W^{reco} distribution would peak sharply around this value. Furthermore it is expected to see a sharp peak around the generated top mass of 172.5 GeV in the m_t^{fit} distribution. With the NN classification the peaks get narrower around this value, which implies that the NN does the selection as anticipated.

5.5. CWoLa

CWoLa means **C**lassification **W**ith**o**ut **La**bels and is a weak supervision training method for NNs. The idea to use it for this analysis is based on [30]. There it is introduced to learn from mixed samples in high energy physics. A brief overview of the essential key points of the training method is given here. They are partly augmented to fulfill the purposes of this analysis, since in said source it was used to identify quark/gluon jets.

Classification problems are about distinguishing two processes from each other, here referred to as signal and background, standing for $t\bar{t}$ and QCD in this specific case. An optimal binary classifier, which distinguishes $t\bar{t}$ from QCD where S means signal ($t\bar{t}$) and B means background (QCD) is the likelihood ratio $L_{S/B}(\vec{x}) = p_S(\vec{x})/p_B(\vec{x})$. Suppose two samples p_s and p_B with pure signal and background events are given. Let $p_S(\vec{x})$ and $p_B(\vec{x})$ be the underlying distributions. Then each event x_i drawn from one of the samples is assigned with a label $u_i \in \{S,B\}$. This case is called **full supervision**. If two samples M_1 and M_2 consist of both signal and background events, but the composition of the two samples is known, e.g

$$M_1 = f_1 \cdot p_S(\vec{x}) + (1-f_1) \cdot p_B(\vec{x}) \tag{23}$$

$$M_2 = f_2 \cdot p_S(\vec{x}) + (1-f_2) \cdot p_B(\vec{x}) \tag{24}$$

with prior known fractions f_1 and f_2. Then the solution is to use several different samples with different fractions [30, 29] along with an augmented loss function, which includes information about the fraction f_1 and f_2. This procedure is counted as weak supervision.

CWoLa is another form of weak supvervision. The only difference between learning from label proportions as described in the paragraph before is the lack of knowledge on the label proportions. Still it is possible to obtain an optimal classifier with this method. A short proof from [30] is given here.

Theorem (CWoLa): Let M_1 and M_2 be defined as in equations 23 and 24 with the only restriction $f_1 > f_2$. An optimal classifier h trained to distinguish M_1 from M_2 is also optimal to distinguish S from B.
Proof: Considering distributions p_{M_1} and p_{M_2}, an optimal classifier is the likelihood ratio $L_{M_1/M_2}(\vec{x}) = p_{M_1}(\vec{x})/p_{M_2}(\vec{x})$. For distinguishing S from B it is $L_{S/B}(\vec{x}) = p_S(\vec{x})/p_B(\vec{x})$. With eq.23 and 24 it leads to an algebraic relation:

$$L_{M_1/M_2}(\vec{x}) = \frac{p_{M_1}(\vec{x})}{p_{M_2}(\vec{x})} = \frac{f_1 \cdot p_S(\vec{x}) + (1-f_1) \cdot p_B(\vec{x})}{f_2 \cdot p_S(\vec{x}) + (1-f_2) \cdot p_B(\vec{x})} = \frac{f_1 \cdot L_{S/B}(\vec{x}) + (1-f_1)}{f_2 \cdot L_{S/B}(\vec{x}) + (1-f_2)} \tag{25}$$

According to the *Neyman-Pearson lemma* an optimal classifier is the likelihood ratio $L_{S/B}$ or a monotonically related classifier to it. The ratio at the end of eq.25 is a monotonically increasing rescaling of the likelihood $L_{S/B}$ since its derivative is positive as long $f_1 > f_2$.

5.5.1. CWoLa on MC Simulated Events

As a test of the reliablility on the results gained via CWoLa, the dependency on the sample size N and the fraction sizes f_1 and f_2 are examined. The $t\bar{t}$ and QCD samples are split into 90% training and 10% validation for k-fold cross validation. To maintain a higher training stability samples are normed to 10^6 weighted events. The training sample then is further split into mixed samples with fractions f_1 and f_2. The validation data is not re-weighted because it has to represent the whole data set and the shares of $t\bar{t}$ and QCD events on the whole sample. Fig. 22 displays the results for different samples sizes N and for fractions $f_1 = 1 - f_2$. After reweighting, N = 250 000 corresponds to around 25% of the training data while N = 10 000 is only 1% of training data. The AUC value increases slightly with the number of training events and a higher number of training events also increases the stability of the training process. The obtained AUC values for a single run then spread less about the mean AUC value. The calculation of the ROC-AUC values were done five times per fractions $f_1 = 1 - f_2$ for each sample size to stay at a reasonable computing time.

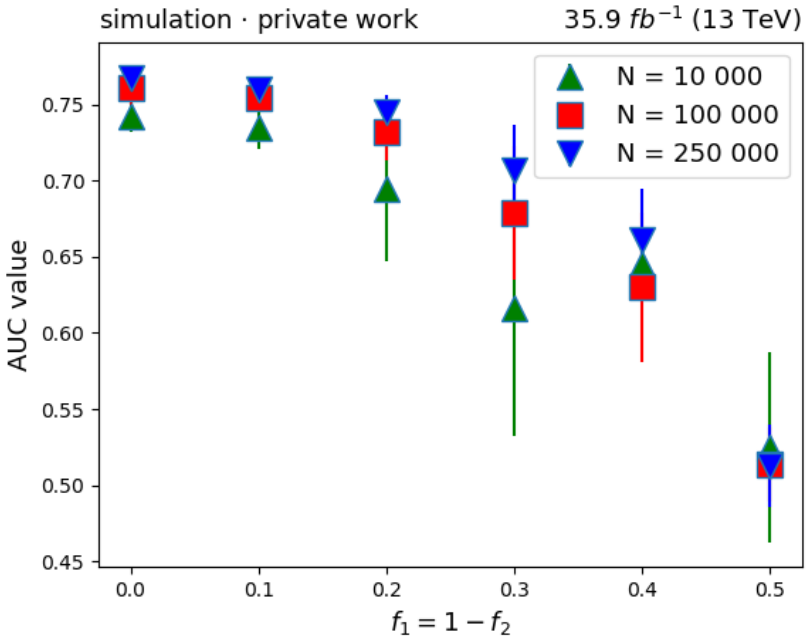


Figure 22: AUC values for different sample sizes and fractions f_1 and $f_2 = 1 - f_2$. The calculation was done five teams. The AUC values represent the mean value and the error bars the standard deviation of the mean value.

The impact of using $\Delta R_{b\bar{b}}$ as a selection step is displayed in tab. 9 as an example on how the signal fractions are affected. The estimated signal fractions after applying all selection steps are calculated while only the cut value of $\Delta R_{b\bar{b}}$ is varied. This leads to an improved difference of the signal fractions f_1 and f_2 of a sample containing one b-tag events in contrast to one with two b-tag events.

If N is big enough the fractions play a smaller role unless $f_1 \approx f_2$. However it is quite unrealistic for this application to have two samples with $f_1 = 1 - f_2$. Fig. 24 shows the correlation

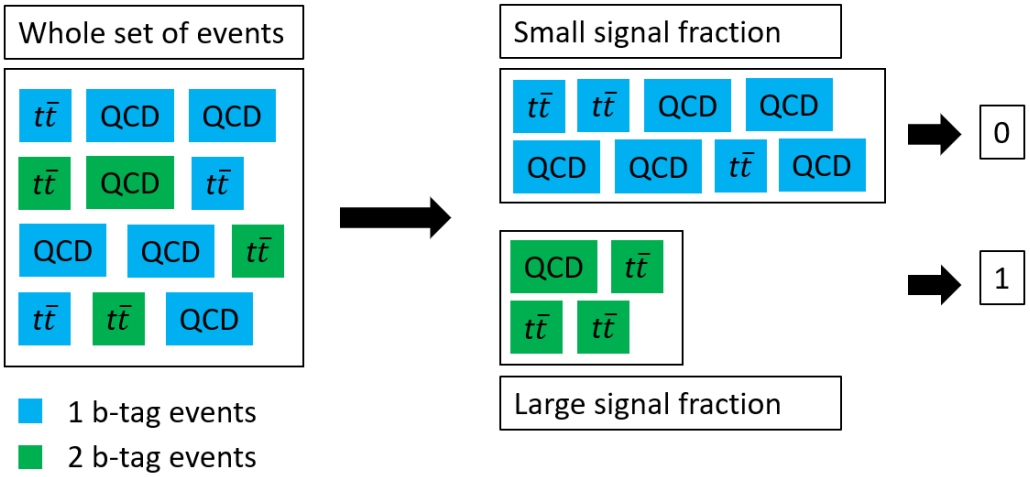


Figure 23: Splitting of the training data by means of the number of b-tagged jets. The whole set of events is split into two samples with either one b-tag value > 0.9535 or two b-tag values > 0.9535. The received samples are marked with 0 (=background) and 1 (=signal) for training.

selection step	$t\bar{t}$	QCD	signal fractions
$\Delta R_{b\bar{b}} > 0.0$	225 442 (1 b-tag) 88 269 (2 b-tag)	2 119 734 (1 b-tag) 360 364 (2 b-tag)	$f_1 =$ 9.6 % $f_2 =$ 19.7 %
$\Delta R_{b\bar{b}} > 1.0$	207 553 (1 b-tag) 82 674 (2 b-tag)	1 825 997 (1 b-tag) 268 682 (2 b-tag)	$f_1 =$ 10.2 % $f_2 =$ 23.5 %
$\Delta R_{b\bar{b}} > 2.0$	161 892 (1 b-tag) 65 448 (2 b-tag)	1 405 325 (1 b-tag) 177 925 (2 b-tag)	$f_1 =$ 10.3 % $f_2 =$ 26.9 %

Table 9: The selection steps are made as described in fig 17. Only $\Delta R_{b\bar{b}}$ is varied. The results display the signal fractions after requiring two leading b-tags (see fig. 23) in both one b-tag and two b-tag sample.

of $f_1, f_2 \in [0, 0.4]$ in steps of 0.025 measured by AUC value. The range of the signal fractions interval is motivated through tab. 9. Splitting of MC simulated $t\bar{t}$ and QCD events into samples with one b-tag, respectively two b-tags, will lead to different signal fractions, because $t\bar{t}$ events are more likely to have two jets b-tagged than QCD multijet events, which is also reinforced by the results displayed in tab. 9.

For now the (pseudo-)purity of samples which are divided in such a way by taking into account all passing permutations per event is estimated. This is strictly speaking not correct, but gives an approximate predictive about the signal fractions. Results are displayed in tab. 5. Before preselection a purity of 11% $t\bar{t}$ events in data is achieved. After the preselection cut flow the purity increases to 15%. These selection steps do not only have the purpose to increase the purity, but to prepare the samples for the splitting into samples with one or two b-tags. The selection process is displayed in fig. 23. The red cross in fig 24 marks the spot of the obtained signal fractions from tab. 9 with $\Delta R_{b\bar{b}} > 2.0$. The AUC value lies between 0.65 and 0.69. The error on it is ± 0.02, which is obtained through 10-fold cross validation for each training pair f_1, f_2. Although increasing AUC values correlate with increasing differences

$|f_1 - f_2|$, regions near $f_1 \approx f_2$ are quite unstable. Thus every selection increasing this difference before the training procedure increases the performance of the NN significantly.

Annother possibility to increase the NN performance and stability is to increase the number of used training data points which was shown before in fig. 22. That's why one wants to rely on a data driven approach, because the number of data points of CMS data taking is larger by approx. a factor of 3 than the ones of MC simulated events.

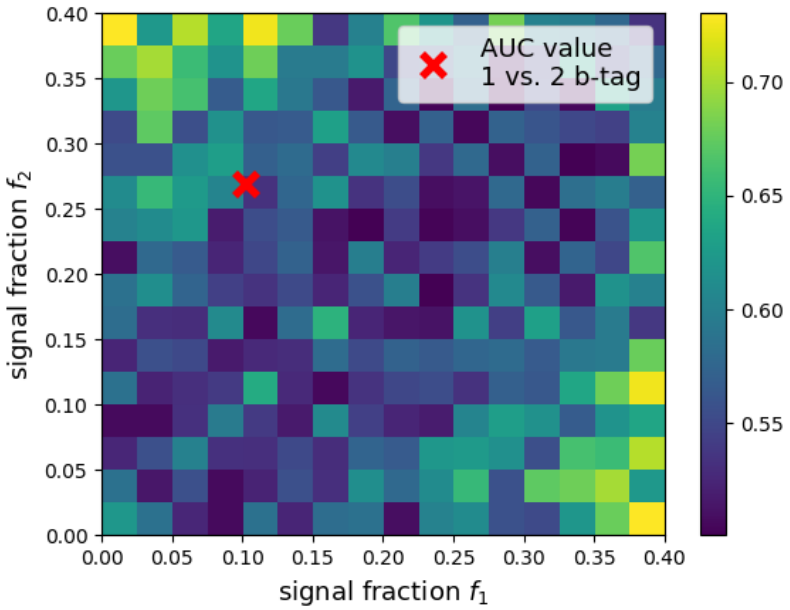


Figure 24: AUC values for different sample size and fractions $f_1, f_2 \in [0, 0.4]$ in steps of 0.025. The expected AUC values for two samples divided into events with 1 b-tag and 2 b-tags is marked in the map. This value corresponds to f1 = 10.3 % and f2 = 26.9 % . AUC values are calculated for fractions above the first bisector and then mirrored at it.

To validate the forecast of the AUC map, a NN is trained according to the splitting done in fig. 23. MC simulated events are split before into 90% training and validation and 10% test data to evaluate the separation into $t\bar{t}$ and QCD events. The result is shown in fig. 25. The test data is the green ROC-Curve, while the blue ROC-Curve shows the validation obtained from 10-fold cross validation training. The blue cross points at a possible selection of P_{GoF} from previous analyses, which is used as a benchmark. If the NN was not learning anything, both ROC-Curves would synchronize with the straight chance line. However, like displayed in the figure, if the NN learns to separate $t\bar{t}$ from QCD at least a bit, the ROC-Curve of the validation sample (1 and 2 b-tag sample) will also differ from chance. The reason is the different signal fractions in the samples with one, respectively two b-tags. In ch. 5.5.2 the output distributions are displayed to verify that no b-tagging NN is trained.

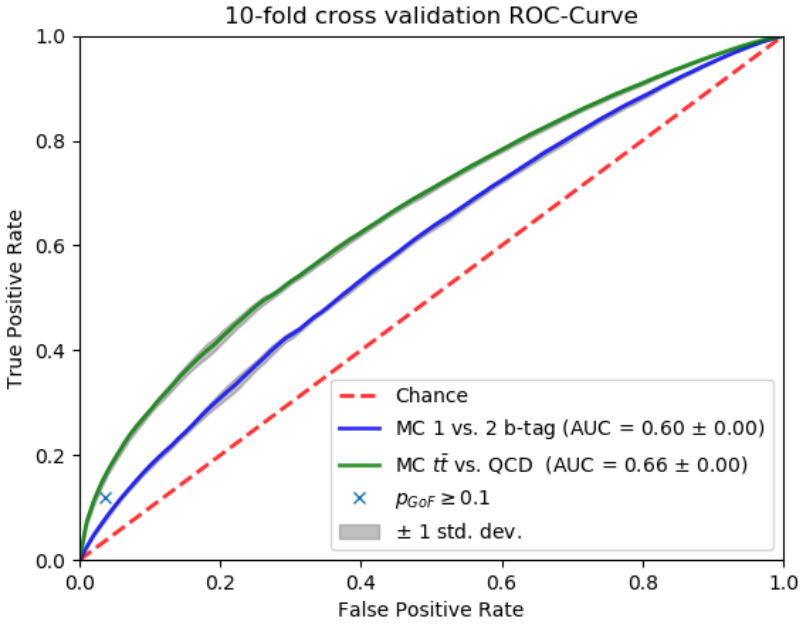


Figure 25: Results of CWoLa on MC simulated events are displayed. The green curve shows the separation of $t\bar{t}$ and QCD events, while the blue one shows the separation of events with 1 and 2 b-tags. The blue cross marks a possible selection step P_{GoF}, regarding $t\bar{t}$ vs. QCD for comparing results.

5.5.2. CWoLa on data

From now on the training is done entirely on data. Fig. 26 shows the training, evaluation and testing procedure, which differs from the one used for simulation. Training and validation are calculated on data, whereas testing is done on simulated events to rate the performance. Training samples are re-weighted to match the equivalent of 10^6 events. For each training, the NN weights leading to the lowest validation loss, which is calculated on 1 and 2 b-tag samples, are used for the further analysis.

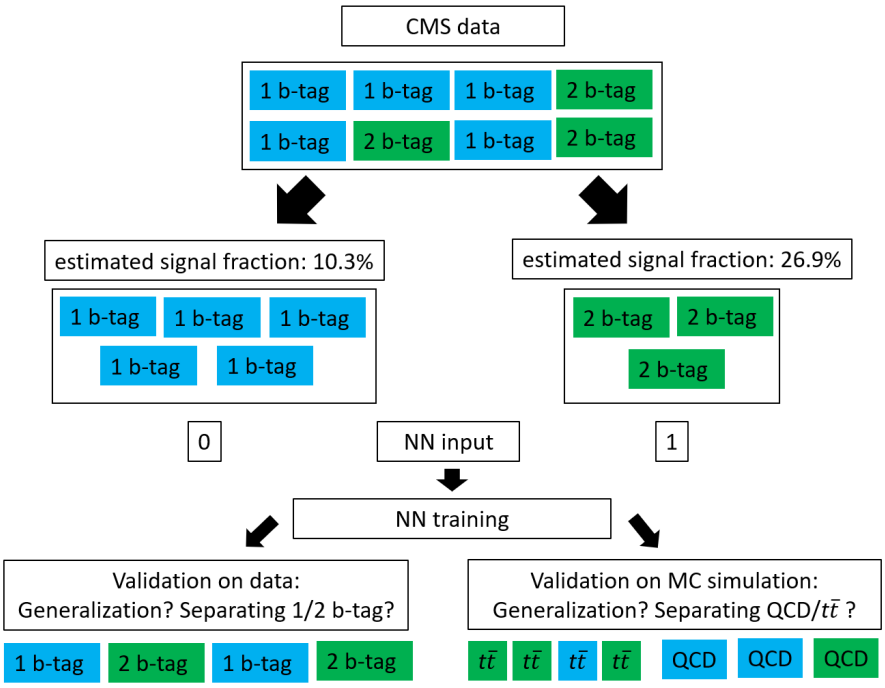


Figure 26: Flow chart for CWoLa on data. Data events are split into samples with either one or two b-tag(s). The NN is evaluated on data and MC. On data generalization and the performance of separating one b-tag events from two b-tag events is checked. On simulated events, separating power for QCD vs. $t\bar{t}$ is measured.

First an overview of the impact of each input feature is presented, called feature importance. This is only done now, because the training stability and also the received ROC-AUC values profit from the larger samples of data events compared to simulated events. The training and evaluation is repeated while one feature is left out at a time. The impact on the AUC-value is then measured. The results are displayed in fig. 27. Most features only have a low impact on the AUC value. Only leaving out m_t^{fit} and m_{W1}^{reco} are making the classification significantly worse. The only value which is not impairing the NN's performance is $\Delta R_{b\bar{b}}$. It seems leaving out this feature would increase the obtained AUC value, but ± 0 is still in the range covered by the uncertainty.

Since m_t^{fit} is an important feature and has a strong influence on the results obtained from the later introduced background estimation method, an additional NN is trained without it for comparison. In the following only the NN with m_t^{fit} as a feature is compared in detail to previous analysis. The results for the one without this feature are summarized in app. A.

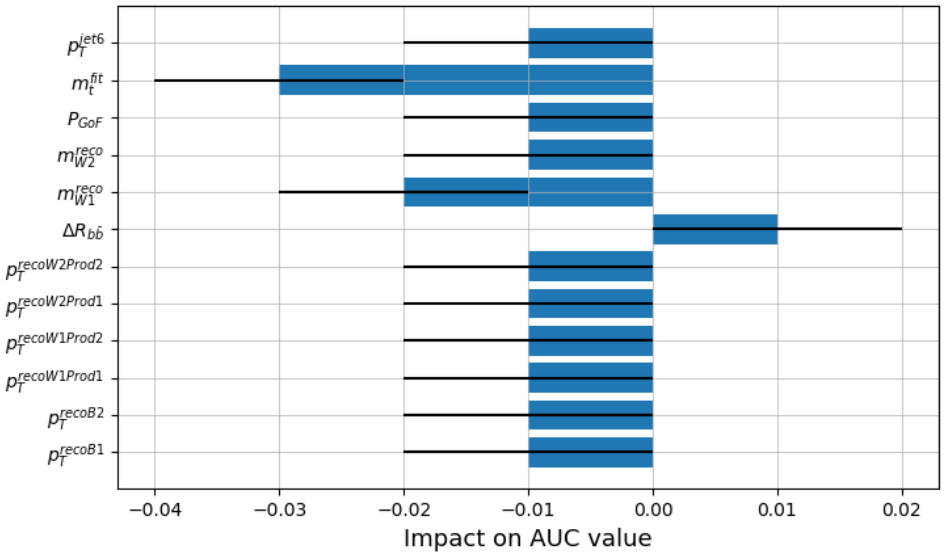
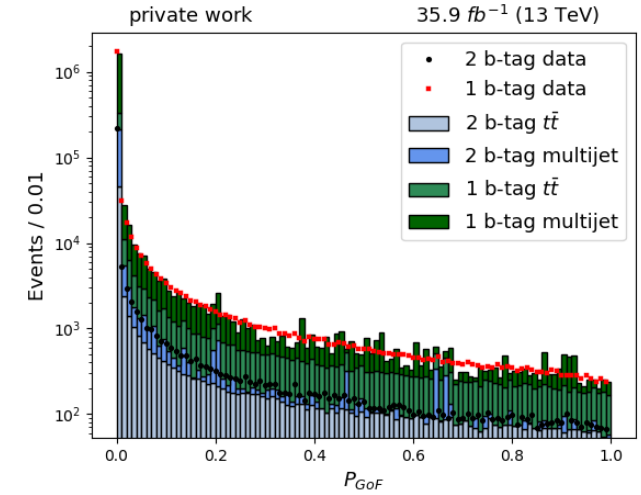


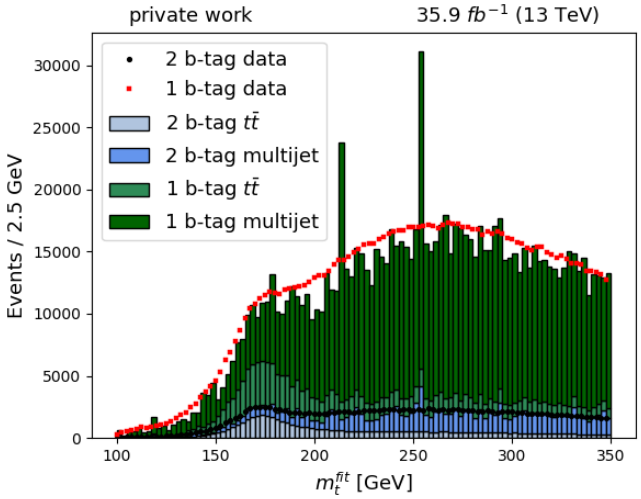
Figure 27: To measure the impact of an import feature, training and evaluation are done without one feature at a time. The importance of a feature is then measured via the difference of the AUC value compared to the default model (0.00), which uses all of the listed features. The black lines show the estimated 1σ error on the AUC values. The error bars all have the same length since rounding is done on the second decimal place and the maximum obtained error for all features is of ± 0.01, this value is estimated as an upper 1σ bound to the results.

Before training the NN, the composition of the input data is investigated. Displayed in fig. 28 are m_t^{fit}, P_{GoF} and m_W^{reco}. The distribution of the remaining input features are shown in fig. 29. Only the relevant range of the distributions are plotted.

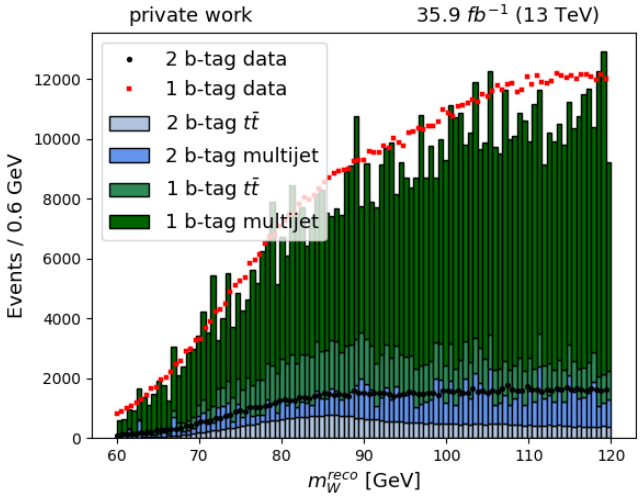
The red and black dots show the data distribution. Similar to the stacked histogram of simulated events, the red dots represent the stacked value of 1 and 2 b-tags events for reasons of presentation. The distribution is rather smooth in contrast to the one with simulated events, also shown in the figure. This is the result of all data events having an event weight of 1. However they show largely agreement between data and simulated events, despite using multiple permutations per event. The 'spikes' at 210 respectively 250 GeV in the m_t^{fit} distributions originate from multiple passing permutations of a single event with a weight of $\approx 10^4$. These high weights result from low statistics after selection efficiency. Especially for QCD the cross section in high H_T bins is low, as displayed in tab. 3.



(a)



(b)



(c)

Figure 28: Input distributions of m_t^{fit}, P_{GoF} and m_W^{reco} for the CWoLa method. Black and red dots are stacked data, whereas the stacked histogram bars are made of MC simulated events. All permutations passing the preselection cut flow are used.

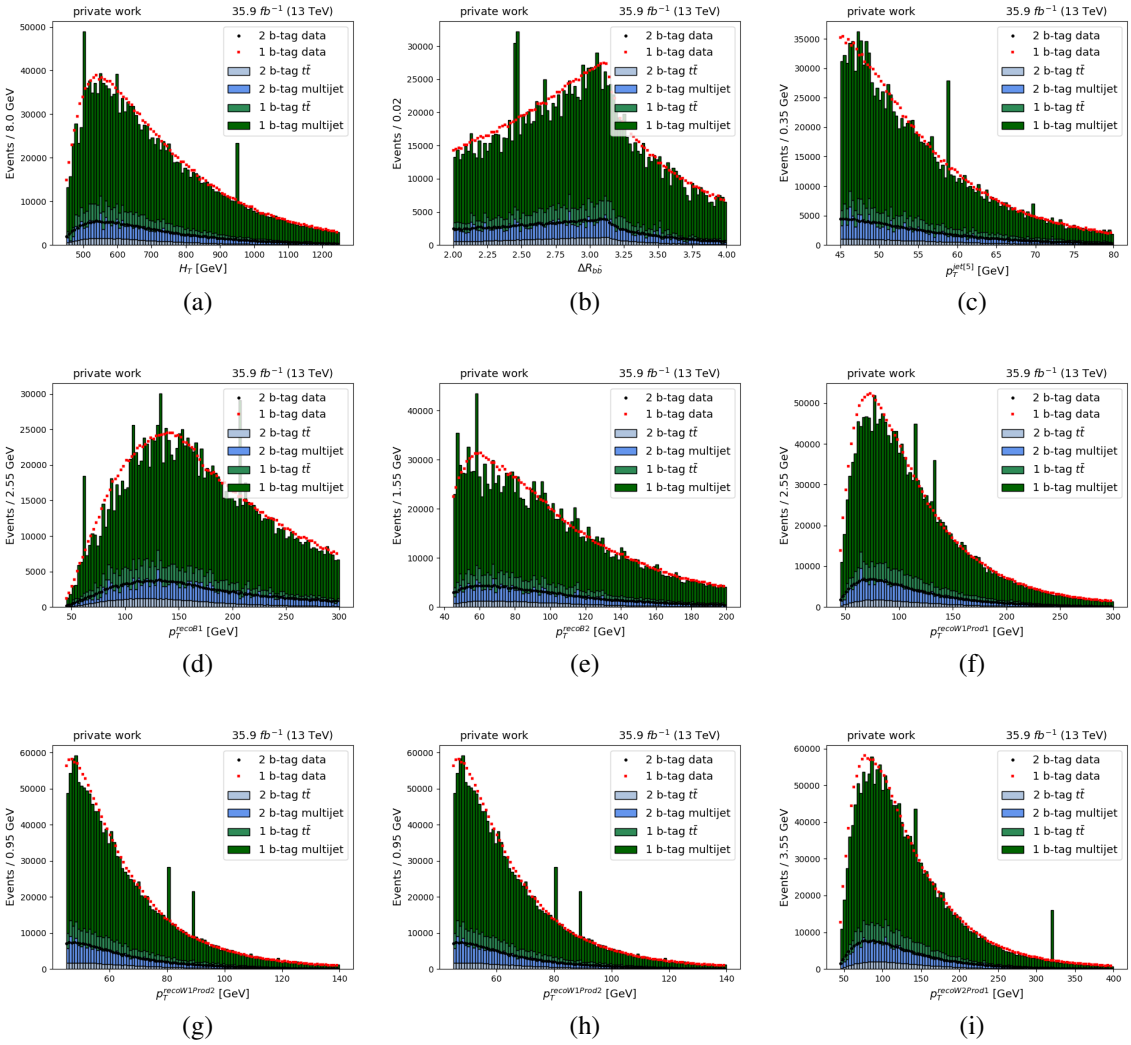


Figure 29: Input distributions for the CWoLa training method. Black and red dots are stacked data, whereas the stacked histogram bars are made of MC simulated events. All permutations passing the preselection cut flow are used.

The results of the training are presented in fig. 30. The AUC value (0.73) is significantly higher as when trained on simulated events (0.66), shown in fig. 25, but it does not reach the result of training on pure samples (0.76). This behaviour underlines the importance of having data with high statistic and equal weighting of each data point.

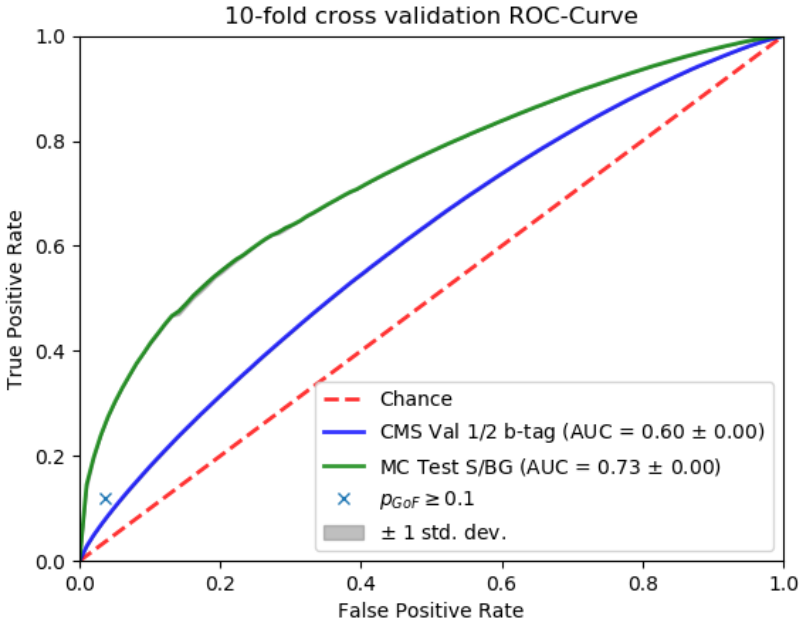


Figure 30: Results of CWoLa method on data. As test sample for $t\bar{t}$ classification, simulated events are used (green curve). The blue curve shows the validation on 1 and 2 b-tag samples. The blue cross marks a possible selection step P_{GoF}, regarding $t\bar{t}$ vs. QCD for comparing results.

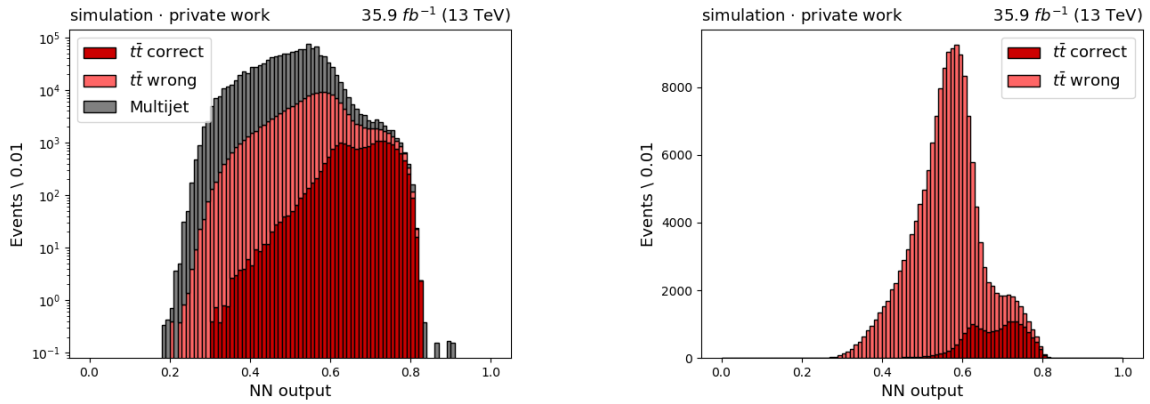


Figure 31: (a) shows the stacked output of the trained classifier for both $t\bar{t}$ and QCD sample. (b) shows the stacked output for the $t\bar{t}$ sample divided into correct and wrong permutations. The model prefers the correct permutations at higher cut values.

The trained NN separates $t\bar{t}$ from QCD multijets, but does not separate by the number of b-tags. The NN output applied on simulation after the preselection cut flow is shown in fig. 31. There the contributions of correct and wrong permutations are displayed. Additionally

fig. 32 shows the NN output by comparing the normalized output shape of simulated events, partitioned into 1 and 2 b-tag events. The separation of $t\bar{t}$ and QCD is observed in (c) and (d).

Though it seems the distributions in (a) and (b) are shifted by a small amount, respectively are not exactly equal. This is either observed because a separation of $t\bar{t}$ and QCD leads to a weak separation of 1 and 2 b-tag events, or because of an input feature, in particular $\Delta R_{b\bar{b}}$. While investigating feature importance, the validation curve on 1 and 2 b-tag data for the model, where $\Delta R_{b\bar{b}}$ was left out, decreased by $0.02(\pm 0.004)$, which was not observed for other features. This might indicate a correlation of $\Delta R_{b\bar{b}}$ and b-tagging efficiency.

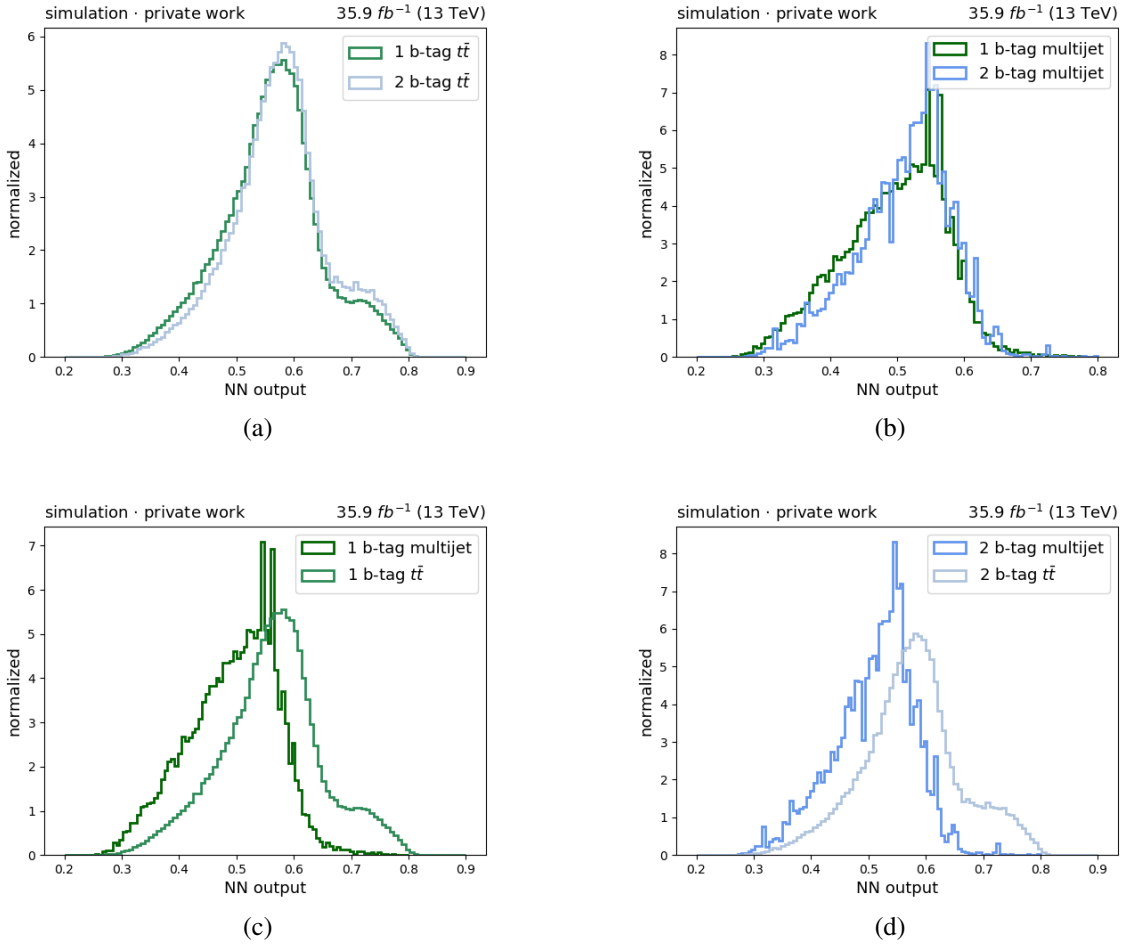


Figure 32: (a) and (b) show the NN output of the trained classifier on simulated events, for signal $t\bar{t}$ and QCD multijet each. The classifier is not able to separate by the number of b-tags. (c) and (d) show the NN output for 1 and 2 b-tag data each. A separation of $t\bar{t}$ from QCD multijet events is observed.

The output distribution is narrower than the one trained with full supervision, displayed in fig. 31. This seems to be a feature of the CWoLa method and might be related to minimizing validation loss in the training procedure on apparently not separable samples with 1 and 2 b-tag. Still it shows a quite similar output distribution as the NN which was trained using full supervision. The correctly matched $t\bar{t}$ events have two peaks. This is apparently due to using m_t^{fit} as an input feature, because the additionally trained NN (see ch.A) does not create such a

behaviour.

Selection method	data	$t\bar{t}$	purity	$t\bar{t}$ correct
$P_{GoF} > 0.1$	10799	8126	75.2%	50.7%
CWoLa NN > 0.722	10827	9016	83.3%	67.6%
CWoLa NN > 0.680	22523	16868	74.9%	59.7%
CWoLa NN > 0.620	69071	37654	54.5%	40.6%

Table 10: Comparison between remaining events of CWoLa NN cut and the results shown in tab. 8. The fraction of $t\bar{t}$ correct is received via $t\bar{t}_{correct}$ / $(t\bar{t}_{correct} + t\bar{t}_{wrong})$.

The results for different cut values are compared to tab. 8. The values 0.722, 0.680 and 0.620 are chosen because a cut value of 0.722 leads to the same amount of remaining events, but at an increased purity of 83.3% instead of 75.2%. At 0.680 the purity is about the same value as in previous analyses, but the total number of remaining events is increased from 10799 to 22523. The value 0.620 is chosen at the first maximum of correctly matched $t\bar{t}$ events according to fig. 31. The later introduced background estimation would also justify this loser cut. Thus more signal events are collected, which has an impact on systematic uncertainties of the top mass measurement. Purities and fractions of correctly matched $t\bar{t}$ events are summarized in tab. 10.

First, the cut at 0.722 is compared to the selection made in tab. 8. In fig. 33 the distributions for m_t^{fit}, m_W^{reco} and P_{GoF} are shown. The selection with the NN results in sharper peaks for both fitted top quark mass and W boson mass. Furthermore nearly no tail at higher masses is observed for the top quark mass distribution.

The shape of P_{GoF} shows an unusual behaviour. While the distribution of wrongly matched events remains rather flat between 0.2 and 1, the correctly matched events have a low point at 0.45. This means that the NN is not entirely relying on this feature, which is another justification for using more information and a NN to select $t\bar{t}$ events in data. This behaviour is related to using m_t^{fit} as a feature. The NN without it (see app.A) shows a rather smooth distribution of P_{GoF} on the chosen cut value.

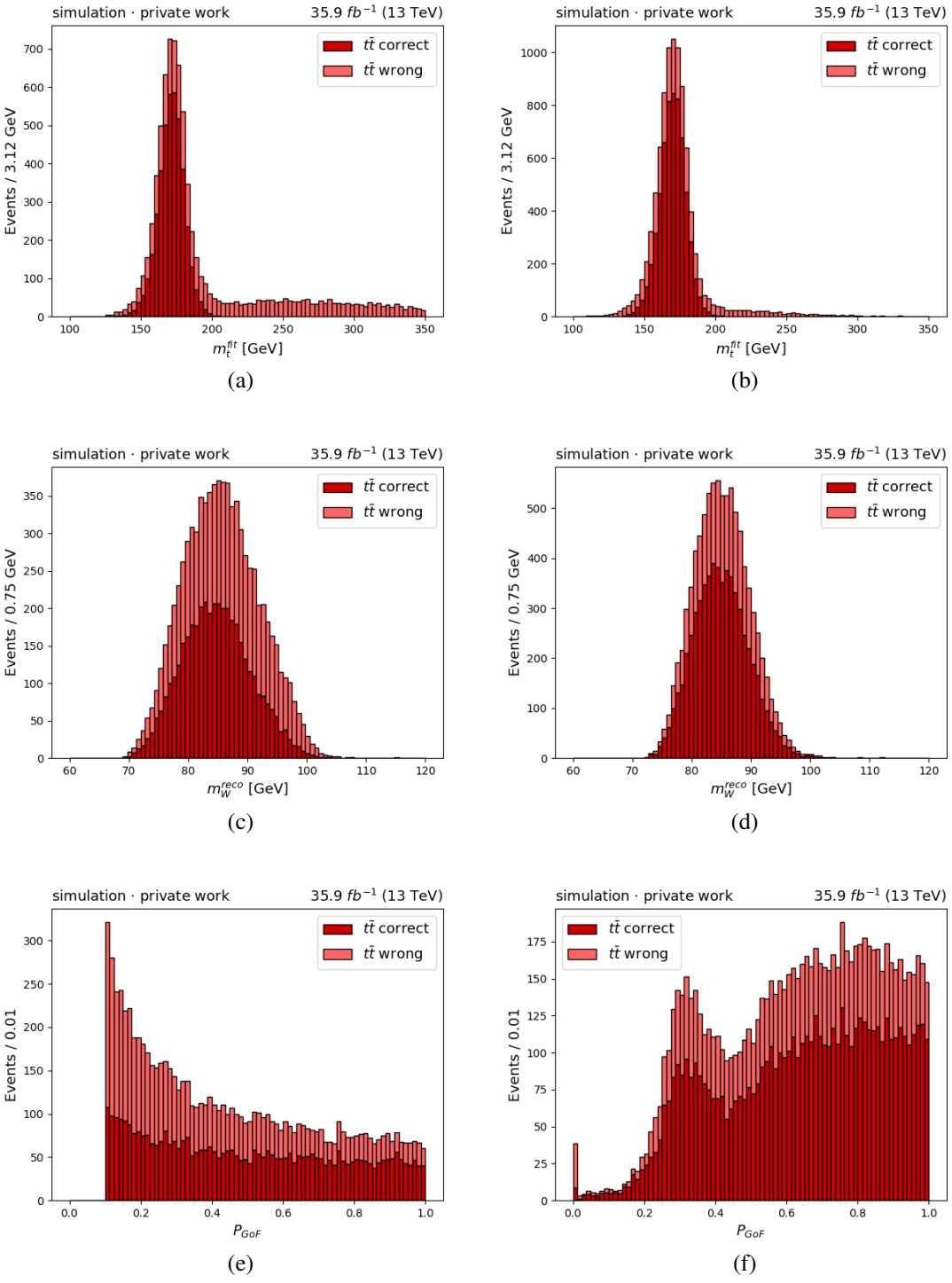
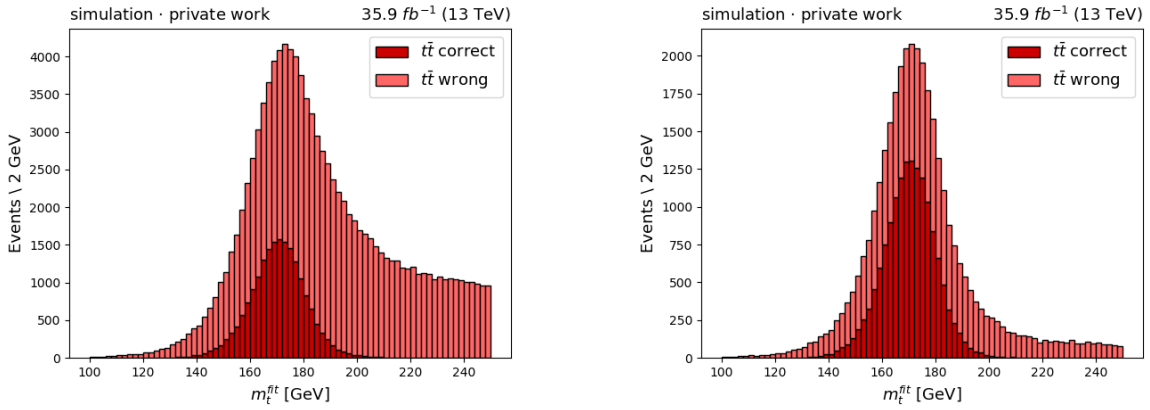


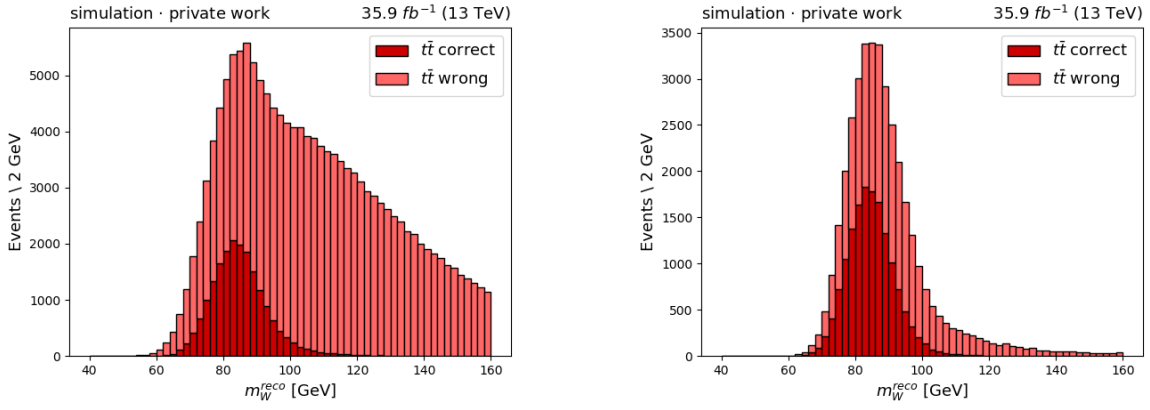
Figure 33: On the left, the distribution after $P_{GoF} > 0.1$ of the cut flow tab. 8 is displayed, on the right the NN output > 0.722 after the preselection of tab. 5 is displayed. The average of both W boson masses is used for m_W^{reco}.

From now on, the cut value 0.620 is investigated further. The application of the NN after the preselection is shown in figs. 34 and 35. Both distributions m_t^{fit} and m_W^{reco} show a long tail towards higher masses. This tail is caused by wrongly matched events. The majority of correctly matched permutations remain after the cut, while the events towards higher masses are sorted out.



(a) (b)

Figure 34: Distributions of m_t^{fit}: (a) shows a selection without a NN cut while (b) shows one with a NN cut of 0.620. Only the best permutation of each event according to the NN is used.



(a) (b)

Figure 35: Distributions of m_W^{reco}: (a) shows a selection without a NN cut while (b) shows a NN cut of 0.620. Only the best permutation of each event according to the NN is used.

Many correctly matched events were cut out in previous analyses because of requiring two b-tags. The CWoLa method used here relies on the number of b-tagged jets for splitting data. That is why correctly matched $t\bar{t}$ events with one b-tag can also be recognized by the NN. Since the majority of events only has one b-tag, the gain lies in these previously not even considered events.

To observe the impact of 1 b-tag events, the distributions in relation to their number of b-tags are displayed in fig. 36. Therefore the cut value 0.620 is compared to the one at 0.722, which

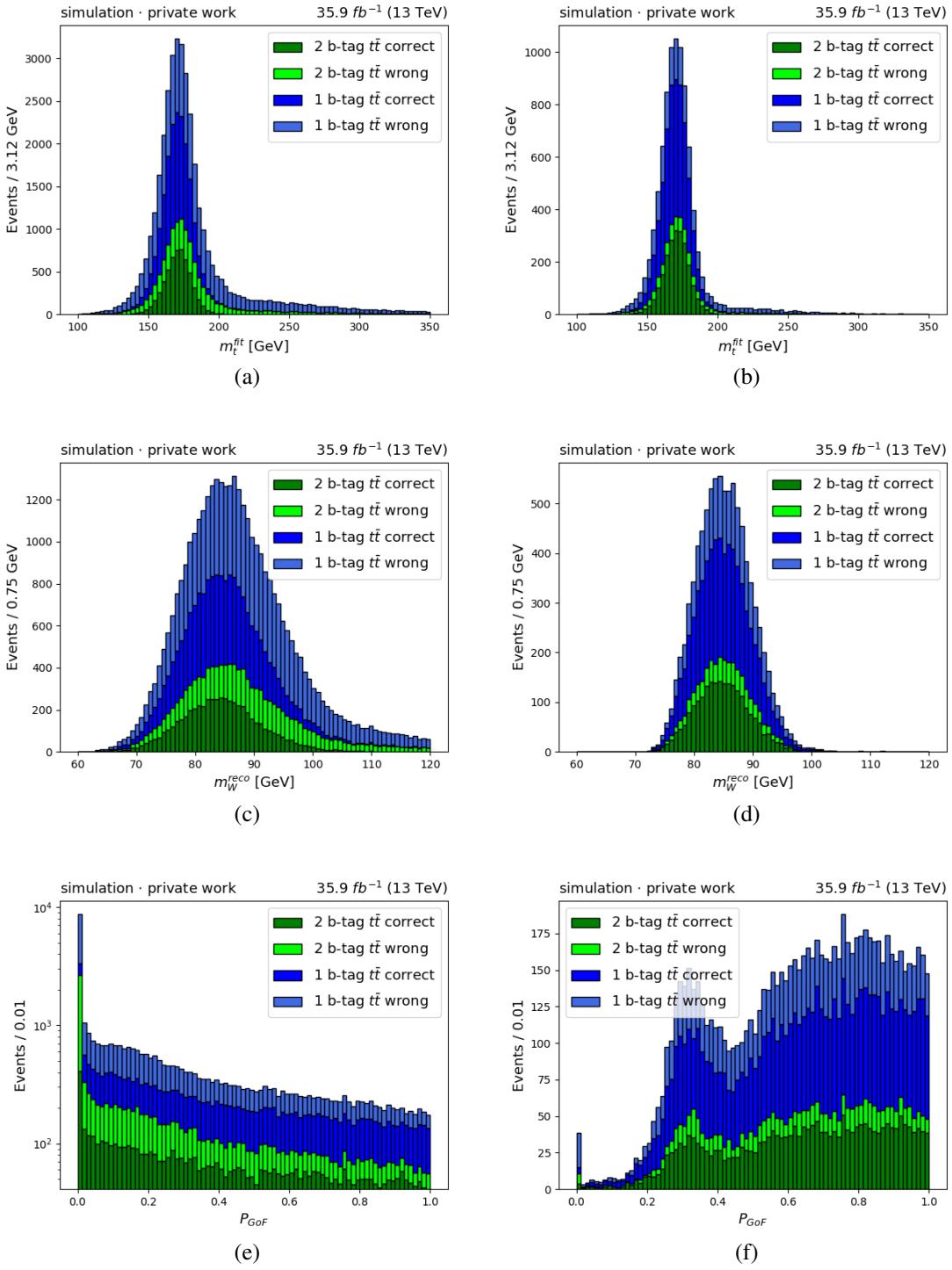


Figure 36: The left side displays the selection with a NN cut > 0.620, applied after the preselection, while the right side displays one with > 0.722. The average of both W boson masses is used for m_W^{reco}. The left side shows a behaviour similar to the selection steps presented in tab. 8. Only the strong peak at low P_{GoF} values is different. Still there are a lot of correctly matched events with only one b-tag, which are recognized by the NN.

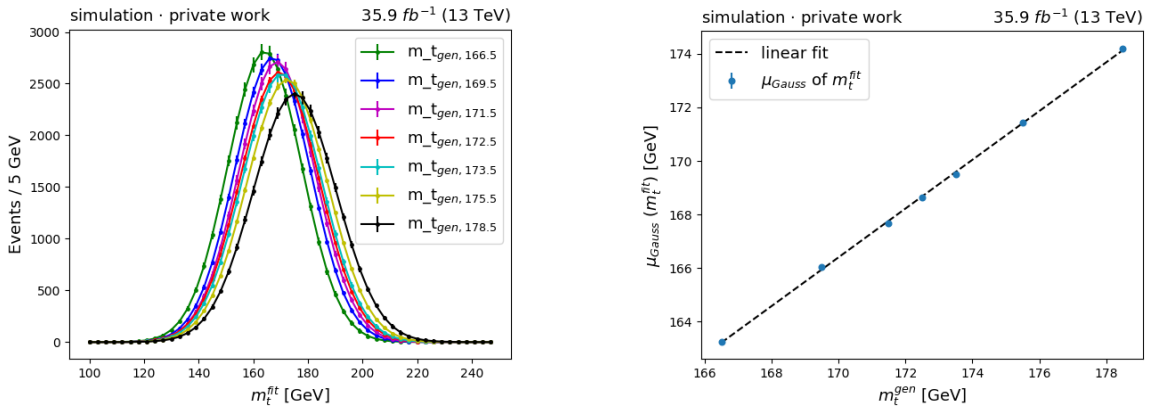
consists of around as many events as the selection in [42], but here the majority of events only have one b-tagged jet. Similar, the fraction of events with one b-tag remains also higher at tighter cuts. Two things are of particular interest here. First, the shape of the distributions of 1

and 2 b-tag correct/wrong resemble each other, even for different cut values. Second, the gain on correctly matched events with only one b-tag contributes the most events at lower cuts to the selection. This is expected, because most of correctly matched $t\bar{t}$ events will only have one b-tag due to only 50% b-tagging efficiency. The ratios resembles the fractions of 1 to 2 b-tag events in the input samples, which is estimated as 2:1[11] Most of events selected by a cut of 0.620 have a P_{GoF} value near zero. The situation for the different cut values are displayed in tab. 11.

cut value	1 b-tag c.m.	1 b-tag w.m.	2 b-tag c.m.	2 b-tag w.m.
CWoLa NN > 0.722	3859	2144	2237	763
CWoLa NN > 0.680	6371	4944	3697	1856
CWoLa NN > 0.620	9701	16113	5591	6249

Table 11: Fractions of correctly matched (c.m) and wrong matched (w.m) events for 1 and 2 b-tag.

Top mass sensitivity



(a) (b)

Figure 37: (a) shows the output distributions of m_t^{fit} for different generated top masses. A gaussian distribution is assumed. In (b) the dependency of the received top mass peak in relation to the generated top mass is displayed. The fitted top mass values are the mean values of the Gaussian fit, the errors on the mean are at a scale below < 0.1 GeV. The equation of the linear regression is provided in eq. 26

One might suspect, that the NN is selecting events according to the fitted top mass. To ensure that it would not just seek events which are distributed around the top mass peak of the training data, the NN is applied to other mass samples. Fig. 37 proofs, that each generated $t\bar{t}$ mass sample peaks at a different mass. The applied linear fit is described through

$$m_t^{fit} = 11.7(\pm 1.4)\ \text{GeV} + 0.910(\pm 0.008) \cdot m_t^{gen} \tag{26}$$

[11] 2 b-tags are assigned with a probability of 0.25 as a subset of the set with one b-tag events while at least one b-tag for a b-jet has the tagging probability of 0.75.

If all samples would peak at the same mass after applying the NN, it would just have learned a single mass value and not a $t\bar{t}$ decay pattern.

5.6. Background estimation

The background estimation follows the one made in [42]. Because of the chance of non-$t\bar{t}$ multijets passing the NN cut, a background estimation is needed. A non-$t\bar{t}$ event passing the criteria can be the result of mis-tagging jets as b-jets or jets originating from gluon splitting into $b\bar{b}$ pairs. It is assumed that light flavor jets can mimic these events by combinatorial chance, such as displayed in fig. 6.

Here, only the background shape is estimated, the normalization is a free parameter. This is done by comparing zero b-tag events (referred to as prediction) with simulated QCD multijet events (direct simulation), which are displayed in tab. 3. Zero b-tags means in this case, that events have no jets with a b-tag value > 0.2. Also a different, pre-scaled trigger, HLT_PFHT450_ixJet40, is used.

The preselection cut flow is applied to the zero b-tag events, too, but a selection step with for example at least one b-tag is not possible. Therefore only the selection steps until $\Delta R_{b\bar{b}} > 2.0$ are made. Eventually, the NN cut is used after this shortened preselection. Signal $t\bar{t}$ events with zero b-tags are neglected because their fraction on all zero b-tag events is only 0.34%. In [42] it is shown, that this ansatz fails for $\Delta R_{b\bar{b}} < 2.0$.

After applying the preselection and the NN also on zero b-tag data events (prediction from data), they are compared to prediction and direct simulation.

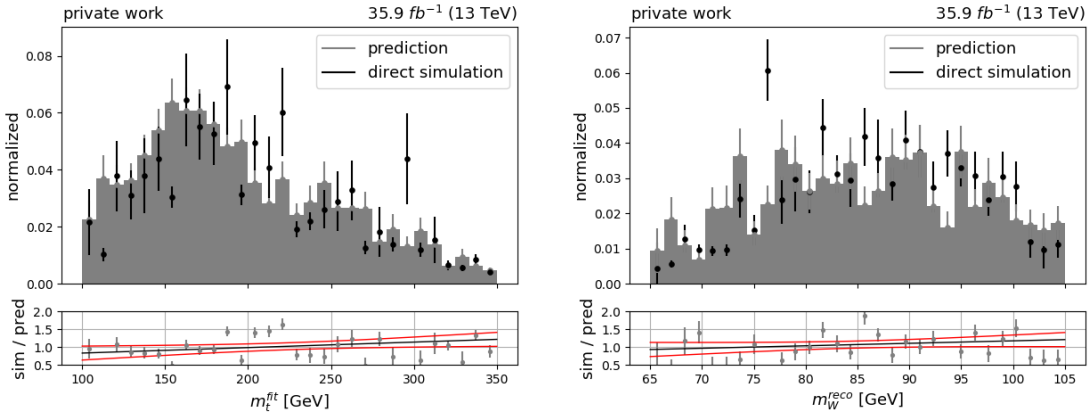
The scaling is done by normalizing the distribution to the difference of obtained signal events from $t\bar{t}$ MC and data. The prediction from data contains of around the same number of events as expected from direct QCD simulation for the NN cut > 0.62.

In fig. 38 the background predictions are shown for the fitted top mass and the reconstructed W boson mass. The histograms are scaled to unity. The lower frame shows the ratio between direct simulation and prediction from zero b-tag QCD events. The errors are calculated as Poissonian error bars.

For m_t^{fit} the slope is 0.0015 ± 0.0014 GeV^{-1} and for m_W^{reco} it is -0.007 $\pm$ 0.009 which is 1.07σ and 0.78σ from zero and covered by the uncertainties. Therefore no correction is applied. The background estimation tests for all other used input features and for H_T, which is used in the selection steps is displayed in app.B. The distributions of prediction and direct simulation show reasonable compatibility.

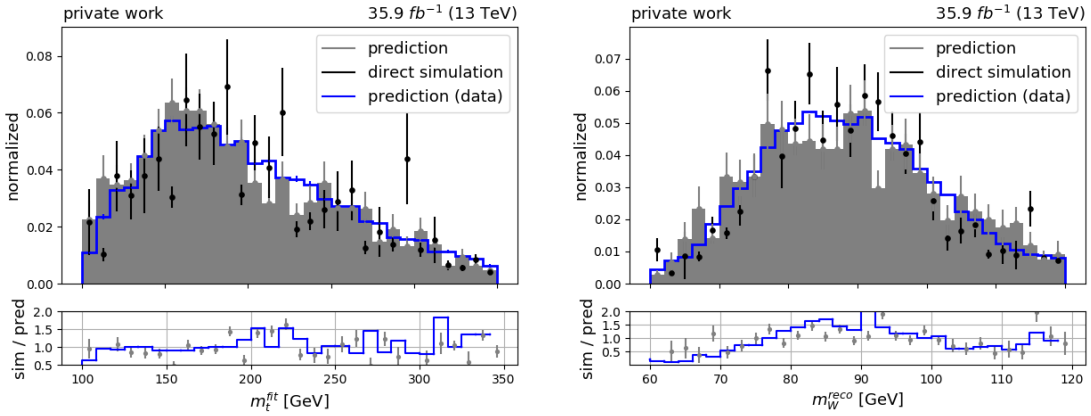
Prediction from simulation and direct simulation is compared to prediction from data. This is displayed in fig. 39. The shape of the background from data resembles the ones modeled via simulation in good agreement. Whereas the reconstructed W boson mass shows some gaps in the MC simulated part. This is most likely a problem with the low statistics. However the prediction from data shows a softer spectrum and an overall good agreement with simulation.

Fig. 40 shows the final distribution of both top quark mass and W boson mass. The distri-



(a) (b)

Figure 38: Test of the background prediction for m_t^{fit} and m_W^{reco}. The grey bars display the distribution obtained from direct QCD multijet simulation, while the black dots represent the distribution for QCD multijet events with zero b-tags. In the ratio plot the black line is the best least squares fit, whereas the red lines mark the 1σ area for slope and intercept.



(a) (b)

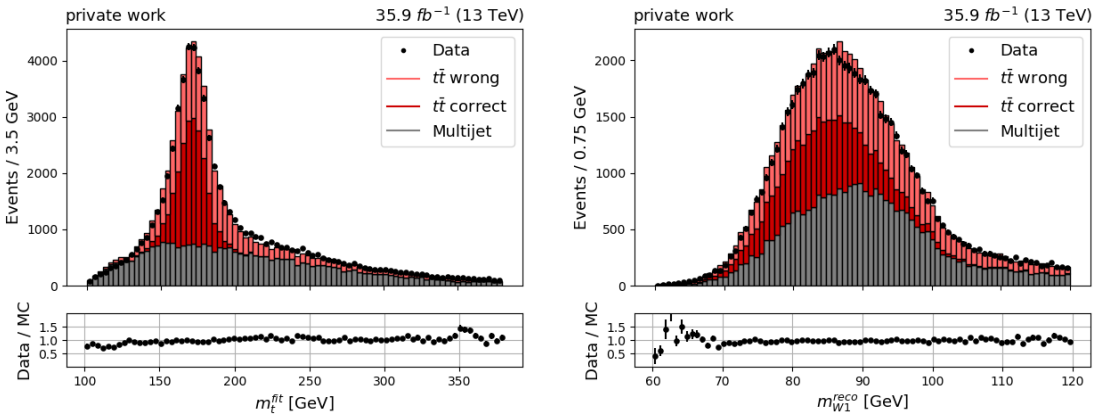
Figure 39: Test of the background prediction for m_t^{fit} and m_W^{reco}. The grey bars display the distribution obtained from direct QCD multijet simulation, while the black dots represent the distribution for QCD multijet events with zero b-tags. The blue steps show the background prediction from data.

butions of the other used features are displayed in app. B. The multijet background estimation contributes about 45% to the selection. For the top mass extraction, background and top mass peaking at nearly the same value is unpleasant. The modeling shows an overall good agreement of data and simulation.

Background estimation without m_t^{fit} as a feature of the NN

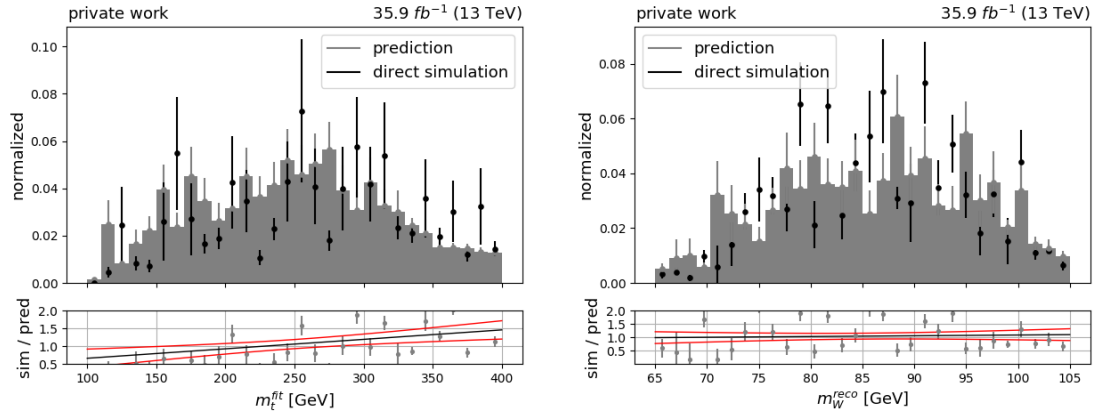
To avoid background and fitted top mass peaking at the same value, another NN is trained without m_t^{fit} as an input feature. Hyperparameters and training procedure are not changed.

The chosen cut value of 0.64 of this alternative model leads to 34850 selected events in data



(a) (b)

Figure 40: Final distributions of data compared to signal $t\bar{t}$ and the multijet background estimate. The average for both W boson masses is used.

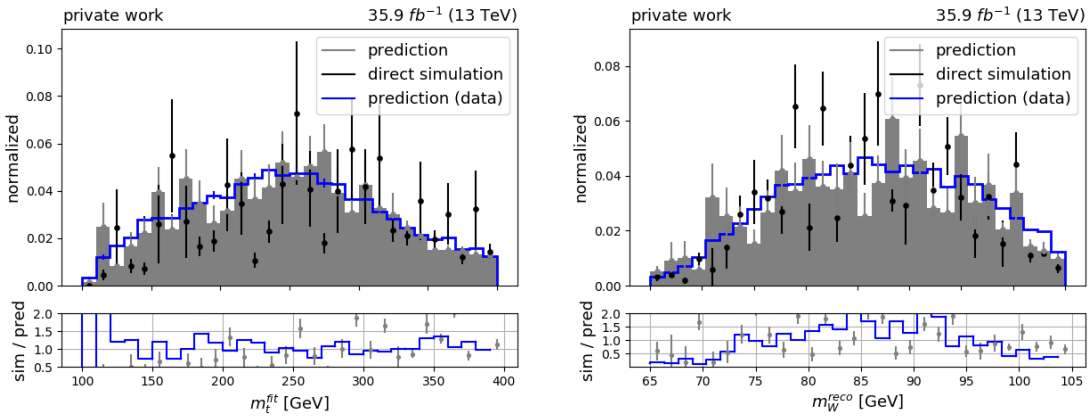


(a) (b)

Figure 41: Final distributions of data compared to signal $t\bar{t}$ and the multijet background estimate for a binary classifier which was trained without m_t^{fit} as input feature. The average for both W boson masses is used.

and 19626 $t\bar{t}$ events, which results in a purity of 56.3%. This value is chosen, because the gained purity is similar to the one of the previous discussed model, so results can be compared. The new distributions for m_t^{fit} and m_W^{reco} are shown in fig. 41. The slope of the linear fit on m_t^{fit} is 0.0027 ± 0.0015 and for m_W^{reco} it is 0.003 ± 0.010. The background estimation for the fitted top mass is significantly worse (1.8σ) than the one received with the other NN model whereas the slope of m_W^{reco} is covered by the uncertainty. This effect is mainly resulting from having lower statistics compared to the selection from the previous discussed model. Again, prediction and direct simulation is compared to data. This is displayed in fig. 42

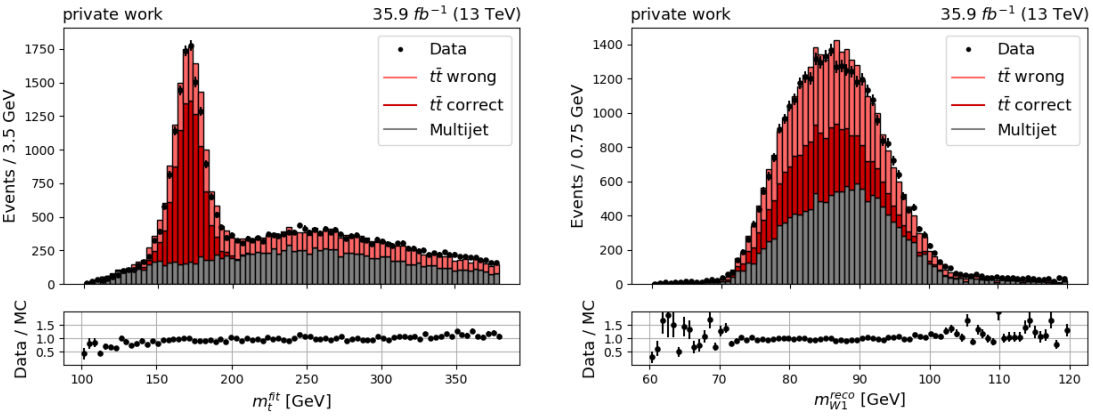
The final distribution for the fitted top quark mass and the reconstructed W boson mass are shown on fig. 43. The background estimation now peaks at a value of 250 GeV, which is far away of the fitted top mass peak. This leads to an better description of the data distribution in the range of the fitted top mass peak when using the background estimation as a template.



(a) (b)

Figure 42: Final distributions of data compared to signal $t\bar{t}$ and the multijet background estimate for a binary classifier which was trained without m_t^{fit} as input feature. The average for both W boson masses is used.

The distributions for other input features are shown in app. B. However the total number of remaining events is significantly lower at the same purity compared to the trained classifier with m_t^{fit} as an input feature. The decision, if this feature should be used should be made by looking at the systematic errors on the top quark mass measurement. An outlook is given in the next chapter.



(a) (b)

Figure 43: Final distributions of data compared to signal $t\bar{t}$ and the multijet background estimate. The average for both W boson masses is used.

6. Summary and Outlook

As described in ch. 5 the selection of $t\bar{t}$ pairs in multi-jet events is realised through a data-driven approach. Therefore the CWoLa training method was used. CWoLa relies on learning from samples with different signal fractions. To separate data with unknown labels (signal/background), b-tagging information is used. In previous analyses (see [42, 39]) only events with at least 2 b-tags are used, because the probability of these events stemming from $t\bar{t}$ decays is higher than the one from events with 1 b-tag. The idea of the data-driven approach is to make use of this knowledge and to simultaneously take 1 b-tag events into account.

It is shown that the number of selected signal events can be increased from 75.2% to 83.3% for a similar number of remaining data events. Also at a similar purity, more than two times of remaining data events can be achieved. Additionally the fraction of correct permutations of $t\bar{t}$ events is increased. It is remarkable, that using m_t^{fit} as the most important feature is not affecting the selections on simulated mass samples with different generated masses. Finding more suitable input features or a possible improvement of the NN hyperparameters might increase the selection efficiency further.

The modelling of the background is more demanding than in previous selections. A single feature, such as m_t^{fit} has a strong influence on the background estimation. Also a bias towards higher masses is observed for both the model with and the one without fitted top mass as an input feature. Further development of this method may be needed in particular when using a NN.

The systematic effects on the top quark mass mesurement are included as nuisances in a binned maximum-likelihood fit to the m_t^{fit} and m_W^{reco} distributions. The impact of the nuisances on the mass measurement are evaluated from pseudoexperiments using the default simulation and the data-driven background estimate. The impacts are shown for both the NN with the fitted top mass (cut value 0.620) and without it (cut value 0.64) as a feature and are compared to the ones from [42]. A slight improvement from 0.71 GeV to 0.69 GeV (NN with m_t^{fit}) and 0.67 GeV (NN without m_t^{fit}) is observed. The results are received through a correspondence with H. Stadie.

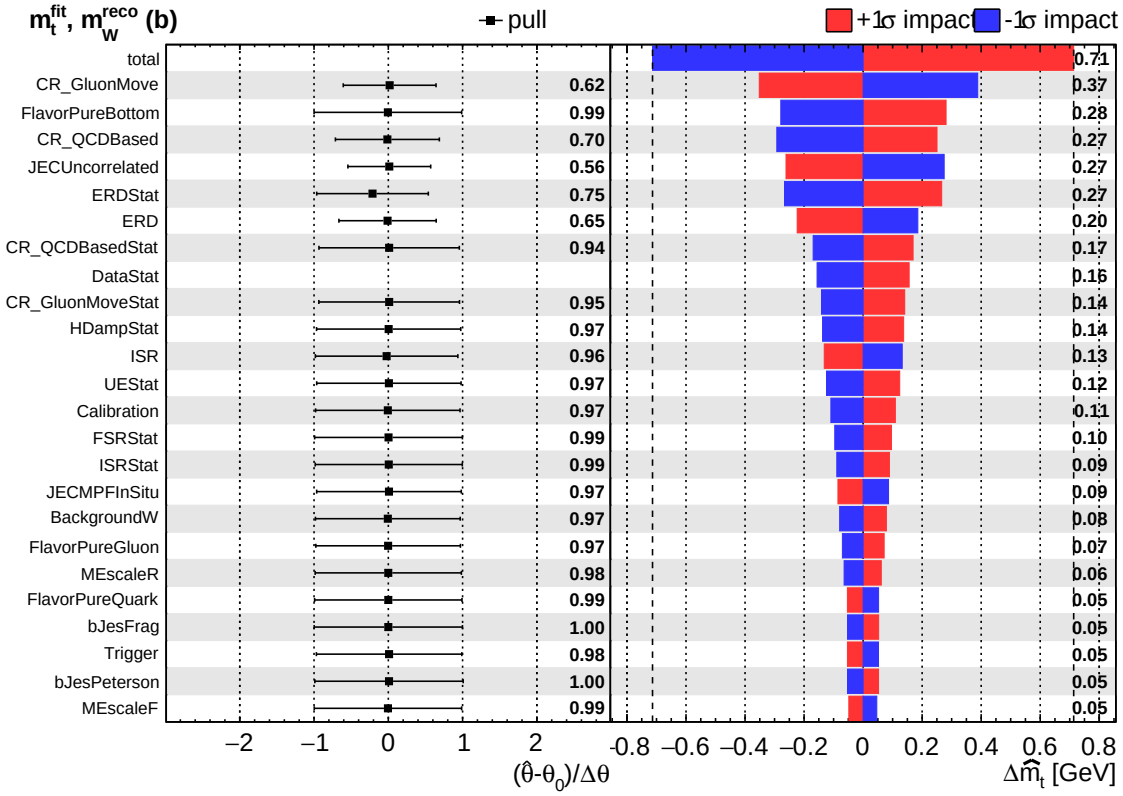


Figure 44: The systematic uncertainties of the selection from tab. 8 are shown with their impact.

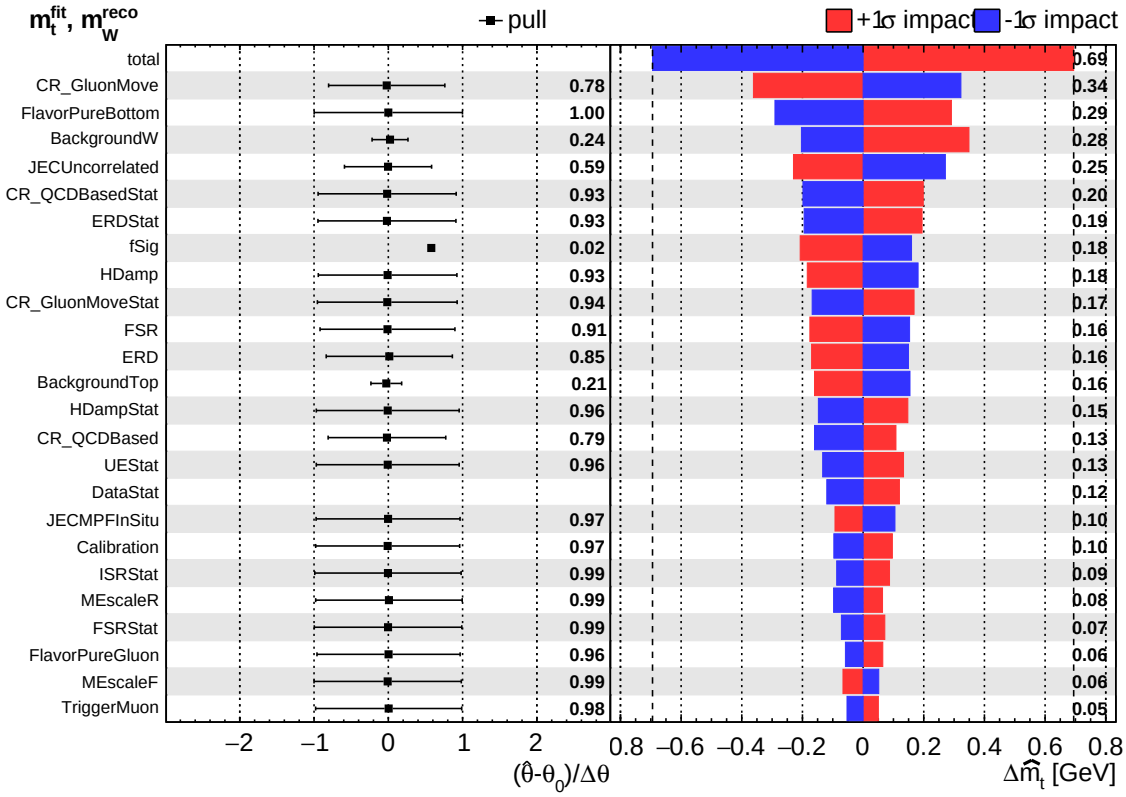


Figure 45: Result for the NN with m_t^{fit}. The systematic uncertainties are shown with their impact at a NN cut > 0.620.

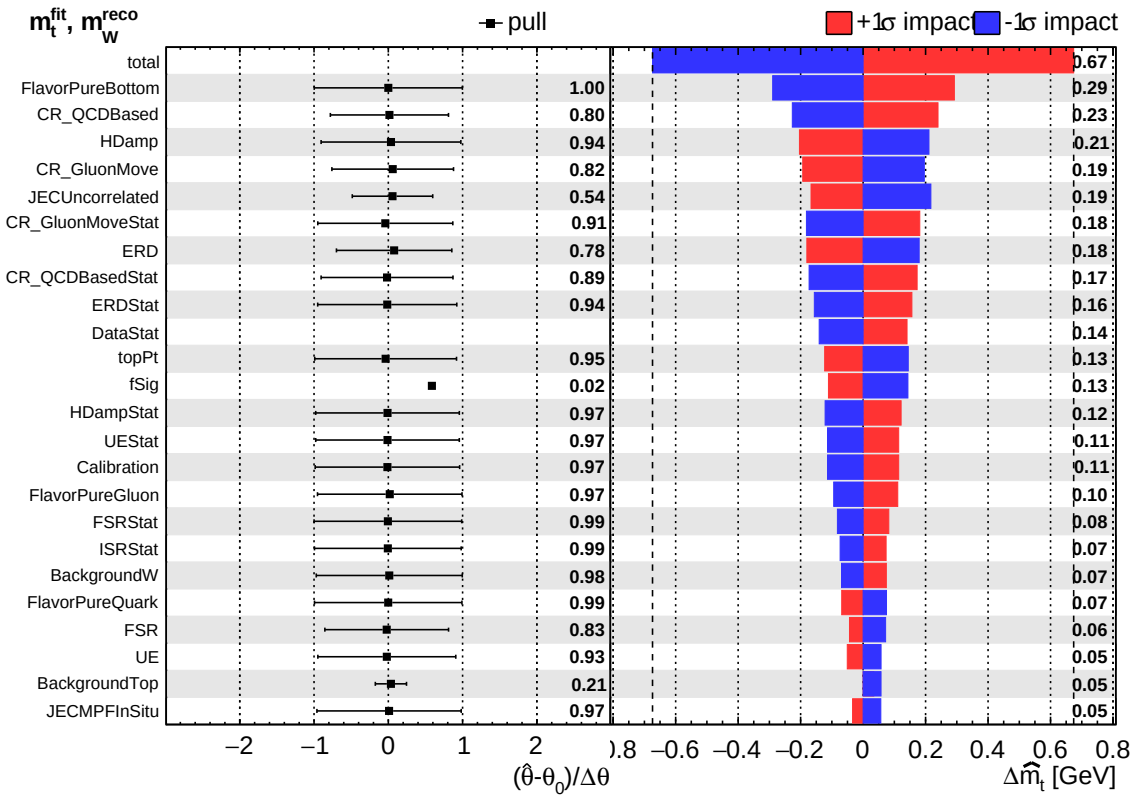


Figure 46: Result for the NN without m_t^{fit}. The systematic uncertainties are shown with their impact at a NN cut > 0.640.

7. Appendix

A. NN without fittet top mas as input feature

ROC-AUC value and output distributions

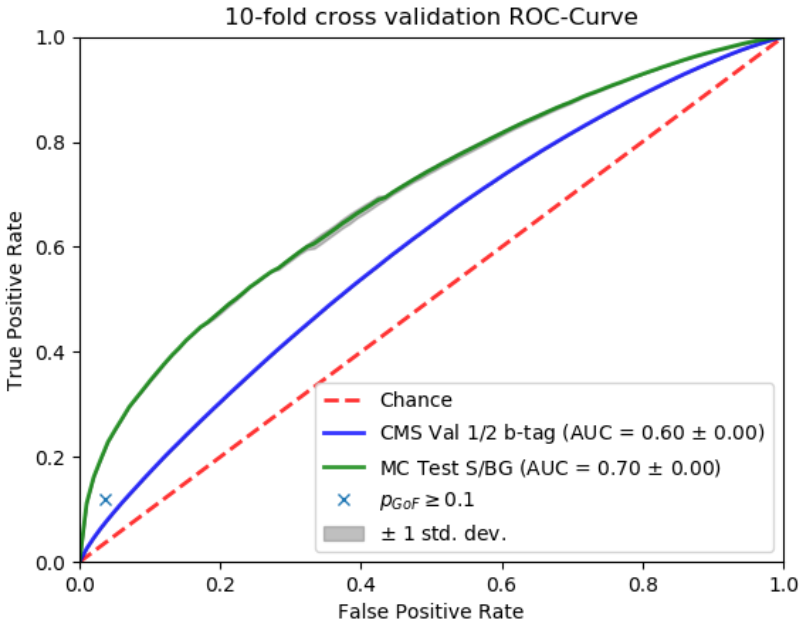
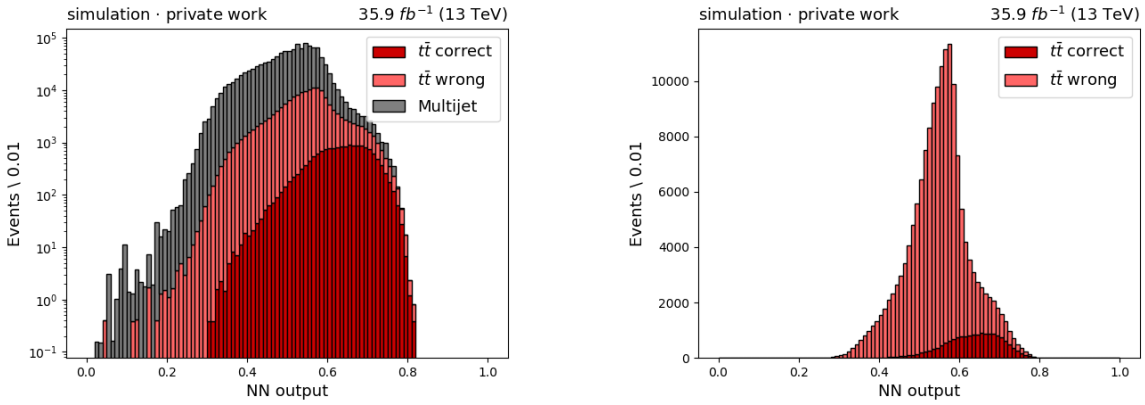


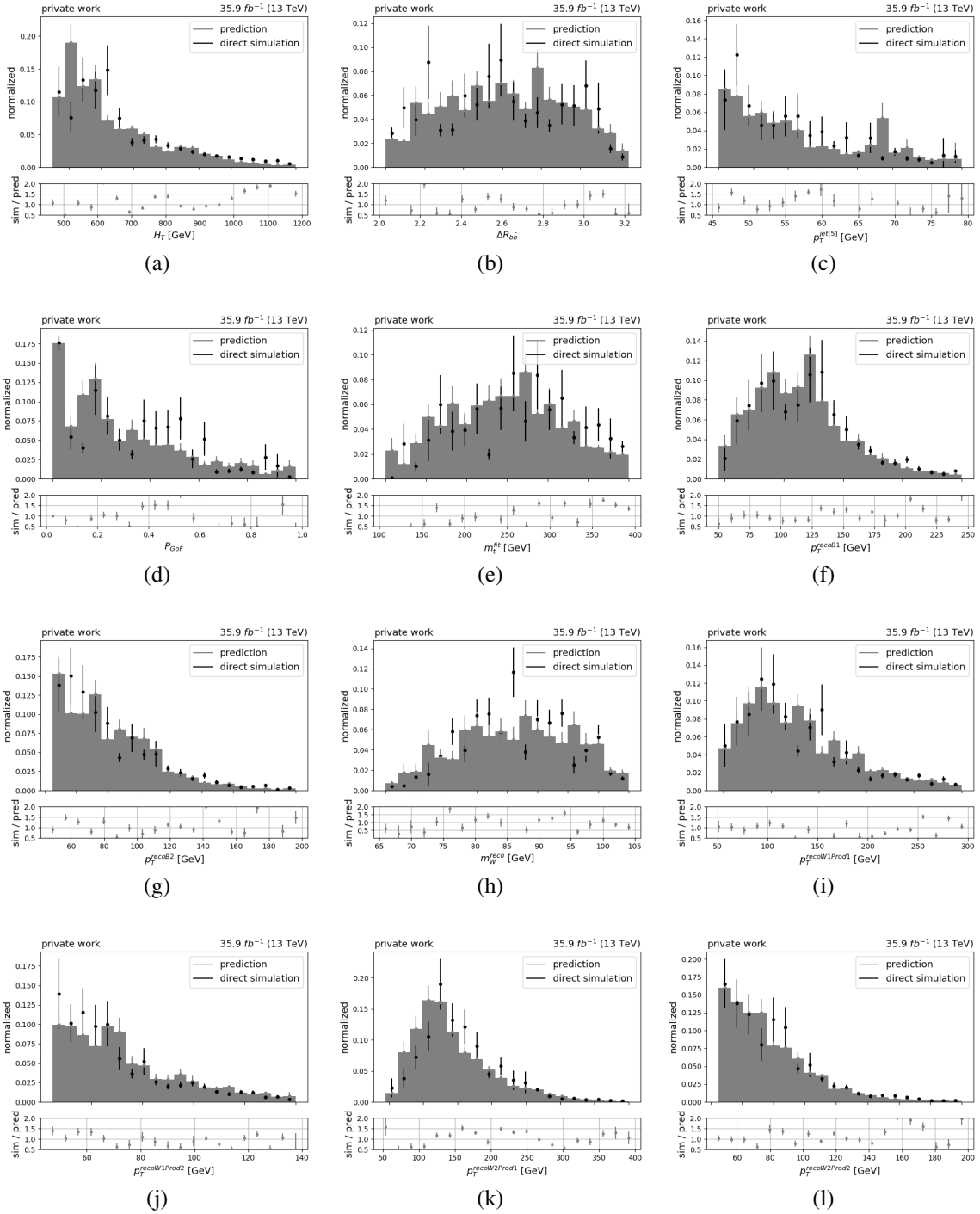
Figure 47: ROC-AUC value of the NN without using m_t^{fit} as an input feature.



(a) (b)

Figure 48: (a) shows the prediction distribution of the trained classifier used on both $t\bar{t}$ and QCD sample. (b) shows the prediction distribution for the $t\bar{t}$ sample, divided into correct and wrong permutations. The model rates correct permutations higher than wrong permutations on average.

Background estimation validation



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Figure 49: Background estimation validation for all input features of the NN.

Background prediction

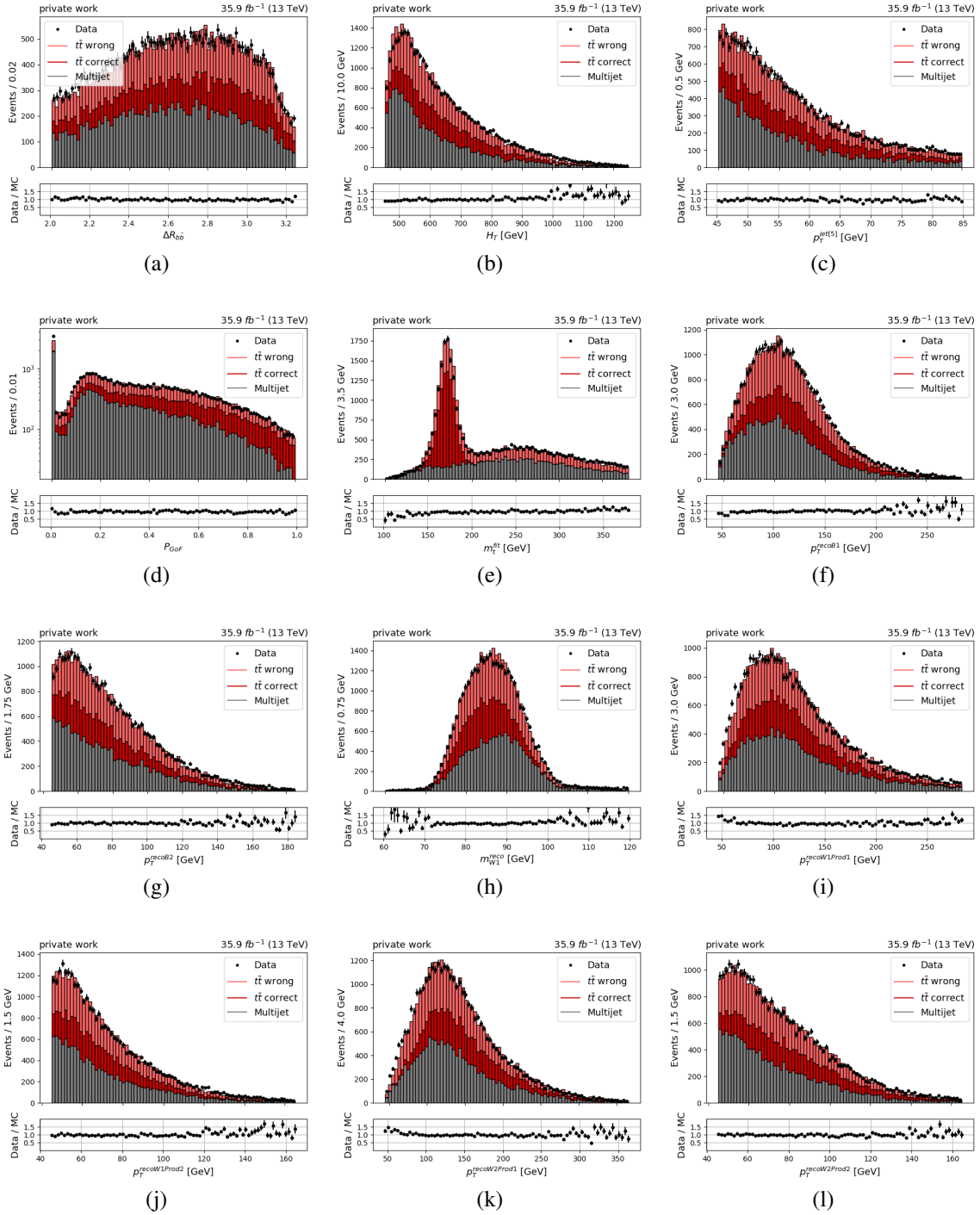


Figure 50: Background prediction for all input features and H_T of the NN at a cut value of 0.64.

Feature distributions

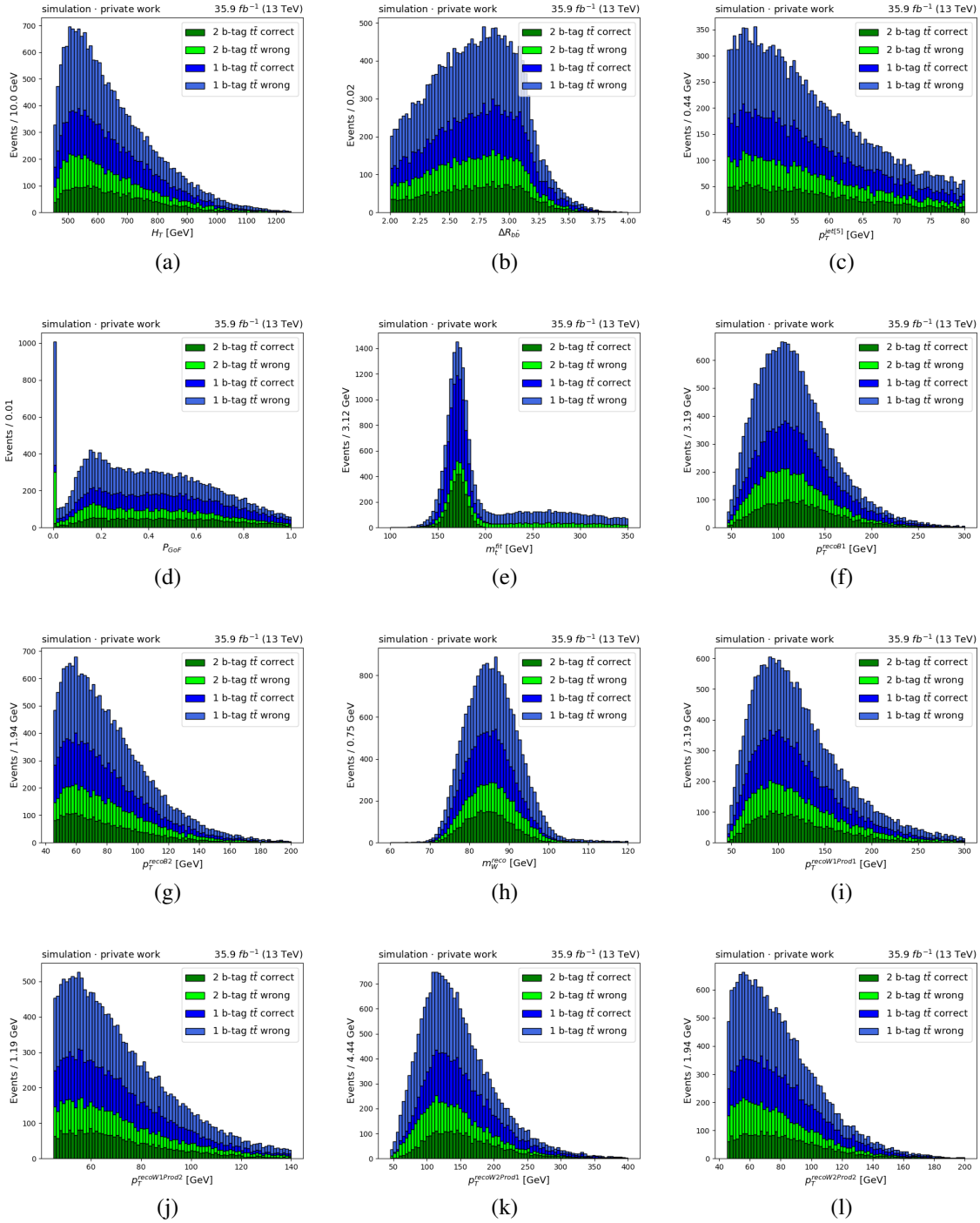
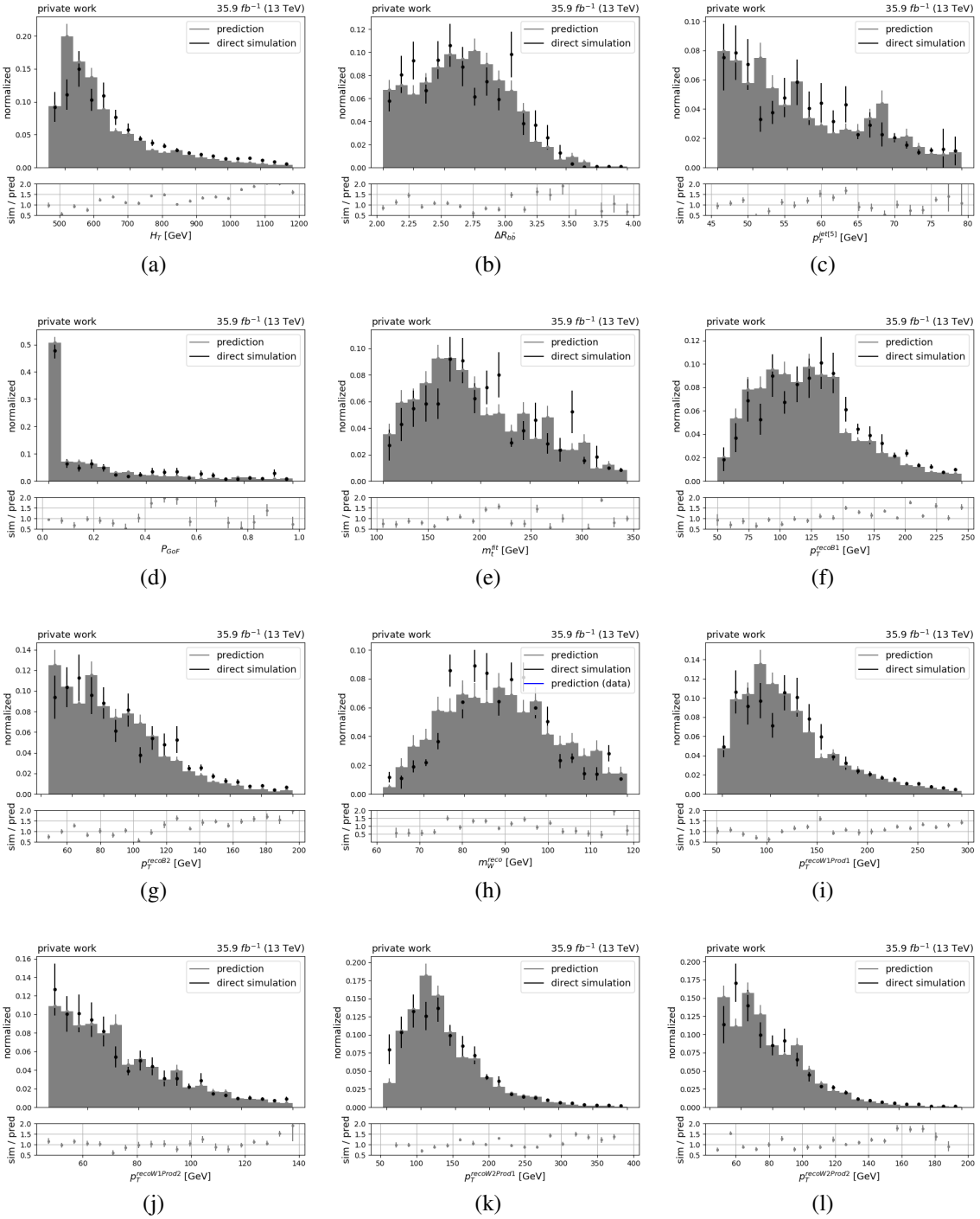


Figure 51: Feature distribution for NN cut = 0.64.

B. NN with fitted top mass as a feature

Background estimation validation



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Figure 52: Background estimation validation for all input features of the NN.

Feature distribution

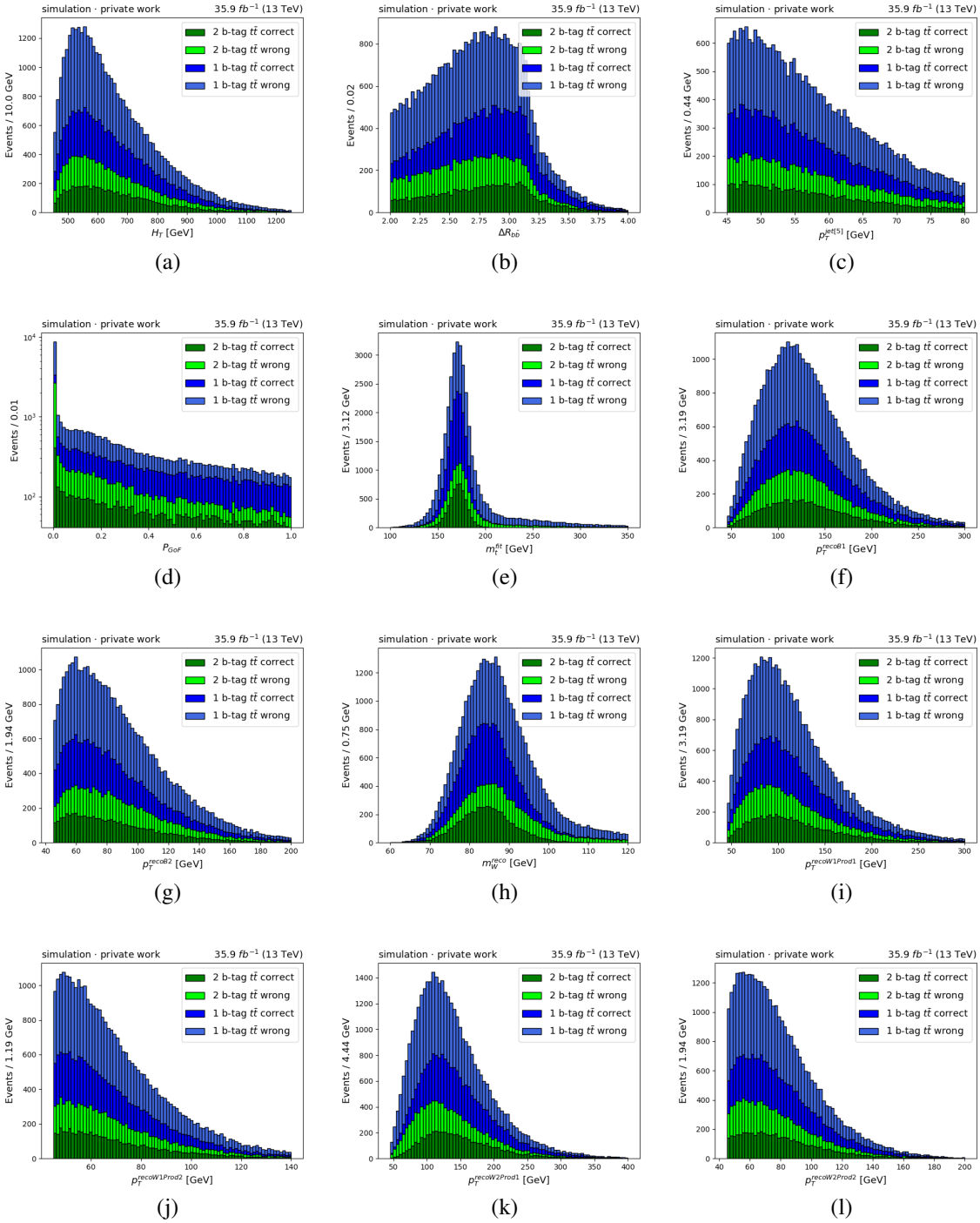


Figure 53: Feature distribution for NN cut = 0.620.

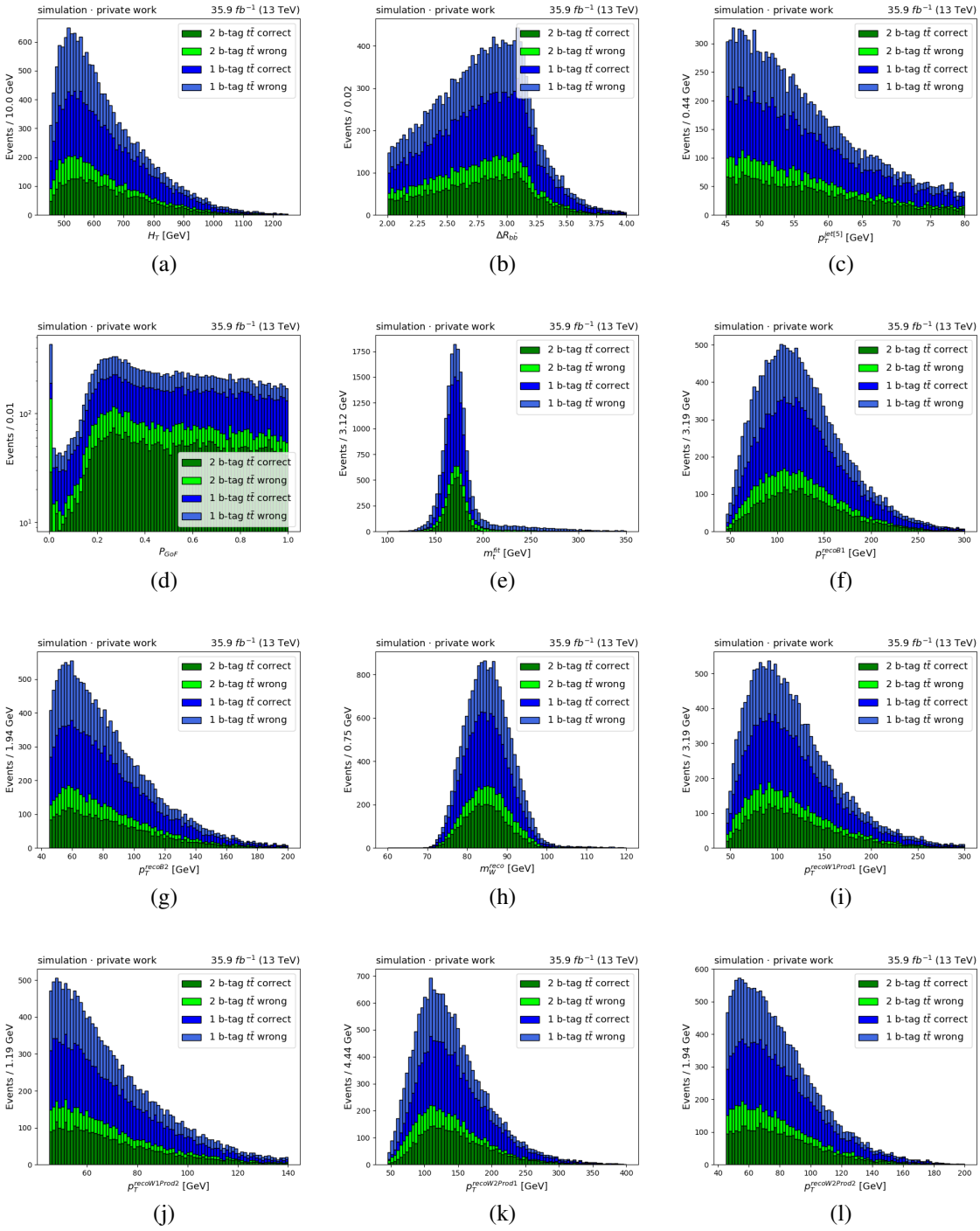


Figure 54: Feature distribution for NN cut = 0.680.

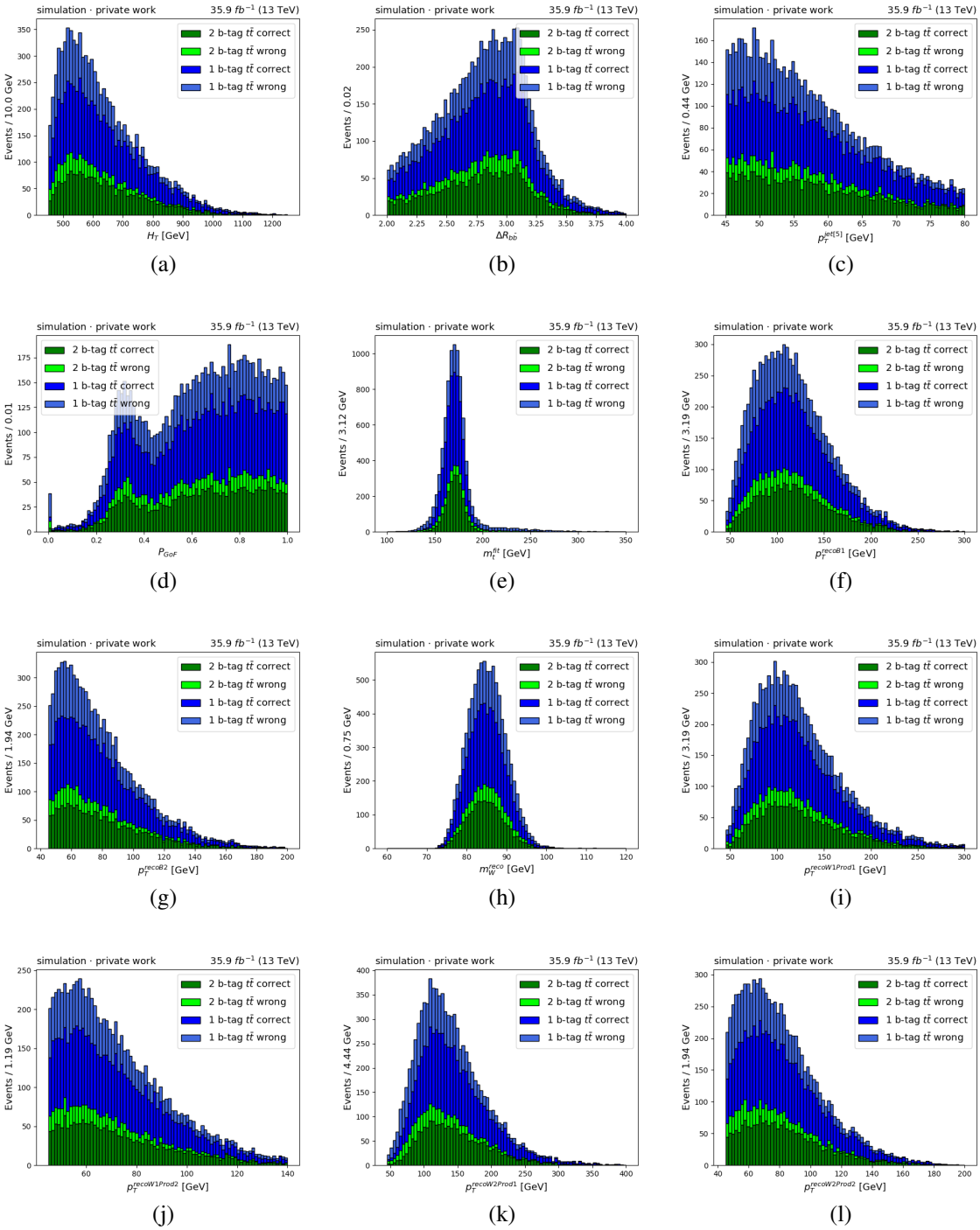


Figure 55: Feature distribution for NN cut = 0.722.

Background prediction

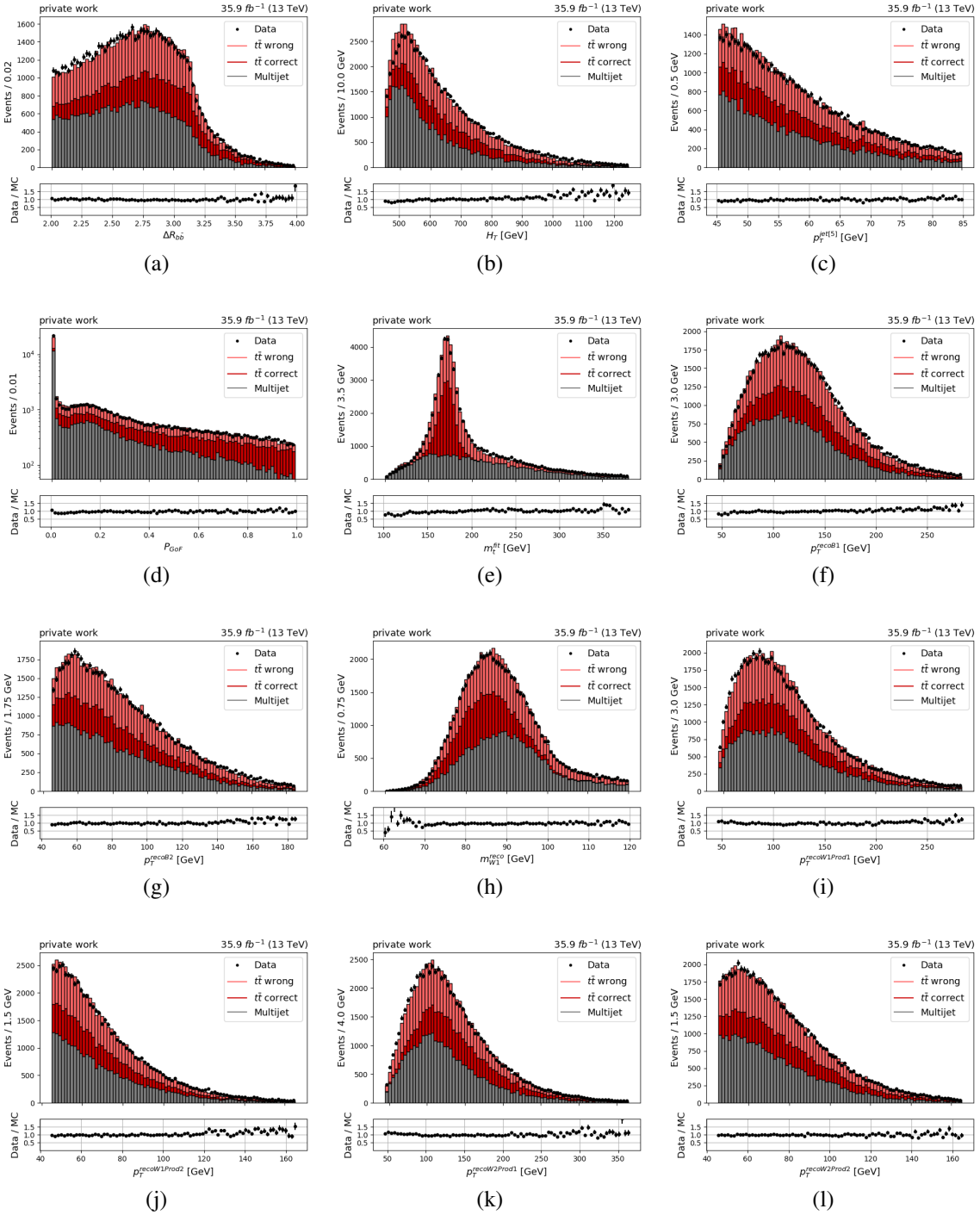


Figure 56: Background prediction for all input features and H_T of the NN at a cut value of 0.62.

References

[1] M. Beneke et al. *Top Quark Physics*. 2000. arXiv: hep-ph/0003033 [hep-ph].

[2] S. Agostinelli et al. "Geant4—a simulation toolkit". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (2003), pp. 250–303. ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(03)01368-8. URL: http://www.sciencedirect.com/science/article/pii/S0168900203013688.

[3] Paolo Nason. *A New Method for Combining NLO QCD with Shower Monte Carlo Algorithms*. 2004. arXiv: hep-ph/0409146 [hep-ph].

[4] J. D'Hondt et al. *Fitting of event topologies with external kinematic constraints in CMS*. 2006. eprint: CMSNOTE2006/023.

[5] J. Alwall et al. *Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions*. 2007. arXiv: 0706.2569 [hep-ph].

[6] Stefano Frixione, Paolo Nason, and Carlo Oleari. *Matching NLO QCD computations with Parton Shower simulations: the POWHEG method*. 2007. arXiv: 0709.2092 [hep-ph].

[7] Q Ingram. "Energy resolution of the barrel of the CMS Electromagnetic Calorimeter". In: *Journal of Instrumentation* 2.04 (Apr. 2007), P04004–P04004. DOI: 10.1088/1748-0221/2/04/p04004. URL: https://doi.org/10.1088%5C%2F1748-0221%5C%2F2%5C%2F04%5C%2Fp04004.

[8] Marion Lambacher. "Study of fully hadronic $t\bar{t}$ decays and their separation from QCD multijet background events in the first year of the ATLAS experiment". dissertation. Ludwig-Maximilians-Universitat München, 2007.

[9] Torbjörn Sjöstrand, Stephen Mrenna, and Peter Skands. *A Brief Introduction to PYTHIA 8.1*. 2007. arXiv: 0710.3820 [hep-ph].

[10] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. *The anti-k_t jet clustering algorithm*. 2008. arXiv: 0802.1189 [hep-ph].

[11] D0 Collaboration. *Evidence for production of single top quarks*. 2008. arXiv: 0803.0739 [hep-ex].

[12] The CMS Collaboration et al. In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08004–S08004. DOI: 10.1088/1748-0221/3/08/s08004. URL: https://doi.org/10.1088%5C%2F1748-0221%5C%2F3%5C%2F08%5C%2Fs08004.

[13] Simone Alioli et al. *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*. 2010. arXiv: 1002.2581 [hep-ph].

[14] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. *FastJet user manual*. 2011. arXiv: 1111.6097 [hep-ph].

[15] Michal Czakon and Alexander Mitov. *Top++: a program for the calculation of the top-pair cross-section at hadron colliders*. 2011. arXiv: 1112.5675 [hep-ph].

[16] *Performance of b tagging at sqrt(s)=8 TeV in multijet, ttbar and boosted topology events*. Tech. rep. CMS-PAS-BTV-13-001. Geneva: CERN, 2013. URL: http://cds.cern.ch/record/1581306.

[17] Mark Thomson. *Modern Particle Physics*. Cambridge University Press, 2013. ISBN: 978-1-107-03426-6.

[18] 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*. 2014. arXiv: 1405.0301 [hep-ph].

[19] M. Baak et al. *The global electroweak fit at NNLO and prospects for the LHC and ILC*. 2014. arXiv: 1407.3792 [hep-ph].

[20] Florian Beaudette. "The CMS Particle Flow Algorithm". In: 2014.

[21] John M. Campbell et al. *Top-pair production and decay at NLO matched with parton showers*. 2014. arXiv: 1412.1828 [hep-ph].

[22] The NNPDF Collaboration et al. *Parton distributions for the LHC Run II*. 2014. arXiv: 1410.8849 [hep-ph].

[23] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2014. arXiv: 1412.6980 [cs.LG].

[24] Peter Skands, Stefano Carrazza, and Juan Rojo. *Tuning PYTHIA 8.1: the Monash 2013 Tune*. 2014. arXiv: 1404.5630 [hep-ph].

[25] Sergey Ioffe and Christian Szegedy. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. 2015. arXiv: 1502.03167 [cs.LG].

[26] Roderik Bruce et al. "LHC Run 2: Results and challenges". In: CERN-ACC-2016-0103 (July 2016), MOAM5P50. 7 p. DOI: 10.18429/JACoW-HB2016-MOAM5P50. URL: https://cds.cern.ch/record/2201447.

[27] *Investigations of the impact of the parton shower tuning in Pythia 8 in the modelling of* $\mathrm{t\bar{t}}$ *at* $\sqrt{s} = 8$ *and 13 TeV*. Tech. rep. CMS-PAS-TOP-16-021. Geneva: CERN, 2016. URL: https://cds.cern.ch/record/2235192.

[28] CMS Collaboration. *Particle-flow reconstruction and global event description with the CMS detector*. 2017. arXiv: 1706.04965 [physics.ins-det].

[29] Lucio Mwinmaarong Dery et al. *Weakly Supervised Classification in High Energy Physics*. 2017. arXiv: 1702.00414 [hep-ph].

[30] Eric M. Metodiev, Benjamin Nachman, and Jesse Thaler. *Classification without labels: Learning from mixed samples in high energy physics*. 2017. arXiv: 1708.02949 [hep-ph].

[31] Andrej B. Arbuzov. *Quantum Field Theory and the Electroweak Standard Model*. 2018. arXiv: 1801.05670 [hep-ph].

[32] CMS Collaboration. *Measurement of the top quark mass in the all-jets final state at $\sqrt{s} = 13$ TeV and combination with the lepton+jets channel*. 2018. arXiv: 1812.10534 [hep-ex].

[33] Nataliia Kovalchuk. "Top quark mass measurement and color effects at the LHC". dissertation. Universiät Hamburg, 2018.

[34] *Measurement of the top quark mass in the all-jets final state at $\sqrt{s} = 13$ TeV*. Tech. rep. CMS-PAS-TOP-17-008. Geneva: CERN, 2018. URL: https://cds.cern.ch/record/2628540.

[35] Owe Philipsen. *Quantenfeldtheorie und das Standardmodell der Teilchenphysik Eine Einführung*. SpringerLink: Bücher. Springer Berlin Heidelberg, 2018. URL: https://doi.org/10.1007/978-3-662-57820-9.

[36] A.M. Sirunyan et al. "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV". In: *Journal of Instrumentation* 13.05 (May 2018), P05011–P05011. DOI: 10.1088/1748-0221/13/05/p05011. URL: https://doi.org/10.1088%5C%2F1748-0221%5C%2F13%5C%2F05%5C%2Fp05011.

[37] M. Tanabashi et al. "Review of Particle Physics". In: *Phys. Rev. D* 98 (3 Aug. 2018), p. 030001. DOI: 10.1103/PhysRevD.98.030001. URL: https://link.aps.org/doi/10.1103/PhysRevD.98.030001.

[38] Gopinath Rebala. *An Introduction to Machine Learning*. Ed. by Sanjay Churiwala. Springer eBooks: Engineering. Springer, 2019, p. 263. URL: https://doi.org/10.1007/978-3-030-15729-6.

[39] A.M. Sirunyan, A. Tumasyan, W. Adam, et al. "Measurement of the top quark mass in the all-jets final state at $\sqrt{s}$ = 13 TeV and combination with the lepton+jets channel". In: *Eur. Phys. J. C (2019) 79:313* (3 Apr. 2019). DOI: 10.1140/epjc/s10052-019-6788-2. URL: https://link.aps.org/doi/10.1103/PhysRevD.98.030001.

[40] Twiki Cern. *NNLO+NNLL top-quark-pair cross sections. ATLAS-CMS rec ommended predictions for top-quark-pair cross sections using the Top++v2.0 program (M. Czakon, A. Mitov, 2013)*. Accessed: 14.2.2020. 2020. URL: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/TtbarNNLO.

[41] Twiki Cern. *Top quark pair cross section summary of CMS measurements in comparison with the theory calculation at NNLO+NNLL accuracy. The Tevatron measurements are also shown.* Accessed: 14.2.2020. 2020. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsTOPSummaryFigures.

[42] Johannes Lange. “Measurement of the top quark mass in the all-jets final state at $\sqrt{s}$ = 13TeV and combination with the lepton+jets channel”. dissertation. Universität Hamburg, 2020.

[43] BruceBlaus. Own work, CC BY 3.0. Accessed 09.08.2020. URL: https://commons.wikimedia.org/w/index.php?curid=28761830.

[44] CERN. Accessed: 08.08.2020. URL: https://home.cern/.

[45] developers.google.com. *Classification: ROC Curve and AUC*. Accessed: 13.08.2020. URL: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

[46] DØcollaboration. *Observation of Single Top Quark Production*. Accessed: 14.08.2020. URL: https://www-d0.fnal.gov/Run2Physics/top/singletop_observation/singletop_observation_updated.html.

[47] Glosser.ca. Own work, Derivative of File:Artificial neural network.svg, CC BY-SA 3.0 Accessed 09.08.2020. URL: https://commons.wikimedia.org/w/index.php?curid=24913461.

[48] Keras. Accessed: 17.8.2020. URL: https://keras.io/.

[49] Rinat Maksutov. Accessed 09.08.2020. URL: https://towardsdatascience.com/deep-study-of-a-not-very-deep-neural-network-part-2-activation-functions-fd9bd8d406fc#:~:text=Fig.12%5C%20Comparison%5C%20of%5C%20various%5C%20activation%5C%20functions.

[50] MissMJ. *Standard Model*. Own work by uploader, PBS NOVA, Fermilab, Office of Science, United States Department of Energy, Particle Data Group. Accessed: 11.2.2020. Last updated due to new findings: 17.9.2019. URL: https://en.wikipedia.org/wiki/Standard_Model.

[51] pathmind. *Neural Network*. Accessed: 09.08.2020. URL: https://pathmind.com/wiki/neural-network.

[52] peltarion. *categorical crossentropy*. Accessed: 13.08.2020. URL: https://peltarion.com/knowledge-center/documentation/modeling-view/build-an-ai-model/loss-functions/categorical-crossentropy.

[53] scikit-learn.org. *Cross validation*. Accessed: 13.08.2020. URL: https://scikit-learn.org/stable/modules/cross_validation.html.

[54] TensorFlow. Accessed: 18.8.2020. URL: https://www.tensorflow.org/.

[55] CERN TWiki. Accessed: 12.08.2020. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults?rev=161.

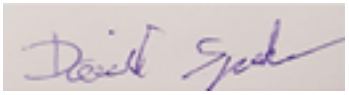
[56] Julia Woithe1, Gerfried J Wiener, and Frederik F Van der Veken. *Let’s have a coffee with the Standard Model of particle physics!* Accessed: 12.2.2020. URL: https://https://iopscience.iop.org/article/10.1088/1361-6552/aa5b25/pdf.

List of Figures

Eidesstattliche Erklärung

Ich versichere, dass ich die beigefügte schriftliche Masterarbeit selbstständig angefertigt und keine anderen als die angegebenen Hilfsmittel benutzt habe. Alle Stellen, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen sind, habe ich in jedem einzelnen Fall unter genauer Angabe der Quelle deutlich als Entlehnung kenntlich gemacht. Dies gilt auch für alle Informationen, die dem Internet oder anderer elektronischer Datensammlungen entnommen wurden. Ich erkläre ferner, dass die von mir angefertigte Masterarbeit in gleicher oder ähnlicher Fassung noch nicht Bestandteil einer Studien- oder Prüfungsleistung im Rahmen meines Studiums war. Die von mir eingereichte schriftliche Fassung entspricht jener auf dem elektronischen Speichermedium. Ich bin damit einverstanden, dass die Masterarbeit veröffentlicht wird.

Hamburg, den 20.8.2020



Ort, Datum Unterschrift