

Universität Bonn

Physikalisches Institut

Improvement of the Mass Resolution in the $H \rightarrow b\bar{b}$ Decay Using Multivariate Regression in ATLAS Data

Elisabeth Schopf

In the context of Higgs searches in the $H \rightarrow b\bar{b}$ decay channel at ATLAS a good invariant di-jet mass resolution is important for the improvement of the signal to background ratio. This thesis demonstrates a method to improve the invariant mass resolution by using boosted decision trees. These decision trees were trained to give a better estimate of the true b -jet energy based on several input variables describing the jet. The influence of the input variables was investigated and several test were performed in order to optimize the setup that was used in this multivariate regression technique. As a result the invariant mass resolution of two b -jets was improved by 30% to a value of 10.9%. The multivariate regression technique was also successfully applied to other physics processes containing b -jets such as top-quark pair production. In addition, a comparison of the multivariate regression to other jet energy corrections shows that its performance is similar or even better.

Physikalisches Institut der
Universität Bonn
Nussallee 12
D-53115 Bonn



BONN-IB-2014-07
Juli 2014

Acknowledgements

I would like to thank Norbert Wermes for giving me the opportunity to work on this interesting topic within the research of his group. I would like to thank Eckhard von Törne for the supervision and that he never became tired of explaining, answering questions, troubleshooting, giving advice and coming up with new ideas and inspiration. I would like to thank Stephan and Jan for being great office mates, answering questions, helping me with problems and for funny conversations during coffee breaks. I would like to thank Götz and Vadim for helping me from far away and being critical and coming up with new ideas in the meetings. I would like to thank Jochen Dingfelder who also supported me during my master thesis. I would like to thank Eckhard, my parents, Kristof and Jan for comments and corrections for my thesis. Of course an extra "thank you" is addressed to my parents for their support in any situation, their encouragement and their interest even when they listened to my detailed explanations of jet energy measurements. I would also like to thank many others who helped and supported me on my way, especially my fellow-sufferers since the first day of studies. Last but not least I would like to thank everybody who ever shared a coffee/chocolate/tea/cake/beer with me in the last year since sometimes this is a very welcome change.

Contents

1	Introduction	1
2	Introduction to High Energy Particle Physics	3
2.1	Physics of Hadron Collisions	3
2.2	The Large Hadron Collider	5
2.3	The ATLAS Detector	7
2.4	Reconstruction of Kinematic Properties of Particles	9
2.5	Higgs Physics	10
2.5.1	Theory of the Higgs Boson	11
2.5.2	The Mass of the Higgs Boson	11
2.5.3	The Decay Channel $V, H \rightarrow b\bar{b}$	12
2.6	Physics of Jets	13
2.6.1	Identification of b -Jets	14
3	Technical Introduction	17
3.1	Monte Carlo Generators and Samples	17
3.2	The Data	19
3.3	Analysis Frameworks	19
3.3.1	Overkill	20
3.4	The Bonn D3PDs	21
3.5	The Standard $V, H \rightarrow b\bar{b}$ Analysis	21
4	Multivariate Analysis	23
4.1	Multivariate Techniques	23
4.2	Boosted Decision Trees (BDTs)	24
4.3	The TMVA Framework	25
5	b-Jet Energy Correction Using Multivariate Regression	27
5.1	Motivation	27
5.2	Training Samples	28
5.3	Target Variable	30
5.4	Input Variables	31
5.4.1	Physical Motivation of the Input Variables	31
5.4.2	Choice of a Set of Input Variables	32
5.5	Event and Object Selection	35
5.6	Preparation of the Samples, Variables and the BDTs	36

5.7	Training Results: the Jet E_T Resolution	37
5.8	Application of the Trained Model	38
5.9	Application to non- $W^\pm H \rightarrow l^\pm \nu^{(\pm)} b \bar{b}$ Samples	40
5.10	Application to Data	43
5.11	Comparison with other Jet Correction Methods	44
5.11.1	Comparison to the Standard $V, H \rightarrow b \bar{b}$ Correction	44
5.11.2	Comparison to an Alteration of the Standard Correction	45
5.11.3	Comparison to Other Methods	47
5.12	Consistency Test of the Input Variables	47
5.12.1	Importance of the Input Variables	48
5.12.2	Test of an Alternative Target Variable	50
5.12.3	Test of Alternative Sets of Input Variables	51
6	Summary	55
7	Outlook	57
7.1	Determination of a Final Set of Input Variables	57
7.2	Extension of the Regression to the Full Data Set and Compatibility with ATLAS Analyses	58
A	Appendix	59
A.1	Discussion of Different Measures for the $m_{b\bar{b}}$ Resolution	59
A.2	Additional Figures and Tables for the Analysis	62
A.3	Modeling of the Input Variables for the Training	66
	Bibliography	75
	List of Figures	79
	List of Tables	81

Introduction

"... for the theoretical discovery of a mechanism that contributes to our understanding of the origin of mass of subatomic particles, and which recently was confirmed through the discovery of the predicted fundamental particle, by the ATLAS and CMS experiments at CERN's Large Hadron Collider"

Nobel Prize committee, Oct. 2013

On the 8th of October 2013 Peter Higgs and Francois Englert were awarded the Nobel Prize for physics for the theoretical description of a mechanism which breaks electroweak symmetry spontaneously. This mechanism is very important for elementary particle physics since it describes how fundamental particles gain their mass. An additional fundamental particle corresponds to this mechanism: the so-called standard model Higgs boson. The ATLAS and CMS experiments at the Large Hadron Collider (LHC) were designed to have among other purposes the potential to discover this boson. Since the operation started in 2010 the LHC as well as the detectors worked very well and provided large amounts of data. In summer of 2012 ATLAS and CMS were finally able to report an significant excess of data which was compatible with the hypothesis of a standard model Higgs boson at a mass of approximately 126 GeV. This was, after over 40 years, a great breakthrough in the history of the Higgs mechanism or as Peter Higgs said: *"It's very nice to be right sometimes."* But the work for the experimentalists is not done yet. Now that the Higgs boson is discovered the work even starts since it has to be studied further. One open issue is the observation of all predicted decays of the Higgs boson. The channel which is the most probable one is the decay into a bottom-antibottom-quark pair. But a $b\bar{b}$ pair is a very prominent signature in hadron colliders not only for the Higgs boson. Therefore much work has to be done to separate the Higgs boson signal for $H \rightarrow b\bar{b}$ decays from the background processes. One important issue is to measure the energy of the b -quarks which manifest as particle bundles — jets — in the detector as precise as possible. The b -jet's energy is measured in the calorimeter of the ATLAS detector via the interaction with the detector material and creating showers of particles. The energy depositions of those particles in the active detector material is a measure for the energy of the jet. Hence the jet energy is reconstructed from the sum of many single energy measurements. The energy measurement is intrinsically limited by the resolution of the calorimeter. The reconstruction of the jet from the single energy depositions in the calorimeter leads to further inaccuracies since a dedicated algorithm has to cluster the energy depositions to jets. This reconstruction technique cannot assure that no jet energy depositions

are missed or that depositions are picked up which do not originate from the jet. Further sources for inaccuracies of the jet energy measurement are regions in the detector, so-called dead regions, where no measurements can be recorded as no detecting material is present. Since the jet consists of various particles the jet energy measurement is deteriorated further due to different responses of the detector to different particles. All those effects result in an inaccurate jet energy measurement which has to be corrected afterwards. This master thesis is dedicated to find a correction for the b -jet energy by using machine learning algorithms. Among standard ATLAS analyses this is a new approach. This so-called multivariate regression uses the machine learning algorithm to reconstructs the true jet energy from measured jet properties. The algorithm learns the relationship between the energy (also called target) and the measured parameters (also called input variables) from artificial data generated by Monte Carlo generators. In those samples the true jet energy is known as well as the input variables. Due to the huge amount of data and its complexity this approach seems to have the chance to give a better performance than hand made step-by-step calibrations. If the b -jet energy resolution is improved also the resolution of the Higgs mass will improve as it is calculated from the four momentum vectors of the two b -jets. The excess of data in the signal region becomes more significant for good Higgs mass resolutions since the amount of signal events in this region will be enhanced.

Introduction to High Energy Particle Physics

This chapter aims at giving a short introduction to elementary particle physics. The theoretical base is the standard model of particle physics. It describes twelve elementary particles (fermions) and their interaction via three forces. The elementary particles are separated into six leptons and six quarks. The three fundamental forces – electromagnetic, weak and strong force — are transferred via force carrier particles (bosons). The additional inclusion of the Higgs mechanism and its boson in the standard model is supposed to explain the origin of the mass of the elementary particles. To study the standard model and aim for new discoveries high energies are necessary which are achieved in particle collisions.

2.1 Physics of Hadron Collisions

Nowadays much research in the field of particle physics is done by observing processes during particle collisions. One distinguishes between two types of colliders : those colliding elementary particles and those colliding hadrons. In most colliders electrons/positrons or protons/antiprotons are used. The main difference is that at e^+e^- -colliders fundamental particles collide. Therefore these colliders are very suitable for precision measurements since the events are relatively clean. By tuning the center-of-mass energy specific processes may be probed. Since protons are composite objects much more might happen during a collision and the events are less clean. One reason for building proton colliders is the achievable center of mass energy E_{cm} . The center of mass energy for colliding beams where the particles in both beams have the same energy is [1]:

$$E_{\text{cm}} = \sqrt{m_1^2 + m_2^2 + 2E_1E_2 + 2p_1p_2} \approx \sqrt{4E_1E_2} \quad \begin{matrix} E_1=E_2=E \\ \hline \end{matrix} \quad 2E \quad (2.1)$$

with $m_1, m_2 \ll E_1, E_2 \rightarrow E_1 \approx p_1, E_2 \approx p_2$

The center of mass energy is an important parameter since smaller structures can be probed with larger energies according to the de Broglie wavelength $\lambda = 2\pi/k$ where k is the momentum. In addition particles with higher masses can be produced in reactions with higher center of mass energies. The crucial point is the energy loss of charged particles via synchrotron radiation. The energy loss per turn ΔE in the accelerator is proportional to [2]:

$$\Delta E \propto \frac{1}{\rho} \frac{E^4}{m^4} \quad (2.2)$$

where ρ is the bending radius, E the energy of the charged particle and m the mass of the charged particle. As a consequence electrons cannot be accelerated to the same energy as protons due to their low mass of $m_e \approx 0.5 \text{ MeV}$. This is one reason to use protons instead of electrons in high energy colliders. Another reason is the potential for discoveries. Since the proton is a composite object the center of mass energy during the collision is not fixed but covers a large energy range (see formula 2.3). Therefore a large amount of processes can be probed [1, 2].

A proton is build up of three valence quarks: two up-quarks (u) and one down-quark (d). Quarks are fundamental particles with charges of $2/3e$ (for up-type quarks) and $-1/3e$ (for down-type quarks). The quarks in the protons are bound together by the strong force whose force carriers are the gluons. The gluons in the proton can split into quark-antiquark-pairs. These quarks are called sea quarks. Therefore the proton consists of three different types of particles which are all in motion. Valence quarks, sea quarks and gluons are called partons. If the energy during a collision is high enough the individual constituents of the proton interact with each other instead of the proton on the whole. During these inelastic collisions, also called hard interactions, the partons are considered as quasi free. Due to the proton structure it is difficult to predict which particles in the colliding protons interact and which fraction of the protons' momentum they carry. Experimentally determined parton distribution functions (pdfs) describe the probability of finding a parton of a given flavor with a momentum fraction x of the proton momentum in a reaction with a momentum transfer Q^2 . Although the pdfs are dependent on the momentum transfer during the collision a valence quark carries on average about 15% and the gluons about 50% of the proton's momentum [1]. Therefore the center of mass energy in the collision of a parton with momentum fraction x_1 of the momentum of the first proton and a parton with x_2 of the momentum of the second proton reduces to $\hat{E}$ [1]:

$$\hat{E} = \sqrt{x_1 x_2} E \quad (2.3)$$

with E being the center of mass energy of the incoming protons. The particles which were produced during the hard interaction have a high momentum transverse to the beam direction and are considered as "interesting" since in this processes new particles can be produced.

The quarks in the protons interact via three fundamental forces: electromagnetic force, weak force and strong force. The photon, which is the massless, uncharged force carrier of the electromagnetic force couples to every particle with an electromagnetic charge. In addition the quarks carry a so-called weak hypercharge to which the weak force couples through the $W^\pm$ and Z^0 bosons. Quarks and gluons have a color charge (red, green or blue) which is the charge of the strong force. Therefore they couple to the force carrier of the strong force: the gluon. A feature of the strong force which is called confinement states that observable objects have to be color neutral. Color neutral states are composed of three quarks with different colors (red, blue and green) which are called baryons (e.g. the proton). Quark-antiquark systems are called mesons. After the collision the proton is destroyed as one or more partons were detached from the original proton structure during the hard interaction. As a consequence of confinement all partons coming from the protons and all partons originating from the proton-proton collision have to find partners to form color neutral objects. This process is called hadronization which will be discussed further in sections 2.6 and 3.1. During the hadronization a bunch of additional particles are produced. While the hadrons coming from partons produced in the hard interaction have a high transverse momentum the hadrons coming from the proton remnants have a high longitudinal momentum along the beam axis. The latter is called underlying event. Another effect which leads to less clean signatures of the hard interaction process are initial and final state radiation (ISR/FSR). ISR and FSR is the radiation of photons and gluons off the quarks. The gluons undergo hadronization afterwards. A further complication found in high luminosity (see formula 2.5) colliders arises due to multiple interactions per crossing

of proton bunches. This is called pile-up. At the LHC (section 2.2) a single bunch of protons contains 10^{11} protons which leads to approximately 30 simultaneous proton-proton collisions for luminosities of $10^{34} \text{ cm}^{-2}\text{s}^{-1}$ [3]. The large number of protons per bunch and thus the high luminosity is necessary to increase the probability for a hard interaction during a collision to reasonable levels. The pile-up interactions are soft, i.e. quasi elastic, and produce particles with low transverse momenta. These soft proton-proton collisions are also called minimum bias events. It is very unlikely that two hard interactions happen during one collision of proton bunches. A schematic picture of a proton-proton collision which illustrates the above mentioned effect is shown in fig. 2.1. [1, 2, 4]

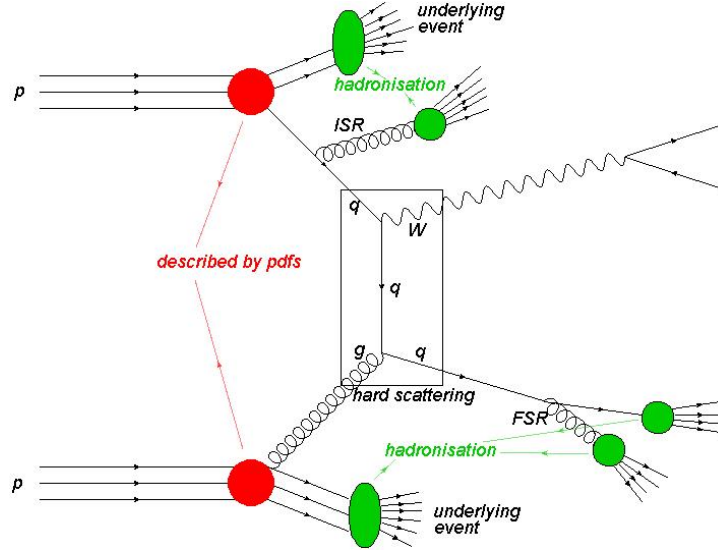


Figure 2.1: Schematic view of a proton-proton collision: the parton structure during collision is described by pdfs, the hard interaction process is $qg \rightarrow Wq$, the hard event is polluted by the underlying event of the proton remnants, ISR and FSR, all partons in the event undergo hadronization.

2.2 The Large Hadron Collider

The LHC (Large Hadron Collider) is a circular accelerator with a circumference of 26.7 km. It is located at CERN (Conseil Européen pour la Recherche Nucléaire) at the border between Switzerland and France near the city of Geneva. The collider was built 45 m to 170 m under the surface in the already existing tunnel of the LEP (Large Electron Positron) collider. The purpose of the LHC is to collide protons with a center of mass energy up to 14 TeV to search for physics beyond the standard model. Besides accelerating protons the LHC also runs with heavy ions. Up to now the LHC operated at a center of mass energy of 7 TeV during 2011 and 8 TeV during 2012. The design energy was not reached due to technical difficulties. After the currently ongoing, 2 year long shutdown, during which the LHC and the detectors will be upgraded, the design center of mass energy of 14 TeV might be reached. The protons which are extracted from a hydrogen source are first accelerated up to an energy of 50 MeV in a linear accelerator (LINAC). After the LINAC the protons are accelerated in three consecutive synchrotrons up to an energy of 450 GeV. With this energy the protons are injected into the LHC and accelerated up to an energy of 8 TeV. The schematic view of the accelerator chain is shown in fig. 2.2. The LHC is built with two beam pipes where the proton beams circulate in opposite directions. The beams cross each other at four interaction points where the Experiments ATLAS, CMS, LHCb and ALICE

are located. The proton beam consists of bunches of protons each containing 10^{11} protons. The design spacing between the bunches can be as low as 25 ns. For the operation with protons with such high energies high magnetic fields are required to manipulate the beam. The magnetic field strength has to be approximately 8 T which makes it necessary to work with superconducting accelerator components. [2, 3, 5]

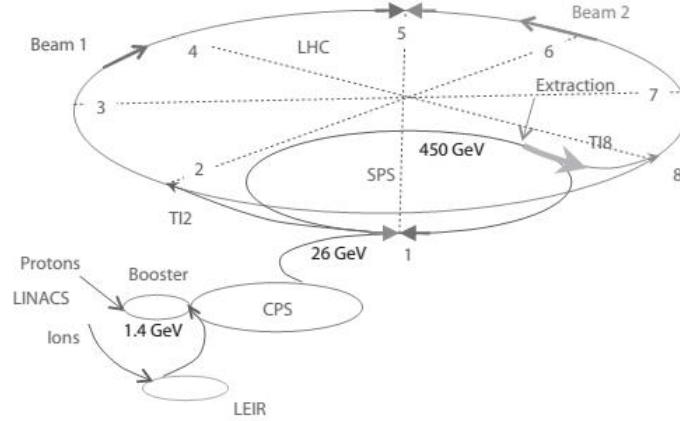


Figure 2.2: The LHC and its pre-accelerators which accelerate the protons up to 450 GeV, after injection to the LHC ring they are accelerated up to an energy of 8 TeV [2].

An important parameter for colliders is the luminosity L from which the event rate N_{event} may be calculated [3]:

$$N_{\text{event}} = L\sigma_{\text{event}} \quad (2.4)$$

where σ_{event} is the cross section for the physics process under study and L is the machine luminosity. The luminosity describes the number of collisions per area and per unit of time. It is a purely accelerator dependent quantity [3]:

$$L = \frac{N_b^2 n_b f_{\text{rev}} \gamma_r}{4\pi \epsilon_n \beta^*} F \quad (2.5)$$

where N_b is the number of particles per bunch, n_b the number of bunches per beam, f_{rev} the revolution frequency, $\gamma_r = (\sqrt{1 - v^2/c^2})^{-1}$ the relativistic gamma factor, ϵ_n the normalized transverse beam emittance, β^* the beta function at the collision point and F the geometric luminosity reduction factor. The emittance and the beta function are measures for the focusing properties of the accelerator and are related to the beam sizes $\sigma_{x,y}$ for Gaussian shaped beams. F is a reduction factor which originates from the fact that the proton beams do not collide head on but cross under a certain angle at the collision points. The design peak luminosity of the LHC is $L = 10^{34} \text{ cm}^{-2}\text{s}^{-1}$. Formula 2.4 implies that physics processes might produced with a higher rate if the luminosity of the collider is higher. This also means that very rare processes can only be studied with high luminosity machines. The ability to study new physics was a main design criterion for the LHC. In figure 2.3 the cross section for various physics processes in proton-(anti)proton collisions in dependence on the center of mass energy is shown. This shows that at the LHC energies the cross sections for Higgs production are several orders of magnitude higher than for the Tevatron energy. The Tevatron was the hadron collider with the highest collision energy before the LHC. Also the cross section for other processes which should be studied further at the LHC (bottom and top quark physics as well as vector boson physics) are higher. The over all cross section for a proton-(anti)proton collision is 100 mb which is several orders of magnitudes higher than

the processes under study. One challenge the LHC detectors have to address is the selection of rare, but interesting events out of the huge number of collisions [2–4].

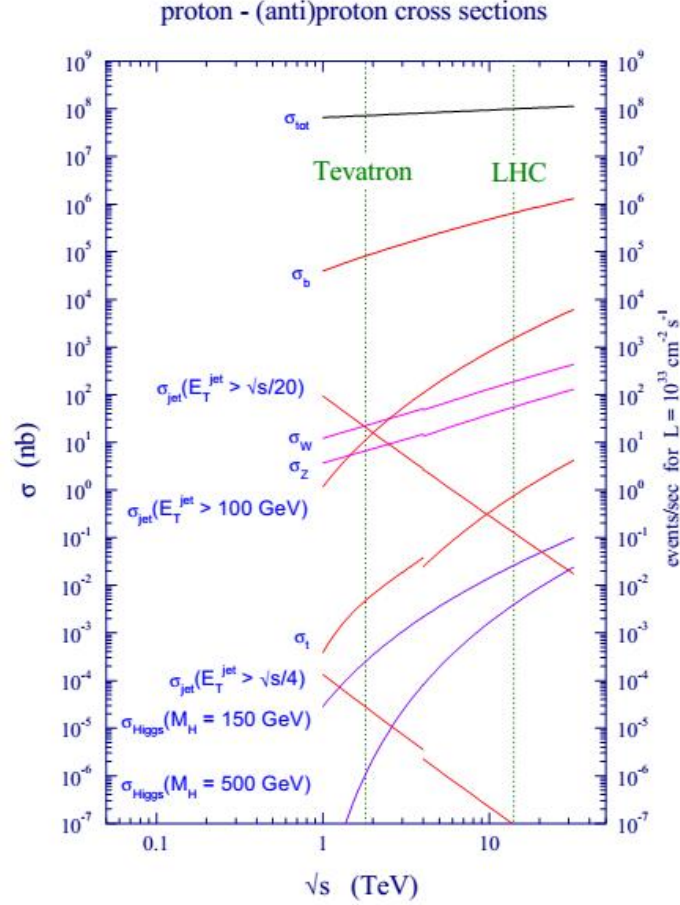


Figure 2.3: Cross sections for several physics processes in proton-(anti)proton collisions in dependence on the center of mass energy E_{cm} ; all illustrated cross sections are higher for the LHC E_{cm} compared to the Tevatron E_{cm} including the cross section for Higgs production [4].

2.3 The ATLAS Detector

The ATLAS (A Toroidal LHC ApparatuS) detector is a multi-purpose detector at the LHC which is designed to precisely measure high energetic hadron collisions with large luminosities.¹ Figure 2.4 shows a side view of the ATLAS detector as well as a sector of a cross-section. The ATLAS detector is build as a magnetic spectrometer following a cylindrical symmetry. The detection layers of the ATLAS detector are (from inside to outside):

Tracker: The tracker is used for track reconstruction of charged particles. It is built within a solenoidal magnetic field in order to measure the momentum p of charged particles. The innermost layers consist of silicon pixel detectors followed by four layers of silicon strip detectors. Because the highest particle density is reached close to the interaction point a pixel detector is used which

¹ The CMS experiment at the LHC is also a multi-purpose detector with a very similar physics program.

has a high granularity and a very good spatial resolution. In addition the used material has to be radiation hard to withstand damage due to the high particle flux. The outer part of the tracker is a transition radiation tracker which consists of several thousand gas filled drift tubes. The gaps between the tubes are filled with fibers which are chosen such that only electrons emit transition radiation. The transition radiation photons lead to a larger signal for electrons/positrons which is used to separate them from other charged particles [6].

Calorimeters: The calorimeters are built to measure the energy of long-lived particles. There are two calorimeters for measuring the particles' energy: the inner one for electrons/protons and photons and the outer one for hadrons, e.g pions, neutrons. The calorimeters are designed such that the incoming particles induce showers by interacting with the detector material and deposit all their energy in the calorimeter. Both calorimeters are sampling calorimeters which means that absorbing and detecting layers alternate. The electromagnetic calorimeter has a accordion-like sampling structure which improves its resolution. In the absorber layers (lead, iron) the particles shower and in the detecting layers (liquid argon, scintillator) the signal is produced. The amount of photons which are produced in the detecting layers is proportional to the energy of the particle. The relative energy resolutions σ_E/E in the barrel region ($|\eta| < 2.5$) for the electromagnetic (ECAL) and the hadron calorimeter (HCAL) are [6]:

$$\text{ECAL: } \frac{\sigma_E}{E} = \frac{(10.0 \pm 0.4)\%}{\sqrt{E(\text{GeV})}} \oplus (0.2 \pm 0.1)\% \quad (2.6)$$

$$\text{HCAL: } \frac{\sigma_E}{E} = \frac{(56.4 \pm 0.4)\%}{\sqrt{E(\text{GeV})}} \oplus (5.5 \pm 0.1)\% \quad (2.7)$$

Muon spectrometer: Muons are (excepts for neutrinos) the only known particles which escape the detector. Therefore the outer part of the ATLAS detector is designed as a muon spectrometer composed of 3 layers of muon detectors within a toroidal magnetic field. It is essential for the identification of muons and gives additional to the tracker a measurement of the muon momentum [6].

As already mentioned neutrinos escape the detector undetected because they only interact via the weak interaction. They are reconstructed indirectly using the missing transverse energy in events (see formula 2.10). In fig. 2.4 the signatures of different particles in the ATLAS detector are shown. The particles which are drawn in the picture are basically all directly observable particles except for the neutrino. For defining kinematic properties of the objects measured in the detector a coordinate system has to be defined. The origin of the coordinate system is per definition the vertex of the hard interaction. The direction along the beam line is the z -direction, the y -direction points upwards and is perpendicular to $\vec{z}$ and the x -direction points from the interaction point to the center of the LHC ring and is orthogonal to $\vec{z}$ and $\vec{y}$. Furthermore the azimuthal angle ϕ and the polar angle θ are used to describe directions in the coordinate system [8].

Another challenge the ATLAS detector has to face is the data acquisition. Based on the peak luminosity and the total cross section for proton-proton collisions 10^9 events per second are expected. This corresponds to a huge amount of data which cannot be saved fast enough until the next collision happens. Hence a fast three stage trigger system is used to select interesting events for storage. The first level trigger (L1) uses a limited amount of the detector information and just searches for muons, electrons, photons, jets and τ -leptons as well as for events with high missing transverse energy. This decision is made within $2.5 \mu\text{s}$. The events which pass the L1 trigger are processed by the second level trigger (L2). The L2 trigger uses the information in regions in $\eta\phi$ where the L1 trigger has identified

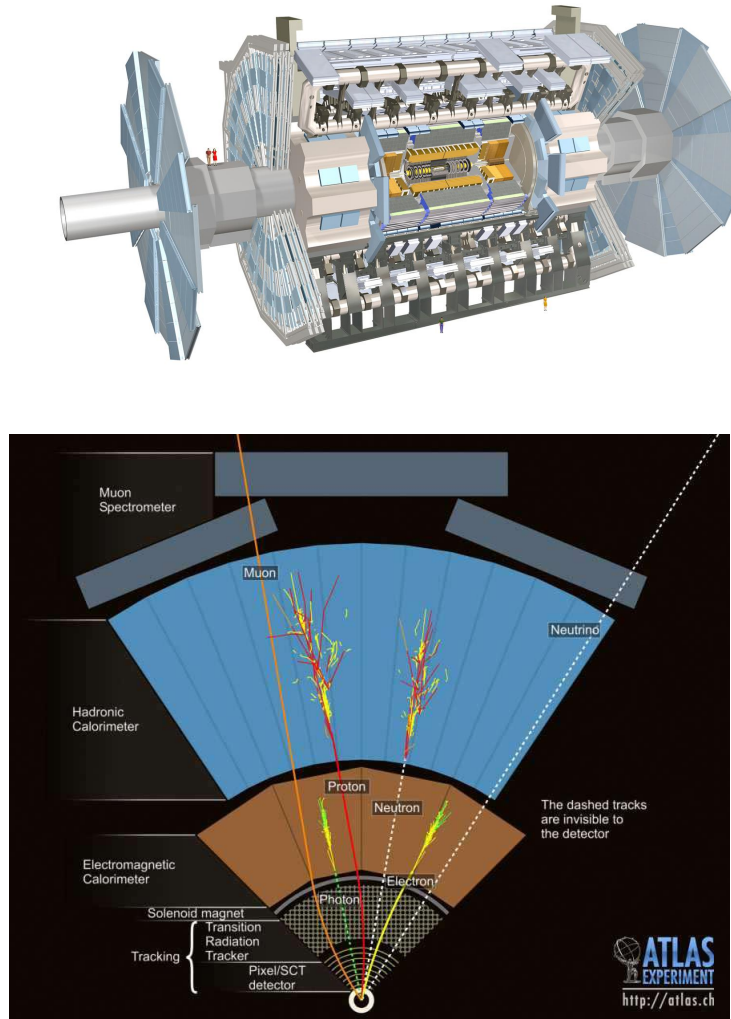


Figure 2.4: The ATLAS detector: above a computer generated view of the whole detector and below a cross section which shows how different particles manifest in the different layers of the detector [7].

trigger objects (ROIs = regions of interests). With the information about the ROIs the L2 trigger limits the amount of data which must be transferred from the detector readout. The third trigger level is the high level trigger which uses offline analysis procedures on the L2 objects to decide which events may include physics processes to study. This reduces the amount of events which are recorded further. At the end the recorded event rate is 200 Hz [6].

2.4 Reconstruction of Kinematic Properties of Particles

Most of the particles which are produced in the proton-proton collisions decay (or hadronize) before they enter the detector. The particles which live long enough to leave a signature in the detector are shown in fig 2.4. Therefore short-lived particles have to be reconstructed from their decay products. The mass of a decayed particle can be reconstructed from the four momentum vectors $\mathbf{p}$ of the decay

products by calculating their invariant mass M :

$$M^2 = (\mathbf{p}_1 + \mathbf{p}_2)^2 = \left[\begin{pmatrix} E_1 \\ \vec{p}_1 \end{pmatrix} + \begin{pmatrix} E_2 \\ \vec{p}_2 \end{pmatrix} \right]^2 \quad (2.8)$$

with $\mathbf{p}_{1,2}$ being the four momentum vectors of decay product 1,2. For defining directions ϕ and instead of the polar angle θ more commonly the pseudorapidity η is used [9]:

$$\eta = -\ln \left(\tan \frac{\theta}{2} \right) \quad (2.9)$$

The pseudorapidity has the advantage that differences in η are invariant under Lorentz transformation along $\vec{z}$ and the amount of particles per interval of η is approximately constant. Distances ΔR between objects are usually calculated in the $\eta\phi$ -plane: $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$ [9]. Another special measuring technique is the usage of transverse quantities such as the transverse momentum $p_T = |\vec{p}| \sin \theta$. This relies on the fact that the colliding partons normally have different momenta which leads to a boost along the z -axis for the produced particles. Hence the kinematics of the whole event are only meaningfully defined in the plane transverse to the beam. The incoming partons initially only have a longitudinal momentum component and a negligible momentum transverse to the beam. This means that the sum of the transverse momenta of all objects in the detector have to be zero. With this assumption the missing transverse momentum $\vec{p}_T^{\text{mis}}$ is calculated:

$$\vec{p}_T^{\text{mis}} = - \sum_i \vec{p}_T^{(i)} \quad (2.10)$$

Small missing momentum in every event originates from inaccuracies in the measurements and undetected parts of the event since not the whole solid angle around the interaction point is covered with active detector material. However, a large missing transverse momentum in an event implies that an undetected particle left the detector, e.g. a neutrino. In ATLAS the missing transverse momentum is called missing transverse energy E_T^{mis} because the energies in the events are large enough that the relation $E \approx p$ is valid. [8, 9]

2.5 Higgs Physics

It was in the 1960s when Higgs, Englert and Brout[†] first described a spontaneously symmetry breaking mechanism, also called Higgs mechanism, and other persons developed the model further so that it fits into the mathematical description of the elementary particles and forces — the standard model. Over the years the theoretically predicted particles and processes in the standard model were discovered and studied and all fitted well into the standard model. Just the Higgs boson which is an indicator for the Higgs mechanism was still elusive and searches in a low mass region of several GeV did not lead to a discovery. Recently when the LHC was built the Higgs boson mass region of several hundred GeV came into reach and the ATLAS and CMS detectors are designed such that they have the potential to discover the Higgs boson. In the summer of 2012 the ATLAS and the CMS experiments announced the discovery of a boson with a mass of approximately 126 GeV [10, 11]. So far this boson has all properties of the predicted standard model Higgs boson. Several more exotic theories predict more than one Higgs boson with different properties for which no evidence have been found yet. This chapter will just focus on the so-called standard model Higgs boson and will also include the discovery of 2012.

2.5.1 Theory of the Higgs Boson

The elementary particles and their interactions are described by the standard model of particle physics. The whole theory is described by one Lagrangian which contains the description of interactions between particles for different forces via fields. Throughout the years it was found that the standard model describes electroweak and strong interactions very well. But the standard model without Higgs terms in the Lagrangian cannot explain why the observed elementary particles have a mass. Mass terms cannot be added to the Lagrangian since they would violate gauge invariance which is not allowed.² Hence the masses have to be generated dynamically. This is achieved by the so-called Higgs mechanism which leads to spontaneous symmetry breaking of the electroweak symmetry by introducing a scalar field ϕ . This field is only locally gauge invariant. The potential $V(\phi)$ of this field has a shape reminding of a Mexican hat with the property that at $\phi = 0$ the potential is non zero. Via spontaneous symmetry breaking the field gets to its minimum [12]:

$$\phi_{\min} = \frac{v}{\sqrt{2}} = \sqrt{\frac{\mu^2}{2\lambda}} \quad (2.11)$$

where v is the non-vanishing vacuum expectation value, μ and λ are parameters of the potential. Fluctuation around this minimum manifest in a observable massive boson — the Higgs boson. Its mass is [12]:

$$m_H = v \sqrt{2\lambda} = \sqrt{2}\mu \quad (2.12)$$

where $v = 246 \text{ GeV}$ is calculated from electroweak theory and μ and λ are introduced by the Higgs mechanism and cannot be calculated. Therefore the mass of the Higgs boson can only be determined by measurements. If the Higgs field is introduced into the Lagrangian one obtains coupling terms for the Higgs boson with itself, the massive bosons and the fermions which gives them their masses [12, 13].

2.5.2 The Mass of the Higgs Boson

As already mentioned, the mass of the Higgs boson cannot be predicted by theory. Nevertheless theory and measurements at previous accelerators already provide some constraints on the Higgs mass. Assuming the standard model holds up to the Planck scale ($m_{\text{pl}} \approx 10^{19} \text{ GeV}$) theoretical constraints on the minimum and maximum Higgs mass are given. The lower theoretical bound which is given by the stability of the Higgs potential is at $m_H^{\min} \approx 115 \text{ GeV}$ [14]. The upper bound is connected to the scale (Planck scale) at which the standard model breaks down and new physics arise and constrains the Higgs mass to $m_H^{\max} \approx 180 \text{ GeV}$ [14]. If new physics enters earlier the bound for the maximum Higgs mass is expendable to several hundred GeV. Nevertheless some mass ranges were either already excluded or favored by results from other accelerators. The LEP e^+e^- -collider and the Tevatron $p\bar{p}$ -collider excluded a mass of $m_H < 114.4 \text{ GeV}$ and a mass range of $147 \text{ GeV} < m_H < 179 \text{ GeV}$ respectively. Tevatron saw a small excess in data for $115 \text{ GeV} < m_H < 135 \text{ GeV}$. The exclusion from LEP and Tevatron were made with 95% confidence level whereas the excess seen at Tevatron was not statistically significant [14]. According to this knowledge the challenge at the LHC is not only the search for the Higgs boson but also determining its mass. The large center of mass energies at the LHC allow the coverage of a large Higgs mass range up to several hundred GeV. Since the Higgs boson decays immediately after production only its decay products can be detected. Which decay channel is predominant also depends on the Higgs mass as it shown in fig. 2.5. Since the coupling of the Higgs boson to particles should

² Gauge invariance means that symmetries are conserved and the physics which are described remain the same under gauge transformations.

depend on the particle's mass the decay into the heaviest possible particle given by half of the Higgs mass is the favored decay channel. The decay into massless particles such as $H \rightarrow \gamma\gamma$ is also possible via loop diagrams. But the branching ratios for these decays are very small because the appearance of a loop is suppressed. In summer 2012 the ATLAS and the CMS experiments at the LHC announced that they found a 5σ excess over the background estimation for a mass with a particle with approximately 126 GeV. Clear signals are seen in the decay channels $\gamma\gamma$, ZZ and W^+W^- [10, 11]. Although these are not the predominant decay channels, they do not suffer from much background in the signal region or the backgrounds are well predictable. The dominant decay channel for a Higgs boson with this mass is the decay to $b\bar{b}$. This master thesis is dedicated to searches within this decay channel. [2, 12, 14]

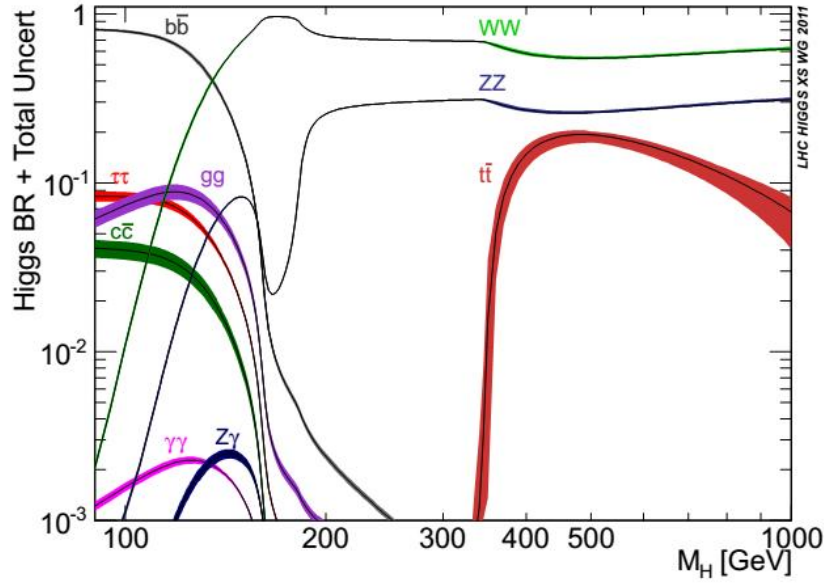


Figure 2.5: Branching ratios for the different Higgs decay channels in dependence of the Higgs mass; for small masses decays into fermion pairs dominate, for high masses decays into W^+W^- dominate [14].

2.5.3 The Decay Channel $V, H \rightarrow b\bar{b}$

In the Higgs mass regime of 126 GeV the decay into a pair of bottom-quarks is the most likely one with a branching ratio of approximately 60%. No signal for a Higgs boson was yet seen in this channel as it suffers from some difficulties. The main problem is the two jets signature of this decay. The distinction of b -jets from jets originating from other quark flavors or gluons is only possible to a certain extent (see section 2.6). Hence the measurements in this decay channel are disturbed by a lot of background processes. Because of initial and final state radiation many events contain at least one to two jets. Additionally the production cross section of quark or gluon pairs is very high in hadron collisions. Another problem appearing is the difficulty to trigger on events with jets. Therefore all efforts to search for the decay $H \rightarrow b\bar{b}$ with the ATLAS detector concentrate on Higgs production modes with associated particles. This thesis will concentrate on the production of a Higgs boson in association with a vector boson V . This production mode is also called Higgs bremsstrahlung since the Higgs boson is radiated off a W or Z boson. The leptonic decays of the bosons: $W^\pm \rightarrow l^\pm \bar{\nu}_l$, $Z \rightarrow l^+ l^-$ and $Z \rightarrow \nu \bar{\nu}$ have a signature with at least one high- p_T lepton or high E_T^{mis} which allows for efficient triggering. Concerning

the lepton just muons and electrons are taken into account because the τ -lepton decays before it enters the detector which is again difficult to trigger on. The diagram for the WH production mode is shown in fig. 2.6a. This looks similar for a Z instead of the W bosons with the corresponding Z decay. However, as shown in fig. 2.6b, a search in this production mode has the drawback of a low cross section of approximately 1 pb for pp -collisions with a center of mass energy of 7 TeV. This is approximately one order of magnitude smaller than the production of the Higgs boson alone. For higher center of mass energies this plot looks similar with a spectrum shifted to higher cross sections. Nevertheless the luminosity of the LHC is large enough that a search in this channel becomes reasonable despite its small cross section. The main backgrounds for this channel are top pair and single top production since a top quark decays into a bottom quark and a W boson, W/Z production with additional bottom quarks and multijet production. All these background processes have significantly higher cross sections than $V, H \rightarrow b\bar{b}$. This and the problem to measure jet energies accurately will make it difficult to achieve a clear Higgs signal in this channel. [1, 8, 12, 15]

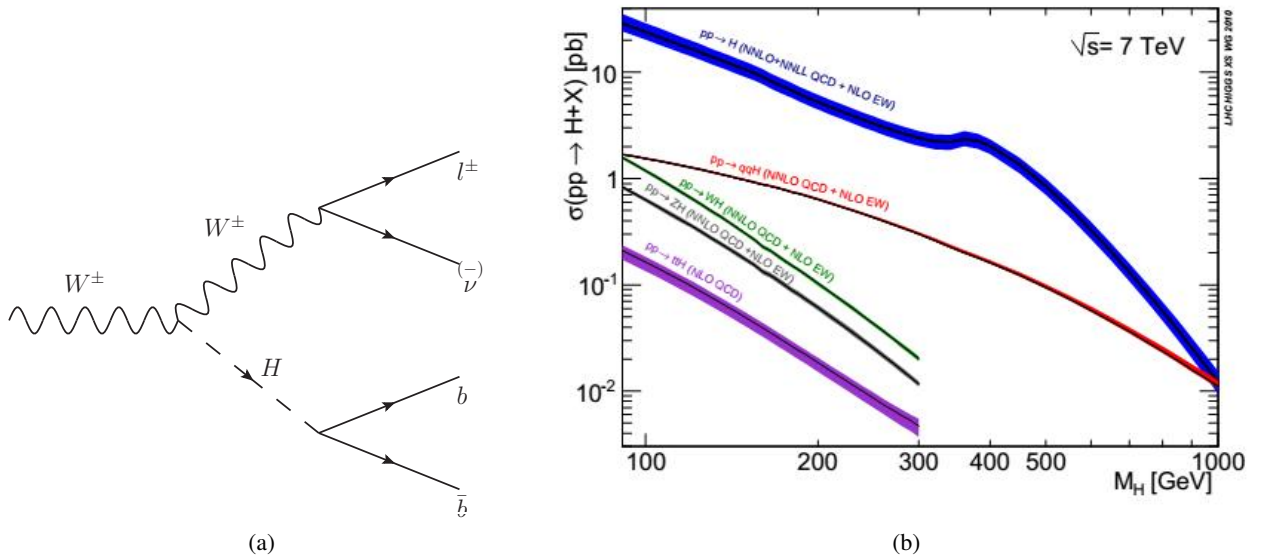


Figure 2.6: (a) $H \rightarrow b\bar{b}$ one lepton decay channel: Higgs production in association with a W boson, (b) cross sections for different Higgs production modes for a center of mass energy of 7 TeV [15].

2.6 Physics of Jets

As already mentioned in chapter 2.1 confinement does not allow quarks to exist freely as they are not color neutral. This can be understood by the nature of the strong coupling constant α_s . For large distances between two color charged objects α_s increases. In a high energetic collision the produced quarks recede very fast from each other and the potential energy stored in the field between them increases. At some point this energy is large enough to form a new quark-antiquark pair. The primary quark might build a color neutral object with those new quarks. If the energy of the quarks is still too large this process will repeat until all quarks are grouped to color neutral objects. This process of hadronization cannot be described by theory, but there exist several models which are used in the generation of Monte Carlo samples (see chapter 3.1). After the hadronization many of the formed hadrons decay due to their

short lifetime. Therefore a quark or gluon manifests as a bundle of hadrons correlated in their direction of flight. [16]

In the ATLAS detector jets are detected in the hadron calorimeter. The hadrons which form the jet lose most of their energy in interactions with the nucleus' of the detector material. During this so-called spallation processes new particles are produced, e.g. photons from disexcitation processes which lead to electromagnetic subshowers in the hadron shower or neutrinos which escape the detector undetected. As a result hadron induced showers are very inhomogeneous and reach deep into the detector. The hadron calorimeter is designed such that all hadrons deposit their whole energy in the calorimeter. Therefore the energy of them is measured. Due to the difficult structure of jets and the different response of the detector depending on the particle the resolution of the jet energy measurement is limited. Additional reconstruction uncertainties worsen the jet energy resolution further. In the ATLAS detector jets are reconstructed in the calorimeter in two steps: first the signals in the calorimeter are clustered together and then a calibration is applied. The clustering itself is separated into two steps. The first step forms so-called topological clusters from calorimeter cells. A cell is a read-out channel of the calorimeter and its size depends on its location in the detector. The idea of topological clustering is to reconstruct three-dimensional energy depositions by adding neighboring cells with an significant energy compared to the energy of noise originating from electronics. Each cluster is defined as a massless pseudo-particle ($E = |\vec{p}|$) with a four momentum ($E, \vec{p}$) and direction η and ϕ pointing from the interaction point to the energy weighted center of the cluster. There are two sorts of clusters: electromagnetic clusters and locally calibrated clusters. The clusters differ in their calibration of the calorimeter response to an energy deposit. The electromagnetic cluster's energy is measured assuming an electromagnetic object (electron, photon) going through the detector whereas the locally calibrated clusters are calibrated to a pion, i.e. a hadron, going through. After the cluster finding a jet finding algorithm will form jets from these clusters. The commonly used jet finding algorithm is the anti- k_t algorithm [17]. It is based on relative distances d_{ij} between objects with a four momentum and analyses them according to their transverse momentum squared. This is compared to the transverse momentum squared relative to the beam d_i . If the transverse momentum of an object i is high enough so that $d_i < d_{ij}$ it will be defined as a new jet. If this is not the case the objects i and j will be combined to a new object. This method has the advantage that low- p_T objects do not influence the jet shape but high- p_T objects do. The size of the jets is controlled via the distance to which the calculated distances are normalized. After the jet finding some further calibrations and corrections are applied since the measured energy of the jet still differs from the true jet energy. The reasons for this are: systematic errors in the measurement of the detector, parts of the jet lay in dead regions (no active detector material) of the detector, parts of the jet which lay not in the jet cluster found by the jet finding algorithms, calibration of the calorimeter cells to specific particles, etc. The main goal of this thesis was to find a good way to correct the energy of b -jets regarding these effects in order to obtain a better input for the Higgs boson reconstruction [6, 8, 17].

2.6.1 Identification of b -Jets

A crucial point in the analysis of $H \rightarrow b\bar{b}$ decays is the reconstruction and identification of jets originating from b -quarks. There are techniques to distinguish the flavors of the jet to a certain extent. The identification of b -jets is called b -tagging. There are several algorithms which rely on the fact that jets originating from the hadronization of a b -quark contain a B -meson. A B -meson consists of a b -quark and another quark and has a rather long lifetime of $\tau \approx 10^{-12}$ s [18]. The flight length is $L = \beta\gamma c\tau \approx \beta\gamma 10^{-4}$ m with $c = 3 \times 10^8$ m s $^{-1}$ and the relativistic factors $\beta = v/c \approx |\vec{p}|/E$ and $\gamma_r = (\sqrt{1 - v^2/c^2})^{-1} \approx E/m$. Often the B -mesons have transverse momenta of several 10 GeV which leads to a flight length of several mm. Distances of this order can be resolved in the tracker of the AT-

LAS detector. The point in space where the B -meson decays is called secondary vertex and is displaced from the interaction point — the primary vertex. The secondary vertex is reconstructed from the tracks in the inner detector and its distance to the primary vertex is calculated afterwards. This property, and dependent on the tagging algorithm, further information which have discriminating power between the jet flavors are used to calculate a b -tagging weight between 0 and 1. This weight corresponds to a probability of identifying a certain percentage of the b -jets as such. The price to pay for a large amount of b -jets, i.e. a cut on a large weight, is less purity. Since there is some overlap between high- p_T c -jet and low- p_T b -jet signatures some c -jets will be misidentified as b -jets as the flight length of particles depend on their momentum. Hence a good working point for the b -tag weight has to be found, identifying as many real b -jets as possible without selecting too many misidentified jets. In this thesis the MV1 tagging algorithm is used [19]. MV1 is a multivariate algorithm based on a neural network. The input for the neural network are the weights from several other tagging algorithms. The neural network uses this information to provide a new b -tagging weight. MV1 yields a better performance compared to a single, cut-based tagging algorithm since it is a combination of three tagging algorithms [8, 19, 20].

Technical Introduction

3.1 Monte Carlo Generators and Samples

An important tool for modern high energy particle physics are Monte Carlo generated data samples. Monte Carlo generators model a certain physics process and produce artificial data which are an image of what will happen in the collision and will be recorded in the detector afterwards. Most often the generators share the event reconstruction and analysis framework with the data analysis. This allows for detailed Monte Carlo studies to develop an analysis strategy, study the effects of certain steps in the analysis and find the best background rejection. This thesis was mainly a Monte Carlo study to develop a method to correct the b -jet energy and to test this method afterwards. Initially Monte Carlo algorithms are randomized algorithms which is a good feature for elementary particle physics since the outcome of a single measurement is not predictable. It is just possible to give a probability for a certain outcome. Generators use (pseudo)random numbers to make choices intended to reproduce the probabilities for different outcomes at various stages of the process. The stages of a process are considered sequentially. For each step a set of rules is defined to iteratively construct the final state of a process. The final state may be very complex containing several hundred particles each one with many degrees of freedom (momentum, flavor, mass, etc.). The stages for the generation of one event generated from a single proton-proton collision are [21]:

1. The incoming protons are described as collection of partons using pdfs.
2. One parton from each proton collides with each other in a hard interaction and gives the process of interest. This can be calculated by theory.
3. In case a short-lived particle, e.g. a $W^\pm$ or Higgs boson, was produced, it decays.
4. The two colliding partons might produce initial state radiation. They might radiate off more than one gluon and these gluons may split to quark-antiquark pairs which might again radiate off gluons. A cascade of partons is triggered. This is also called parton shower and has to be modeled because it cannot be calculated.
5. The outgoing partons or decay products of the short-lived particle may produce final state radiation which also has to be modeled by parton showers.

6. The proton remnants may interact with each other, produce further particles and also produce ISR and FSR.
7. The proton remnants form color neutral states.
8. The receding color charged, "free" partons stretch a "color field" between their color charge and the matching anticolors. The energy stored in this field will grow if the partons recede further due to the nature of the strong coupling. The color fields or tubes cannot be described by theories and must be modeled.
9. If the energy of the color field is high enough they will break up and produce a quark-antiquark pair. These group with quarks from other break-ups to form color neutral, primary hadrons. This is called hadronization and has to be modeled since it cannot be predicted by theory.
10. Many of the primary hadrons are unstable and decay further and their decay is recorded in the detector. Some of them live long enough to directly produce a signal in the detector. At this stage the event generation and the detector simulation meet and have to be matched.
11. The experimental information is obtained and used to reconstruct what happened at the initial state, e.g. production of a Higgs boson.

Figure 3.1a shows some of these stages exemplary for a $H \rightarrow b\bar{b}$ decay.

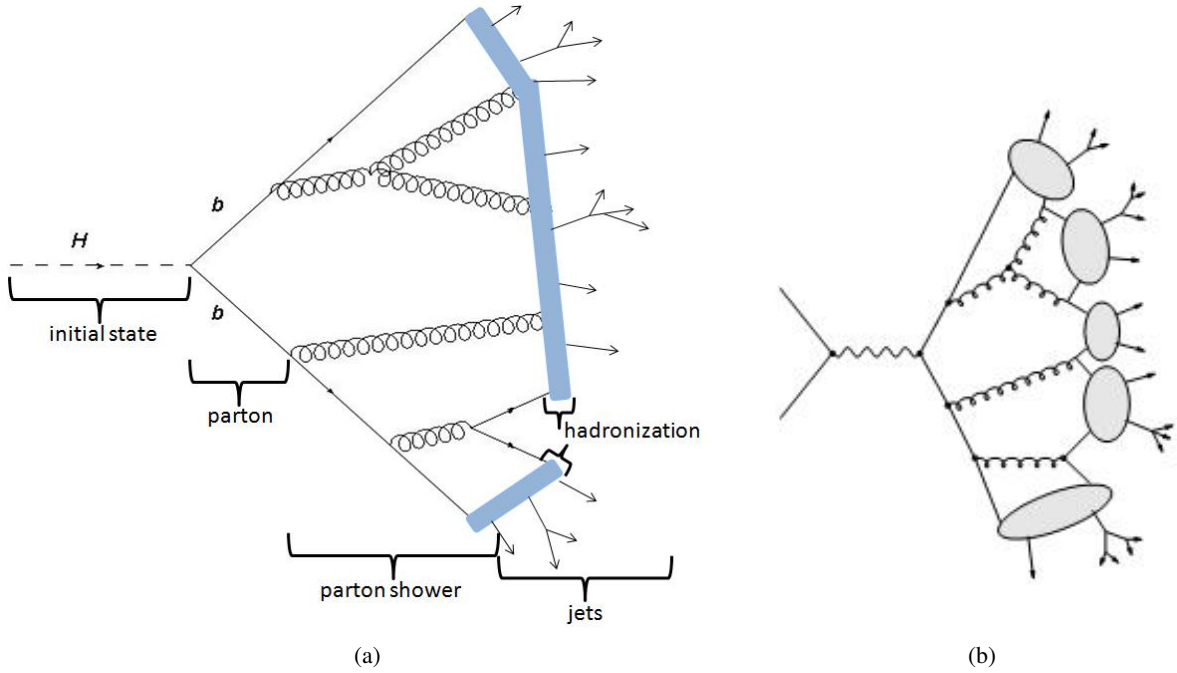


Figure 3.1: Shown is the fragmentation of quarks to jets for (a) the string fragmentation (b) the cluster fragmentation [22]. (a) Shows the fragmentation of $H \rightarrow b\bar{b}$ and the stages on which information can be provided in the Monte Carlo samples.

Processes of the form $2 \rightarrow 2$, i.e. two particles coming in and two particles going out can be predicted very well from theory. The cross sections are calculated from the kinematic properties of the particles, the coupling of the force of their interaction and in the case of partons the pdfs. But this process is just

the lowest order one. Processes of the form $2 \rightarrow 3$ or processes with intermediate loops may also happen and probably should also be studied. Including those processes (called next to leading order — NLO — processes) in the calculation allows a more accurate prediction of the outcome. Also NNLO, NNNLO, ..., processes may happen and including those in the calculations would lead to a further improvement of the prediction. The problem is that the number of integrals which has to be solved to calculate the cross section increases exponentially with every higher order. Therefore a good balance between the accuracy of the prediction of the process under study and the feasibility of computing those processes has to be found. Another challenge the event generator has to face is the modeling of IS and FS parton showers and the hadronization process. The parton shower which is a $2 \rightarrow n$ process is factorized into simple processes like $2 \rightarrow 2$ and then sequentially calculated step-by-step. From one step to the next the momentum transfer decreases until the theory description breaks down. At this point hadronization takes over. [21] There are two common models to describe hadronization: cluster and string fragmentation. Cluster fragmentation splits all gluons into quark-antiquark pairs. Then the quarks are grouped into clusters which undergo isotropic decay into hadrons. String fragmentation works with the tubes/fields which are stretched between the color and the anticolor. The energy density is assumed to be homogeneous along the tubes which form hadrons by breaking up into quark-antiquark pairs. In this model gluons give rise to kinks in these tubes. A schematic comparison between the two models is shown in fig. 3.1. Which particles can be found in the final state and their momentum distributions depend on the fragmentation model. Both models have their advantages and disadvantages and it depends on the generator which model is used. A generator which works with cluster fragmentation is e.g. HERWIG and a generator which works with string fragmentation is e.g. PYTHIA. [21, 22] In this thesis the signal samples $V, H \rightarrow b\bar{b}$ were produced with the PYTHIA event generator. Figure 3.1a also shows the stages for which truth particle information are present in the Monte Carlo samples. This was important for the choice of variables for my correction method (see chapter 5.3). The background samples were produced with various generators.

3.2 The Data

The available data were recorded with the ATLAS detector. They are divided into two sub-sets. Each sub-set contains the data of one working period under stable conditions. The first sub-set contains data from the operation in the year 2011. This data was recorded at a collision energy $\sqrt{s}$ of $\sqrt{s} = 7 \text{ TeV}$ and the integrated luminosity $\mathcal{L}$ is $\mathcal{L} = 4.7 \text{ fb}^{-1}$. The integrated luminosity is the luminosity integrated over time and is according to formula 2.4 directly proportional to the number of events. In high energy physics $\mathcal{L}$ is often measured in inverse barn[b^{-1}]; $1 \text{ b} = 10^{-24} \text{ cm}^2$. The second sub-set contains the collected data from the year 2012 and was recorded at a collision energy of $\sqrt{s} = 8 \text{ TeV}$. The integrated luminosity for this period of data taking is $\mathcal{L} = 20.3 \text{ fb}^{-1}$. The two sub-sets differ in their collision energy which also has an effect on various processes such as pile-up. Furthermore some issues on reconstruction, calibration, trigger and analysis changed. Therefore the two sub-sets of data cannot be treated the same. The Monte Carlo samples are also separated into 2011 and 2012.

3.3 Analysis Frameworks

One major issue when analyzing collider data and LHC data in particular is the complexity of the analysis which should be performed and the large amount of data which has to be handled. Therefore every analysis is performed within a so called analysis framework. An analysis framework is a collection of code providing basic utilities and services which are required in nearly every high energy physics

analysis, e.g. a four momentum object or a class to create and store histograms. This means that the basic infrastructure for the analysis is ready to use but can be individualized afterwards. The utilities and services may be combined to complete analyses. Furthermore a framework assures that everyone working with it works on common definitions. One widely used framework (and many other framework rely on it) in high energy physics is ROOT. ROOT is a C++ based framework which is dedicated to high energy physics problems. It is the most popular high energy physics framework and provides a large amount of services. Since this thesis is performed within the Overkill framework, ROOT version 5.34.00 was mainly used for storing and providing Data/MC in a ROOT format. [23]

3.3.1 Overkill

Overkill is an analysis framework which was developed at the "Physikalisches Institut" in Bonn mainly by the persons involved in $H \rightarrow b\bar{b}$ analyses. It is written in C++ and resembles Athena which is a widely used ATLAS analysis framework. Overkill is designed to work on D3PD files (see section 3.4) but could in principle be expanded to other formats because it obtains physical objects and their required information for the analysis from ROOT trees. If certain branches should be used the framework will take care of activating them and storing their values. In Overkill groups of branches are represented as objects or vectors of objects. The framework already provides common physics objects like electrons, jets, vector bosons, etc. as C++ classes. An analysis within Overkill is created by combining different analysis modules. These modules work with the physics objects obtained from the root tree and every module executes a certain step of the analysis, e.g. selecting jets, correcting electrons, etc. The functionality of the modules is split into tools. Depending on the complexity of the analysis step the module should perform, more or less tools are necessary. But for example many modules need tools to create the physics objects to work with. The whole analysis is configurable via text files. Every module and tool has configurables whose values are determined in this text file. These configurables allow to personalize the analysis because e.g. it is possible to define cut values, the tools which are used by the different modules, etc. In the configuration file one also defines the prefixes of the branches in the root tree which should build a physics object. Therefore it is possible to use e.g. jets from different clustering algorithms (if they are available in the root tree) and it is very easy and fast to switch between them. Overkill also provides a graphical user interface (GUI) which allows the user to change the configuration values. Furthermore the GUI can be used to create basic analysis skeletons and browse histograms. Overall Overkill is able to perform $H \rightarrow b\bar{b}$ dedicated analyses well. Since Overkill just activates certain ROOT branches containing the necessary objects for the analyses and in a selection process only stores indices of the objects instead of making copies of the objects the framework is fast. In addition the high flexibility provided by the configurable module and tool structure which also allows to make changes in a short time are advantages. The greatest advantage is that Overkill is designed to work on customized, i.e. just the necessary information containing, locally stored D3PDs. Therefore all analyses run on local clusters without the necessity to use the Grid¹ as it is necessary for Athena based analyses with common N-tuples. This results in a much faster execution of the analysis [24], This master thesis was performed within Overkill using basic tools. I created a new Overkill module which should prepare the training of the BDT algorithm and then correct b -jets by using multivariate regression (for more information about multivariate techniques, see chapter 4).

¹ A network of computing centers especially for working with large amounts of LHC data.

3.4 The Bonn D3PDs

The way to provide data from the ATLAS detector as well as Monte Carlo samples is an important point since the recorded data and the Monte Carlo predictions contain huge amounts of information. Not all of these information are necessary or they cannot be used for analysis directly. The most common data format for analyses is the Derived Physics Data or shortly D3PDs. In those files the physics objects, e.g. jets, muons, secondary vertices, are already reconstructed and the information about them, e.g. energy, mass, is assigned to them. The Data and Monte Carlo samples which were used in this thesis are in the D3PD format and were especially produced for the Bonn $V, H \rightarrow b\bar{b}$ analysis. The Bonn D3PDs are compatible to Overkill. They were produced from so called AODs (a data object with more information and lower level of object complexity) provided by the ATLAS collaboration. The D3PDs were produced with Athena which is a high energy physics analysis framework and contains common object definitions and instructions to calculate physical properties. This makes sure that everyone works on common data/Monte Carlo samples. Since the instructions within Athena on how to reconstruct objects and assign properties to them change regularly a new processing of data becomes necessary from time to time. Therefore there are different production cycles/releases of the Bonn D3PDs which correspond to the Athena version which was used to produce them. Before every new production cycle of D3PDs it had to be decided which objects and properties are required for the analysis and are stored in the D3PDs [24].

3.5 The Standard $V, H \rightarrow b\bar{b}$ Analysis

This section will give a short overview of the standard techniques which are used to search for the Higgs boson in the decay into a bottom-antibottom-quark pair in associated production with a W boson. The standard analysis uses two different approaches to reject events from background processes: one relies on cuts on different variables which have characteristic distributions for the desired Higgs decay and the other approach uses a machine learning algorithm which categorizes events as either background- or signal-like based on several input variables. The event and object selections for this thesis are based on the common selections of the analysis. The analysis for the channel with an associated W boson requires exactly one lepton (muon or electron) which has to be tight, two b -tagged jets which were tagged with the MV1 tagging algorithm at a 70% working point, a maximum of one additional jet, a missing transverse energy larger than 25 GeV and a transverse mass of the W boson smaller than 120 GeV. The two latter cuts are varied depending on the transverse momentum of the W boson. Additional criteria for the transverse momenta of the two b -tagged jets require for one jet $p_T > 45$ GeV and for the other one $p_T > 20$ GeV. The tracks associated to the jet and originating from the primary vertex have to carry more than 75% of the jet's energy. A table with these cuts is shown in 3.1. The jet momentum is corrected with the " $\mu + p_T^{reco}$ " correction (see section 5.11.1) afterwards. There are also corrections for the momenta of the other reconstructed objects. After the object selection further topological cuts are defined for different regions of p_T of the vector boson. For the channel with an associated W boson these are cuts on the minimum and maximum distance of the two b -jets from each other, the minimum missing transverse energy and the minimum and maximum transverse mass of the W boson. For all object selections pseudorapidity η cuts are set depending on the coverage of the detector components which are necessary to reconstruct those objects. The whole analysis is performed on the full data set from 2011 and 2012. For the modeling of the background processes Monte Carlo samples from various generators were used. Some of them are scaled with additional scale factors. The considered background processes are: W/Z +jets, $t\bar{t}$, single top and VZ . The background from multijet production

is estimated from data. Using this approach and including systematic uncertainties no signal was seen yet.[25]

The machine learning algorithm which is used in this thesis to derive the correction for the b -jet energy may only be used if every object it uses for building its prediction model is selected in the same way. Otherwise a new correction has to be derived for every set of cuts. Therefore this thesis uses a reduced set of the cuts of the standard analyses for practicability reasons. This reduced selections assure that the correct jets are chosen and that the important features of the decays' signature are considered. The object definitions, e.g. the clustering and reconstruction algorithms for the jets and the quality criteria (loose, medium, tight) for the leptons, are the same as in the standard analysis. The detailed object and event selection for this thesis can be found in chapter 5.5.

Object	Cut
jets	exactly 2 b-tags (MV1, 70% working point)
	$p_T^{\text{jet1}} > 45 \text{ GeV}$
	$p_T^{\text{jet2}} > 20 \text{ GeV}$
	jet vertex fraction $> 75\%$
	$ \eta < 2.5$
lepton	≤ 1 additional jet
	exactly 1 tight lepton
missing transverse energy	$E_T^{\text{mis}} > 25 \text{ GeV}$
W boson	$m_T^W < 120 \text{ GeV}$

Table 3.1: Basic object and event selections for the standard $V, H \rightarrow b\bar{b}$ analysis [25]

Multivariate Analysis

4.1 Multivariate Techniques

At hadron colliders the challenge is to separate a few events of the physics process under study from a lot of background events and to measure the properties of those events as precise as possible. Due to the complexity of this large amount of data it is desirable to find a good technique to deal with these challenges and at the same time make use of the lot of information. A multivariate analysis which is based on machine learning algorithms is able to achieve this. The general idea of a multivariate analysis is to use a machine learning algorithm to predict the outcome of a measurement. For this the algorithm has to be trained on a set of data where the outcome and certain features (i.e. physical quantities) of the set are known. With this training data the algorithm builds a prediction model which should be able to use the equivalent features of an unseen data set to predict its outcome. The predicted outcome can be quantitative — this technique is called regression — or categorical — this is called classification. For this thesis the regression technique is used. Therefore the following descriptions will focus on regression.

The assumption for regression is that an outcome y , also called "target", is related to p features x_i , also called "input variables", via a function $f(\mathbf{x})$ with $\mathbf{x} = (x_1, x_2, \dots, x_p)$ [26]:

$$y = f(\mathbf{x}) + \epsilon \quad (4.1)$$

where ϵ is a random error term. In general the function $f(\mathbf{x})$ which connects the inputs with the outcome is unknown and has to be predicted based on n observations. These observations $(\mathbf{x}_1, y_1)$, $(\mathbf{x}_2, y_2)$, ..., $(\mathbf{x}_n, y_n)$ from the training data are presented to the learning algorithm. Based on this the learning algorithm will build a prediction model $\hat{f}(\mathbf{x})$ which is an approximation for $f(\mathbf{x})$ and will give an estimate for the outcome $\hat{y}$ [26]:

$$\hat{f}(\mathbf{x}) = \hat{y} \approx y \quad (4.2)$$

The deviation between $\hat{y}$ and y stems from two sources: The first is the random error term ϵ (see formula 4.1) which is irreducible. This error term represents the intrinsic inaccuracy of measurements. The second source is the so-called "loss". This loss can originate from providing too little information for the learning algorithm or the learning algorithm is not flexible enough to recognize the relationship between $\mathbf{x}$ and y . This error is reducible and the algorithm will try to make its estimate $\hat{f}(\mathbf{x})$ such that the loss between y and $\hat{y}$ is minimal. This is done point wise for given values of the inputs. The most

common loss is the squared error loss: $(y - \hat{y})^2$. Hence the best estimate is found if the mean of the squared errors (*MSE*) [26]:

$$MSE = \frac{1}{n} \sum_{j=1}^n (y_j - \hat{f}(\mathbf{x}_j))^2 \quad (4.3)$$

is minimal. The final form of the estimate depends on the learning algorithm, i.e. the way the learning algorithm works to find the best estimate. Since the algorithm tries to minimize the loss the estimate also depends on the functional form which is assumed for the loss. There are various algorithms and loss functions. The choice of the most suitable algorithm depends on the problem the learning algorithm has to predict. The assumptions which are made for the form of f do not have to be an explicit function. These so called non-parametric methods have the advantage that they are more flexible in determining the shape of f since they are not restricted to a certain assumption. The drawback is that these methods need in general more training data to make a good estimate for f . One of these non-parametric methods are decision trees which might be combined with boosting (see sec. 4.2). This is the method which is used for this thesis. The large amount of observations in high energy physics which are available in Monte Carlo samples are good inputs for this method. The large amount of observations is also beneficial for avoiding so-called overtraining. Overtraining appears when the algorithm tries to minimize the deviation between the true value y and the estimate $\hat{y}$ and is not restricted to a certain functional form. If there are too few observations the algorithm can easily make up a shape for $\hat{f}(\mathbf{x})$ such that it will match all observations $(\mathbf{x}_j, y_j)$. The less observations are given in the training sample the easier it is to make up any shape for $\hat{f}(\mathbf{x})$ which fits all observations. Doing so the algorithm catches up statistical fluctuations which are unique for the training sample. If this overtrained prediction model $\hat{f}(\mathbf{x})$ is now applied to a sample with the same underlying relations between y and $\mathbf{x}$ the deviation between the true value and the predicted value would be much larger than the minimal one from the training. The reason is that every sample has its own unique statistical fluctuations [26].

4.2 Boosted Decision Trees (BDTs)

Boosted decision trees (BDTs) is a learning algorithm based on a very simple idea. This algorithm builds one decision tree for the first iteration and more trees are added in the following iterations, called boosting. A schematic view of one decision tree is shown in fig. 4.1. The growing of the tree starts, i.e. building of the prediction model, at the root node where the tree splits the training sample into two regions based on a cut value c_i for an input variable x_i . After this there are two subsamples: one for $x_i > c_i$ and one for $x_i < c_i$. Those two subsamples get transferred to the next node where each of them is again separated based on some cut value for some input variable. The input variable and the cut value which are used may change from node to node. The split at each node is determined by finding the variable and the cut value which give the best estimate for the outcome y . Splitting the training sample into smaller and smaller subsamples is continued until a certain stop criterion is fulfilled. This can either be a minimum number of observations left in each subsample or a maximum number of nodes or a maximum depth of the tree. Also other stop criteria are conceivable but the above mentioned ones are the most common. The last nodes where the tree stops to grow, so-called output nodes, represent estimated values $\hat{y}_j$ for the real values y_j . Decision trees are very good "out-of-the-box" estimators. Due to their relatively simple structure of decision taking only little tuning is necessary to give reasonable results. They are also able to ignore variables which do not have a large prediction power. This makes decision trees very robust estimators. The drawback is that their accuracy is not as high as for other estimators. One technique to significantly improve the accuracy of the prediction

model of decision trees is boosting. The basic idea is that the information from one decision tree is used to grow another decision tree. This reuse of information for new trees can be repeated several (hundred) times. All boosting algorithms modify the function $f(\mathbf{x})$ which has to be estimated by the next decision tree. The easiest way of boosting is to reweight the observations in the training sample proportional to the deviation between the estimate $\hat{y}$ and the true value y . This reweighting is done every time before growing the new tree and relies on the estimates from the last tree. Hence more worse estimated values get a higher weight and therefore gain importance in the next tree whereas well estimated values get a smaller weight and are not estimated again. This is called adaptive boosting. The old tree together with the new tree which used the reweighted observations usually provide a better estimate. Therefore the accuracy of the prediction model will sequentially increase with every new tree. But the algorithm will reach an accuracy limit asymptotically. At that point adding new trees will not gain more accuracy. Tests showed that in general small but many trees give the best results. For this thesis a special boosting algorithm — gradient boost — was used. Gradient boost is a more robust algorithm compared to adaptive boost when dealing with long-tailed distributions for the error loss distribution and extremely mismeasured values of y . [26–28]

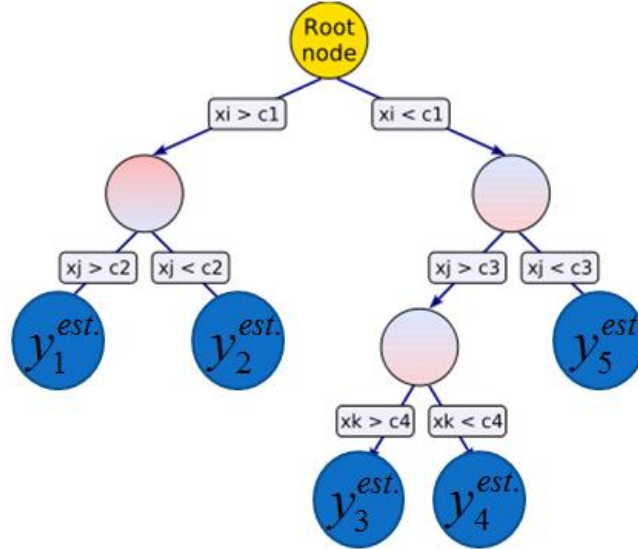


Figure 4.1: Schematic view of a single decision tree; it performs cuts on the input variables x_α and based on that gets to different output node which represent different estimates $y_i^{est.}$ for the output variable y [28].

4.3 The TMVA Framework

TMVA is an open source framework for multivariate analysis and provides a ROOT-integrated environment. The framework is designed for high energy physics and this thesis was performed with it. The data which should be used are handed over to TMVA as ROOT trees. The whole training process is managed by a factory which provides an interface between the user and the TMVA analysis. The user specifies to the factory which input variables should be used, which variable is the target (in case of a regression task), from which tree the algorithm should get its information and which training algorithm should be used. In addition the user can change several algorithm specific options. In preparation of the training the factory splits the input data in a training and a test sample. Then the algorithm is trained just on the data in the training sample and later the trained algorithm is tested on the test sample. During

the training the factory creates several outputs such as a ranking of input variables and average variance between the regression target and the estimation for training and test sample which helps the user to assess the performance of the algorithm. The result of the training is a so-called weight file which gives rules how the target may be determined from the input variables. How these rules look like depends on the training algorithm. The weight file is interpreted by the TMVA reader which manages the application process. The reader needs to know which weight file it is supposed to use, which algorithm was used for the training and which variables in the data set, you want to apply your regression to, correspond to the used input variables of the training. With this information the reader calculates event by event a new value for the target variable one wants to predict from the measured input variables [28].

For this thesis the TMVA implementation of BDTs with gradient boosting was used. The trees are grown minimizing the squared error loss. There are several configuration options to tune the BDTs. The most important ones are the number of trees which should be grown and the stopping criterion for the growth of the single trees. The stopping criterion can be either a maximum number of observations left in the subsamples, a maximum depth or a maximum number of nodes. The tree stops to grow if one of those criteria are reached.

***b*-Jet Energy Correction Using Multivariate Regression**

5.1 Motivation

The idea to use multivariate regression to correct measurements in high energy physics is relatively new. Especially the high amount of data collected at colliders is a good statistical basis to train multivariate algorithms. Two experiments already applied a multivariate method for the correction of their *b*-jet energy successfully — CDF [29] at the Tevatron collider and CMS [30] the other high luminosity experiment at the LHC. This master thesis tries to improve the Higgs mass resolution in the decay channel $W^\pm H \rightarrow l^\pm \bar{\nu} b \bar{b}$ by using similar regression techniques. The idea to use the nominal Higgs mass as a target for the regression would not lead to the desired result which was tested in a bachelor thesis before [31]. Since the mass distribution is a discrete function at the nominal Higgs mass the multivariate algorithm learns that only the nominal Higgs mass is the right answer and creates a mass bias. Therefore this master thesis pursues the CDF and CMS approach to improve the *b*-jet energy resolution since the jet energy has a continuous spectrum. This improves the Higgs mass resolution indirectly since the Higgs mass is reconstructed from the four momentum vectors of the *b*-jets by calculating their invariant mass. CMS claims approximately 15% improvement for their invariant di-jet mass resolution in the $V, H \rightarrow b \bar{b}$ channel using this approach [30]. The four momentum of the *b*-jets is calculated from the measured energy, the measured direction of the jet in η and ϕ as well as the mass which is calculated as the invariant mass of the calorimeter clusters which build the jet. As already mentioned in chapter 2.6 the energy of *b*-jets is not measured very well due to measurement and reconstruction uncertainties. In this chapter I will show that the energy is indeed the worst measured parameter of the *b*-jet four momentum and has to be corrected. Therefore the transverse energy of the initial *b*-quark was chosen as a target for the regression. I will also motivate a set of 30 input variables for the training which consider effects which are related to inaccuracies in the energy measurement. The samples which were used for the training of the BDTs simulate the $W^\pm H \rightarrow l^\pm \bar{\nu} b \bar{b}$ channel with different Higgs masses. Object definitions as well as object and event selections are based on the standard ATLAS analysis for this channel. Due to problems with the truth information of the *b*-quarks in the samples for the 2012 data taking period this thesis could only be performed for the 2011 period of data taking. For a deeper understanding of the regression the influence of the variables, the choice of the target and the training performance will be studied in some detail in this chapter. As a result I will present the resolution of

the invariant mass of the regression corrected *b*-jets, i.e. the reconstructed Higgs mass. This will show that the regression does not only narrow the Higgs mass peak but will also correct the peak position. Since the decay channel has to deal with a lot of background a narrow Higgs peak is desired because it will stand out of the background more. Also the peak position of the peak will be important due to the unknown Higgs mass from theory and to support the results considering the likely discovery of the Higgs boson in 2012 [10]. As a comparison also other methods to improve the Higgs mass resolution will be discussed. This will show that the achievable result for the regression approach which corresponds to an improvement of over 30% in the Higgs mass resolution compares very well with other methods.

5.2 Training Samples

The goal of the regression approach is to improve the Higgs mass resolution in the decay channel $W^\pm H \rightarrow l^\pm \nu^{(-)} b \bar{b}$. The corresponding Monte Carlo samples which are used for the training were all generated with the PYTHIA Monte Carlo generator. There are different samples for the 2011 and the 2012 data taking periods which also differ in the used version of PYTHIA — version 6 and version 8 respectively. In addition there are different releases for those Monte Carlo samples available, meaning that the D3PDs for them are produced in different cycles with different Athena versions (see 3.4). Because the BDTs build the prediction model based on the information presented in the training samples it is important to think about which of the available Monte Carlo samples should be used. There are three crucial points: the truth information (especially the target) should show the desired behavior, all necessary information for the input variables must be available and the statistics (i.e. the number of events) must be high enough to avoid overtraining (see 4). The training samples which were used for this thesis were from release 17.0.6.4.7 and represent the 2011 period of data taking and are therefore generated with PYTHIA 6. Table 5.1 shows the features of the different available $V, H \rightarrow b \bar{b}$ Monte Carlo samples. For the releases 17.0.6.4.7 and 17.2.7.5.14 the truth information is available before final state radiation (FSR), i.e. the *b*-quarks direct after Higgs decay. For release 17.2.7 it was tried to reconstruct the initial *b*-quark four momentum vector by assigning gluons to the *b*-quark after FSR in the same way as it was done for release 17.0.6.4.7. But from PYTHIA 6 to PYTHIA 8 major changes were made, in particular color reconnection was included. This affects the parton showering for FSR with respect to PYTHIA 6 [32]. Therefore the algorithm within Athena for assigning gluons to the *b*-quark is not sufficient anymore. The result is a diluted Higgs peak in the invariant mass spectrum of the two truth partons (see fig.5.1b) and therefore this information could not be used to train the BDTs. In contrary the invariant mass of the two truth partons before FSR for the 2011 LHC working period from release 17.0.6.4.7 shows the desired δ -peak structure for all available $W^\pm H \rightarrow l^\pm \nu^{(-)} b \bar{b}$ Monte Carlo samples which contain different values for the nominal Higgs mass (fig. 5.1a). Another problem with the 17.2.7 release is the absence of the variables for the secondary vertex. These variables are available in the samples of release 17.0.6.4.7. The newest release (17.2.7.5.14) which was produced for the 2012 period of data taking contains the whole decay chain of the event including the initial *b*-quarks which have to be found manually by going through the decay chain. The secondary vertex variables as well as all other important variables for the training are also available in this release. But this release was only available at the end of this thesis which made it impossible to use it for this thesis. Therefore the whole analysis was performed with release 17.0.6.4.7 and all results are shown for the 2011 period of data taking.

release	generator	LHC working period	partonic truth information	available variables for jets
17.0.6.4.7	PYTHIA 6	2011	before+after FSR	jet jet related/track soft lepton secondary vertex jet with cone radius $\Delta R = 0.6$ E_T^{mis}
17.2.7	PYTHIA 8	2012	after FSR	jet jet related/track soft lepton jet with cone radius $\Delta R = 0.6$ E_T^{mis}
17.2.7.5.14	PYTHIA 8	2012	before+after FSR	jet jet related/track soft lepton secondary vertex jet with cone radius $\Delta R = 0.6$ E_T^{mis}

Table 5.1: Available Monte Carlo samples for $V, H \rightarrow b\bar{b}$ and their features regarding the LHC working period, the generator, the truth parton information and the available jet variables.

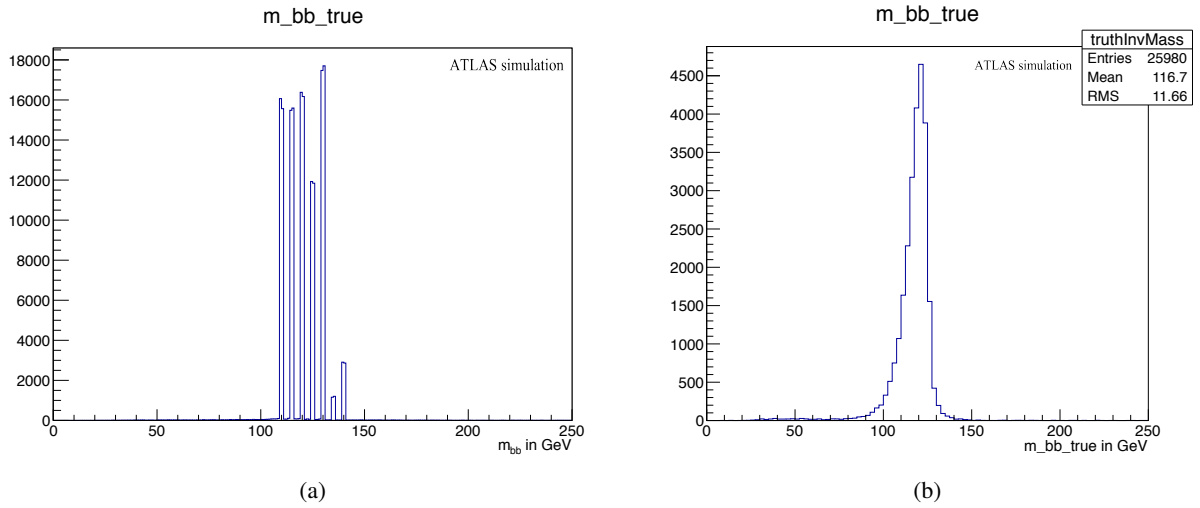


Figure 5.1: Shown is the invariant mass of the two truth b -quarks before FSR for (a) all available $W^\pm H \rightarrow l^\pm \nu b\bar{b}$ Monte Carlo samples for the 2011 LHC data taking period from release 17.0.6.4.7 and (b) for a $W^\pm H \rightarrow l^\pm \nu b\bar{b}$ Monte Carlo sample with $m_H = 125$ GeV for the 2012 LHC data taking period from release 17.2.7. Changes in the PYTHIA event generator lead to a diluted Higgs peak in the latter one.

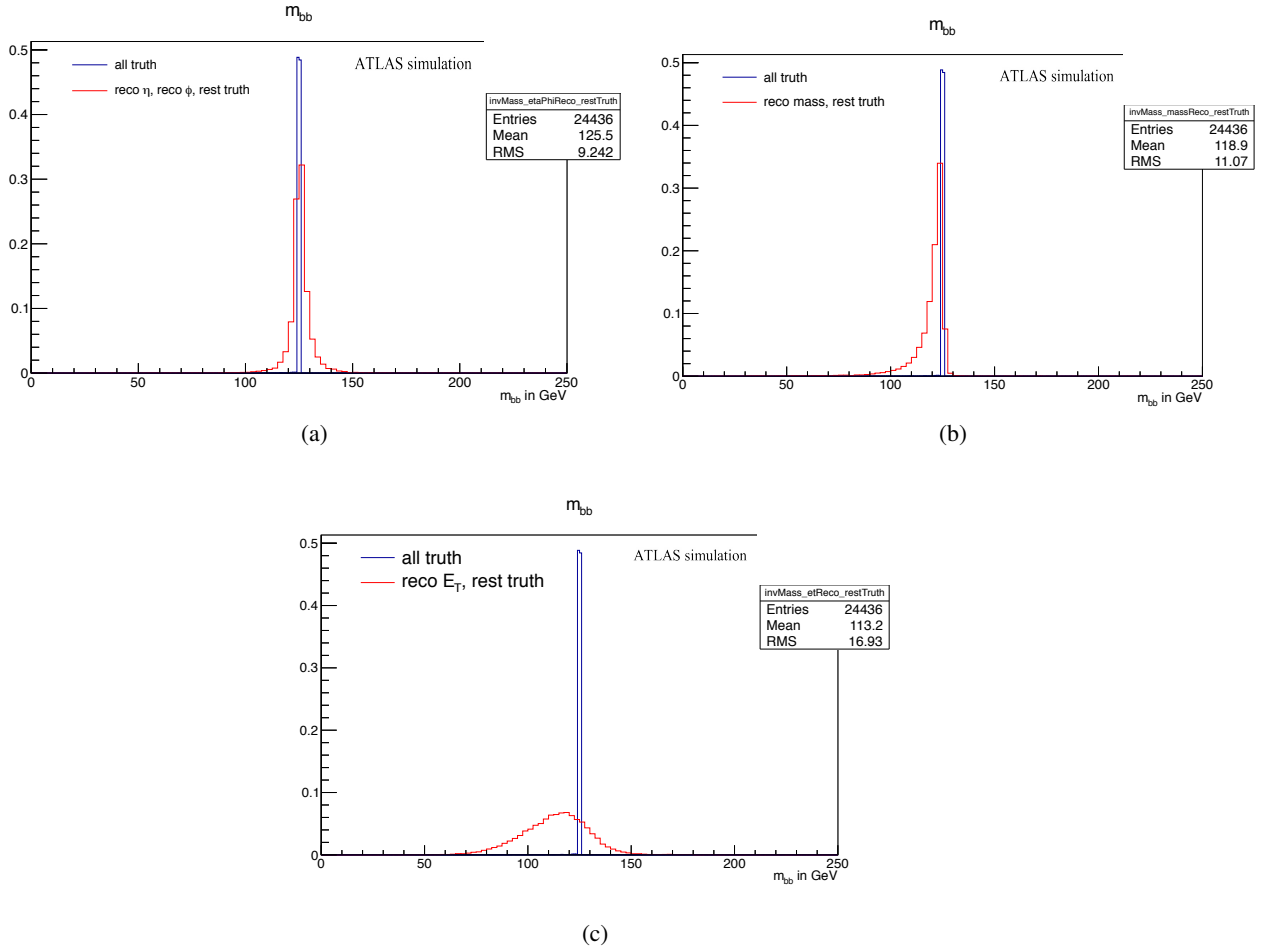


Figure 5.2: $m_{b\bar{b}}$ on truth level compared to the invariant mass of the di- b -quark system where truth η and truth ϕ (a) or truth m (b) or truth E_T (c) were replaced by the corresponding reconstructed quantities

5.3 Target Variable

The aim of the regression based jet energy correction is to have an improved resolution of the Higgs mass. Since the Higgs mass is the invariant mass of the two b -jets ($m_{b\bar{b}}$) it has to be decided which component of the jet four momentum vector should be corrected. The best choice should be the component which was measured worst. Also it has to be decided on which stage the target variables should be calculated. Because the truth information contains in principle the whole decay chain of a physics process, the four momentum vector of the jet is available at every stage of the decay (see 3.1). Again the argument is that the indirect aim is the improvement of the $m_{b\bar{b}}$ resolution. Therefore the truth information should be as close to the information about the initial Higgs boson as possible hence the two b -quarks directly after the Higgs decay are used (see 5.2). A good way to determine the worst measured variable and at the same time see which variable has the most influence on the dilution of the Higgs mass peak is to calculate the invariant mass of two truth b -quarks for whom one component in their truth four momentum vector was replaced by a measured quantity of the corresponding jet. Figure 5.2 shows the $m_{b\bar{b}}$ distributions for $W^\pm H \rightarrow l^\pm \nu^{(-)} b \bar{b}$ with a nominal Higgs mass of 125 GeV if the angles η and ϕ , the mass m or the transverse energy E_T are exchanged in the truth four momentum vector respectively. A

good measure to get a rough estimate for the $m_{b\bar{b}}$ resolution is the ratio between the root mean square root (RMS) and the mean (μ) of the distribution. On truth level the resolution is in this case just limited by the chosen binning which causes that the discrete peak is distributed between two bins. A meaningful RMS value cannot be given but the difference between the truth peak and the other peaks is visible in the given figures. It is easy to conclude that E_T has the most influence on the $m_{b\bar{b}}$ resolution since the resolution is decreased to 15% with replaced E_T . For replaced angles or mass the resolutions just drop to approximately 7% and 9% respectively (the detailed table A.1 is given in the appendix). Figure 5.2c also shows that the reconstructed transverse energy is underestimated since the peak position of the $m_{b\bar{b}}$ distribution is shifted to smaller values. Based on the invariant mass spectra shown in fig. 5.2 it was decided to use the transverse energy as a target for the regression.

5.4 Input Variables

The choice of the variables which are used in the regression is important for the result of the regression since the goal is to use BDTs to build a prediction model based on measured parameters (input variables) and their relation to the truth b -quark energy directly after the Higgs decay (target). Therefore it is important to think about which measured variables are related to the jet energy and which effects influence the measurement of the b -jet energy to find a good set of input variables. The base of this thesis were the variable sets which were used by the CMS and CDF collaboration for their regression based b -jet energy correction. Table 5.2 shows the variables which are used as input variables in the CDF/CMS regression and which are considered in my regression approach. Since the CMS regression is based on the CDF regression the variable set of the CMS regression is similar to the CDF variable set with small enhancements. My set of variables includes the majority of the CDF/CMS variables and further variables which account for additional effects.

5.4.1 Physical Motivation of the Input Variables

The physical effects which play a role in the measurement of b -jets can be split into six categories as it is also done in tab. 5.2. This section should motivate how these categories are related to the measurement of the b -jet energy and which input variables are developed from them.

Jet variables: The most intuitive variables are the reconstructed jet kinematics itself since they are directly connected to the jet energy and the jet energy may be derived from them. The jet variables contain all sorts of variables which may be used to construct a four momentum vector of the jet or can be derived from it (energy, mass, momentum and the corresponding transverse quantities and angles). Additional information is provided by so called raw variables. These variables are the jet kinematics derived on the electromagnetic scale before any corrections or calibrations are applied to the properties which are measured in the calorimeter. Therefore the difference between them and the reconstructed quantities indicate how well the jet was measured.

Jet related/track variables: About 2/3 of the energy of a b -jet is carried by charged particles. Since charged particles are also measured in the tracker which has a better resolution than the calorimeters the difference between these variables and the reconstructed jet quantities can (like the raw quantities) indicate how well the jet was measured.

Soft lepton variables: These variables take into account that the b -quark in the jet can decay into another quark with the emission of a virtual W boson. This is called a semileptonic decay because the virtual W boson decays into a lepton and a neutrino afterwards. Such decays change the

structure of the jet and due to the neutrino a fraction of the energy remains unmeasured. Therefore this has an impact on the reconstructed jet energy. Especially for the decays into a muon an additional fraction of the energy escapes the calorimeter unmeasured. A decay into an electron has also an effect because the electron forms a subshower in the jet. The leptons from such semileptonic decays are indirect decay products of the *b*-quarks and are therefore kinematical connected to the *b*-jet. The larger the *b*-quark energy the larger the lepton momentum and the smaller the distance to the jet axis. The branching ratio for such a semileptonic decay of the *b*-quark is approximately 20% [18]. The information about the escaping neutrino, i.e. the missing transverse energy, cannot be used in the $W^\pm H \rightarrow l^\pm \bar{\nu} b \bar{b}$ decay channel since it is indistinguishable from the missing transverse energy from the neutrino from the *W* boson decay from the Higgs boson production.

Secondary vertex variables: The secondary vertices are reconstructed from tracks. Therefore they provide an additional measurement of the track properties. Information about the flight length, i.e. the distance of the *b*-vertex to the primary vertex, may also be used because the flight length is proportional to the *b*-quark energy due to relativistic time dilatation (see section 2.6).

Jet with cone radius $\Delta R = 0.6$ variables: For the $V, H \rightarrow b \bar{b}$ analysis the standard jets which are used are reconstructed with a cone radius of $\Delta R = 0.4$. But in the Bonn D3PDs information are also provided for jets with a cone radius of $\Delta R = 0.6$. In most cases $\Delta R = 0.4$ is a good choice because most of the energy of the jet is contained in the corresponding cone and not much energy which is not part of the jet is picked up. Nevertheless a part of the jet energy is leaking out of this cone every time, e.g. low energetic charged particles may be bended out of the jet cone, parts of extensive showers may lay outside the cone or the jet may be of a very asymmetric shape. Variables which are derived for jets with $\Delta R = 0.6$ give information about the energy which leaked out of the cone and was not measured for the $\Delta R = 0.4$ jets.

E_T^{mis} variables: If the regression should also be translated to $H \rightarrow b \bar{b}$ decays in associated production with a *Z* boson which decays into leptons afterwards this variables could be added to the training because this process contains no neutrino(s) from the Vector boson decay. Hence the missing transverse energy in the event should either come from inaccurate measurements and the biggest inaccuracies originate from the jet energy measurement in the calorimeter or the missing transverse energy originates from a neutrino from a semileptonic *b*-quark decay. In both cases the missing transverse energy is a measure for a non-measured energy part of the jet.

5.4.2 Choice of a Set of Input Variables

In my approach I first tried to implement all variables which are used by CDF and CMS. But some CMS variables are not available for ATLAS Monte Carlo samples/data because CMS uses a particle reconstruction technique which is called particle flow. CDF also did not use particle flow which is the main reason for the differences in the variable sets of CMS and CDF. The aim of the particle flow is to use information from every subdetector to identify and reconstruct every single, stable particle. Hence in the case of jets single particles and their corresponding momenta which contribute to the jet are known. ATLAS is not using particle flow in their standard analyses. Therefore the only information available for jets is the total energy of all jet constituents¹ which is measured in the calorimeter. The tracks reconstructed in the inner detector may be assigned to the jet and therefore information about

¹ except for contributions from muons and neutrinos in the jet because they escape of the detector, see section 2.3

Variable category	Variable	CMS	CDF	My Analysis
jet	p_T	✓	✓	✓
	raw p_T	✓	—	—
	E_T	✓	✓	✓
	raw E_T	—	✓	✓
	E	—	—	✓
	m_T	✓	✓	✓
	m	—	—	✓
	ϕ	—	—	✓
	θ	—	—	✓
	η	—	—	✓
	MV1	—	—	✓
jet related	lead. track p_T	✓	✓	✓
	$\sum_{\text{tracks}} p_T$	—	✓	✓
	$RMS(\eta(\text{tracks}))$	—	—	✓
	$RMS(\phi(\text{tracks}))$	—	—	✓
	cef	✓	—	✓
	N (constituents)	✓	—	—
	N (tracks)	—	—	✓
	JEC uncertainty	✓	—	—
soft lepton	ρ_{25}	✓	—	—
	p_T	✓	—	✓
	p_T^{rel}	✓	—	✓
	$\sum_{\text{leptons}} p_T$	—	—	✓
	$\sum_{\text{electrons}} p_T$	—	—	✓
	$\sum_{\text{muons}} p_T$	—	—	✓
	$\Delta R(\text{jet}, \text{lep})$	✓	—	✓
secondary vertex	L_{xy}	—	✓	—
	ΔL_{xy}	—	✓	—
	L_{3D}	✓	—	✓
	ΔL_{3D}	✓	—	✓
	p_T	✓	✓	✓
	m	✓	—	✓
jet with cone radius $\Delta R = 0.6$	p_T	—	—	✓
	m	—	—	✓
	E	—	—	✓
	$\Delta R(\text{anti-kt4 jet}, \text{anti-kt6 jet})$	—	—	✓
$Z(\rightarrow l^+l^-)H$ specific	E_T^{mis}	✓	—	✓
	$\Delta\phi(\text{jet}, E_T^{mis})$	✓	—	✓

Table 5.2: List of the input variables for the regression and which of them are used for the CDF/CMS regression and which are considered for the regression approach of this thesis [29, 33]

the number of tracks and their momenta are available. Hence in ATLAS there is no direct information available about the neutral particles in the jet. Two other variables are not accessible via the Bonn D3PDs: the JEC uncertainty which describes the uncertainty on the jet correction and ρ_{25} which is an event property which describes the average density of transverse energy per unit area in the detector. I also did not use some other variables (L_{xy} , ΔL_{xy} and raw p_T) which seemed redundant since they only differ slightly between CMS and CDF. After the examination of the CMS and CDF variables I added additional variables to include more effects which caused inaccurate measurements of the *b*-jet energy. In contrast to the CMS and CDF approach I did not use only the transverse quantities of the jet but also the absolute values of the energy and the mass, and the angles η , θ and ϕ . I also added a *b*-tagging weight (MV1) variable since it depends on the kinematics of the *b*-jet (see chapter 2.6). For the jet related properties I mainly stuck to the CMS and CDF variables which are based on the tracks associated to the jet. These variables take into account the number of tracks, the sum of the transverse momenta of the tracks, the p_T of the track with the highest p_T (lead. track p_T) in the jet and the fraction between the sum of the transverse momenta of the track and the transverse momentum of the whole jet ("charged energy fraction" = "cef"). New are the variables $RMS(\eta(\text{tracks}))$ and $RMS(\phi(\text{tracks}))$ which are a measure for the width of the jet in the $\eta\phi$ -plane. They provide information about the *b*-jet energy because jets become broader with higher energies. The semileptonic decay variables considered the transverse momentum of the lepton, the distance between the lepton and the jet axis and the transverse momentum of the lepton relative to the jet axis (p_T^{rel}). All of them are just used for the lepton closest to the jet axis. p_T^{rel} is just defined for muons because electrons are detected via their showers in the calorimeter and are overlaid by hadronic showers. Therefore non-isolated electrons do not have a very well defined direction. I introduced three new variables $\sum_{\text{leptons}} p_T$, $\sum_{\text{electrons}} p_T$ and $\sum_{\text{muons}} p_T$. They distinguish between electrons and muons from semileptonic decays since those two have very different signatures in the detector which influences the measurement of their properties and the connection to the *b*-jet energy. The new variables also take into account multiple semileptonic decays which can happen for approximately 4% to 5% of the *b*-jets [18]. For the secondary vertex variables I used the same set as CMS including the mass, the transverse momentum and the flight length and its error. Especially the flight length provides new information since the flight length is proportional to the momentum of the *b*-quark (see chapter 2.6). The last set of variables relevant for the $W^\pm H \rightarrow l^\pm \bar{\nu} b \bar{b}$ decay channel are the variables for jets with a cone radius of $\Delta R = 0.6$ instead of $\Delta R = 0.4$. This is a new set of variables which was neither used by CDF nor by CMS. It will be shown below (see section 5.12) that they provide new information to further improve the regression performance. The information which are used are the transverse momentum, the mass and the energy of the larger jets as well as the distance between the standard jet and the larger one. This distance is bigger for larger out of cone leakages and asymmetric jets. All variables together lead to a set of 30 variables as an input for the training (compare tab. 5.2). The BDTs are able to ignore variables if they provide no new information. Hence if redundant variables are used the performance of the regression will not decrease. This will be examined closer in section 5.12. Theoretically there are two additional variables E_T^{mis} and $\Delta\phi(\text{jet}, E_T^{mis})$ which may be used for a training in the $ZH \rightarrow l^+ l^- b \bar{b}$ channel. But due to lack of statistics in the Monte Carlo samples for this channel those additional variables were not examined in detail (see section 5.9). The introduced set of 30 input variables will be used throughout the whole thesis because it covers the most important effects on the *b*-jet energy measurement.

5.5 Event and Object Selection

For the training of the BDTs several objects are needed. These objects have to be selected carefully to avoid picking up misreconstructed objects. In addition a basic event selection based on the $W^\pm H \rightarrow l^\pm \bar{\nu} b\bar{b}$ signature is used to reject events from background processes. To assure compatibility with the standard $V, H \rightarrow b\bar{b}$ analysis common objects definitions as well as a reduced set of the standard object and event selection cuts are used. Especially the b -jets which should be corrected have to have similar kinematic properties as the b -jets for the standard analysis.

Object	Cut
b -jets	exactly 2 $MV1 \geq 0.601713$ (70% working point) $p_T > 20 \text{ GeV}$ jet vertex fraction $> 75\%$ $ \eta < 2.5$ ≤ 1 additional (non b -tagged) jet
jets with cone radius of $\Delta R = 0.6$	$p_T > 20 \text{ GeV}$ jet vertex fraction $> 75\%$ $ \eta < 2.5$ $\Delta R(\text{anti-}k_t 6 \text{ jet}, \text{anti-}k_t 4 \text{ jet}) < 0.2$
lepton in jet	loose+combined muon loose electron $\Delta R(\text{lepton}, b\text{-jet}) < 0.4$
secondary vertex (sv)	$N(\text{tracks}) \geq 2$ $\Delta R(\text{sv}, b\text{-jet}) < 0.4$
W boson	exactly 1

Table 5.3: Selections for the b -jet and the objects used in the regression

The jets which are used as an input for the training are reconstructed using the anti- k_t clustering algorithm with topological electromagnetic clusters and a cone radius of $\Delta R = 0.4$ (for a description of jet reconstruction see chapter 2.6). These jets have to lay in a pseudorapidity interval of $-2.5 < \eta < 2.5$ since the ATLAS tracker just covers this range and information about the tracks associated to the jet should will be used in the regression. In addition its important to make sure that the jets originate from the hard-scatter interaction, in this case the Higgs production and decay, and not from underlying events. Therefore the jet vertex fraction of the jets should be larger than 0.75 meaning that the tracks which originate from the primary vertex and are associated to the jet carry at least 3/4 of the energy of all associated tracks. Since the Higgs boson is assumed to have a mass of approximately 125 GeV the decay products, i.e. the b -jets, should have a high transverse momentum. Based on this assumption it is required that the jets have a minimal transverse momentum of 20 GeV. The asymmetric b -jet p_T from the standard $V, H \rightarrow b\bar{b}$ cuts are not used since all b -jets which are used in the training have to be treated the same by the algorithm to achieve a consistent correction for all b -jets. Therefore only the looser cut of $p_T^{\text{jet}} > 20 \text{ GeV}$ is used. Since the regression should derive a correction for b -jets a b -tagging cut is set. For this the MV1 b -tagging algorithm is used with a working point of 70% which assures that 70% of all b -jets are identified as such. In addition the jets undergo an overlap removal which removes jets

if there are reconstructed objects with a significant energy close by ($\Delta R < 0.4$) which also showered in the calorimeter. When the regression will be applied to the *b*-jets the same cuts are used.

The jets which were selected in the way described above are the base for the regression and all other objects for the regression are associated to them. First the corresponding truth *b*-quarks whose E_T is the target variable have to be found. A truth *b*-quark is assigned to a jet if the distance ΔR between the quark and the jet axis is smaller than 0.4. For the variables which consider the out of cone leakage jets are used which are clustered in the same way as the base jets but with a cone radius of $\Delta R = 0.6$. They undergo the same cuts except for the *b*-tagging cut. These jets are assigned to the base jets if the difference between the jet axes is smaller than $\Delta R = 0.2$. For the variables which consider leptons from semileptonic decays in the jet the leptons have to have a maximum distance of $\Delta R = 0.4$ to the jet axis. Both, electrons and muons, have to be at least loose. The definition of a loose lepton is chosen such that objects which were reconstructed as leptons but are presumably no real leptons are rejected (for a detailed definition of a loose lepton see [34]). In addition the muons have to be combined, i.e. were reconstructed in the inner detector as well as in the muon spectrometer. A secondary vertex is assigned to a jet if it has more than one track and has a maximum distance to the jet axis of $\Delta R = 0.4$.

The event selection requires exactly two *b*-tagged jets, exactly one leptonically decaying *W* boson and not more than one additional jet. The *b*-tagged jets are selected with the criteria mentioned above. The third jet in the event has to fulfill the same criteria as the *b*-jets except for the *b*-tagging cut. The *W* boson is identified via the lepton which has to be tight. The definitions of tight follow [25] and require essentially a high transverse momentum and the isolation of the lepton. Isolation means that in a cone of defined radius around the lepton just a small amount of energy must be measured. Table 5.3 gives a summary of the above mentioned selections.

5.6 Preparation of the Samples, Variables and the BDTs

Before the training starts the information which will be used by the BDTs have to be prepared and the BDTs have be configured. It was decided that the *WH* PYTHIA samples from the 2011 period of data taking (release 17.0.6.4.7) will be used (see section 5.2). Additionally it was decided that the sample which was generated for a nominal Higgs mass of 125 GeV will not be used in the training since the regression will be tested on that. The exclusion of this sample from the training makes sure that no bias will be created. This leaves samples with Higgs masses of 110 GeV, 115 GeV, 120 GeV, 130 GeV, 135 GeV and 140 GeV which amount to an input for the training of 215 054 *b*-jets with an associated truth *b*-quark. This is enough statistics for a reasonable training result. Some of the 30 input variables are derived from objects associated to the jet. It might happen that in some cases no associated object can be found either due to reconstruction or physical reasons. A typical case for reconstruction issues are the jets with the larger cone since in for every jet the same jet just with a larger cone should be found. The reason for not associating a lepton to every jet has a physical background since semileptonic decays only happen for 20% of the jets. Table 5.4 gives the percentages of cases for which no object of a certain category could be assigned to the jet. Per definition a truth *b*-quark is found for every jet since this is the criterion for filling a jet into the training. If no object of a certain category can be associated to the jet its corresponding variables will be filled with the "dummy" value of -1 . This value makes no sense physical wise but the jets and the variables may still be used in the training. This is important to maintain a reasonable training statistic. The BDTs are capable of ignoring these unphysical values to some extent.

The BDTs are configured with a default configuration given by the TMVA developers. This configuration showed good performance in tests of the developers [35]. The amount of nodes per tree is five

Category	Number of associated objects $N_{\text{assoc.}}$	$N_{\text{assoc.}}/N_{\text{jets}}$ in %
truth b -quarks	215054	100.00
associated tracks	215040	99.99
min. 1 lepton in jet	56825	26.42
secondary vertex	25460	11.84
jets with cone radius $\Delta R = 0.6$	210413	97.84

Table 5.4: Objects which have to be associated to the jets and the absolute and relative amount of successful associations.

which limits the depth of the tree. Each input variables is divided into 20 equally large regions which defines the granularity of the cuts the BDTs can make. The boosting algorithm is gradient boost and the amount of trees are 400. Hence the algorithm works with many but small trees (see chapter 4.2). In the last step the 215 054 b -jets are split into two equally large samples: the training and the test sample. This splitting is done randomly. The prediction model will be built based on the training sample. Since the test sample is unknown to the BDTs it is suitable for evaluating the training's performance.

5.7 Training Results: the Jet E_T Resolution

The results of the training are new estimates for the transverse energies of the jets. The width of the distribution of the differences between the jet E_T before and after the regression and the true (partonic) E_T is a good measure to assess the influence of the regression on a single b -jet. Figure 5.3 shows these distribution before (blue) and after (red) the regression. Besides it shows this for the test tree (lines) and training tree (crosses) separately to get an estimate for the overtraining.

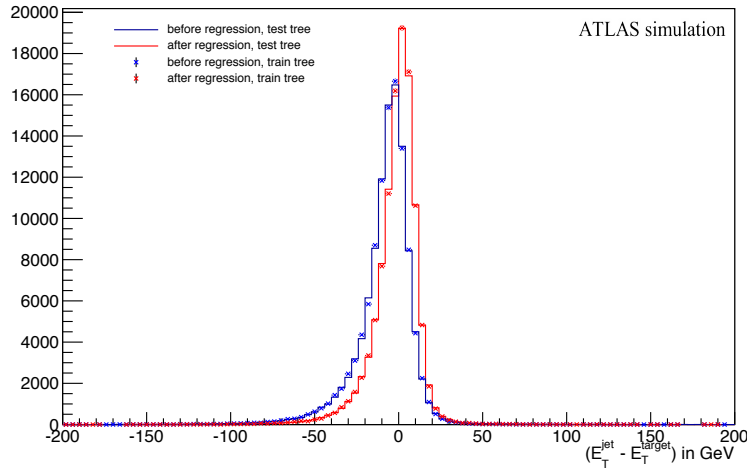


Figure 5.3: Shown is the difference between the jet E_T and the truth parton E_T (target) for the reconstructed jets (blue) and the jets with the estimated E_T from the training (red) separated for the training (crosses) and the test (lines) sample. The training was done on $W^\pm H \rightarrow l^\pm \nu b \bar{b}$ samples with all available Higgs masses except for $m_H = 125$ GeV for the 2011 working period.

The distributions for the test tree is used to estimate the improvement of the regression since the

test tree is an independent sample not used during the training phase but with the same features as the training sample. The RMS of those distribution improves from (17.36 ± 0.04) GeV before the regression to (14.58 ± 0.03) GeV with the regression corrected jets. This corresponds to an improvement of 16%. The figure also shows that the mean which should be at zero theoretically improves from an offset of (-8.36 ± 0.05) GeV to (-1.67 ± 0.04) GeV. This improvements will translate implicitly to the $m_{b\bar{b}}$ resolution. From the good agreement of the results for the training and the test tree it may be concluded that the effect of overtraining is low. The RMS of the training tree distribution after the regression is (14.42 ± 0.03) GeV which corresponds to a deviation of approximately 1% to the result on the test tree. This is slightly higher than the difference between the RMS of the training tree and the test tree before the regression which originates from statistical fluctuations. That leaves the conclusion that this degree of overtraining is tolerable. The detailed results are given in table A.2 in the appendix A.2.

5.8 Application of the Trained Model

After the training the prediction model built by the BDTs may be applied to *b*-jets in any sample. The application is done jet wise. Based on the measured variables of the jet the jet retraces the cuts for the given BDTs grown during the training. Thus a new estimate for the transverse energy of the jet is found. Hence only the measured input variables are necessary to find new jet E_T estimates. This makes this method applicable to data. With the new estimate for the transverse energy of the *b*-jet ($E_T^{\text{reg.}}$) a new four momentum vector is calculated for the *b*-jet:

$$\begin{aligned} E^{\text{corr.}} &= \frac{E_T^{\text{reg.}}}{\sin \theta^{\text{reco.}}} \\ p_x^{\text{corr.}} &= p_T^{\text{corr.}} \cos \phi^{\text{reco.}} \\ p_y^{\text{corr.}} &= p_T^{\text{corr.}} \sin \phi^{\text{reco.}} \\ p_z^{\text{corr.}} &= \frac{p_T^{\text{corr.}}}{\tan \theta^{\text{reco.}}} \end{aligned} \quad (5.1)$$

with

$$p_T^{\text{corr.}} = \sqrt{E_T^{\text{reg.}} - m^{\text{reco.}} \sin \theta^{\text{reco.}}} \quad (5.2)$$

The effect of applying this to a $W^\pm H \rightarrow l^\pm \nu^{(-)} b\bar{b}$ Monte Carlo sample for the 2011 LHC working period with a nominal Higgs mass of $m_H = 125$ GeV is shown in fig. 5.4. The $m_{b\bar{b}}$ distribution before the regression (blue) is significantly wider than the distribution for $m_{b\bar{b}}$ with the regression corrected jets (red). The difference to the nominal Higgs mass is larger before the regression than after the regression. Bukin functions were fitted to the histograms to determine the resolution. The Bukin function is especially designed for high energy physics problems as it is able to model asymmetric tails. This is achieved by a convolution of a Gaussian function and exponential functions which model the tails separately [36]. There are other measures for the resolution which are discussed in some detail in the appendix A.1. The Bukin function was used in this thesis since this is the standard measure for studies of the $m_{b\bar{b}}$ resolution in the $V, H \rightarrow b\bar{b}$ channel and describes the characteristics of the distributions best (see A.1). From the Bukin fit the peak position x_p and the peak width σ_p is derived. The results for the

resolution σ_p/x_p with their statistical errors are therefore:

$$\begin{aligned} \text{before regression: } \frac{\sigma_p}{x_p} &= \frac{(18.18 \pm 0.16) \text{ GeV}}{(115.40 \pm 0.26) \text{ GeV}} \hat{=} (15.76 \pm 0.14)\% \\ \text{after regression: } \frac{\sigma_p}{x_p} &= \frac{(13.68 \pm 0.11) \text{ GeV}}{(125.32 \pm 0.20) \text{ GeV}} \hat{=} (10.91 \pm 0.09)\% \end{aligned}$$

This corresponds to an improvement in resolution of approximately 31%.

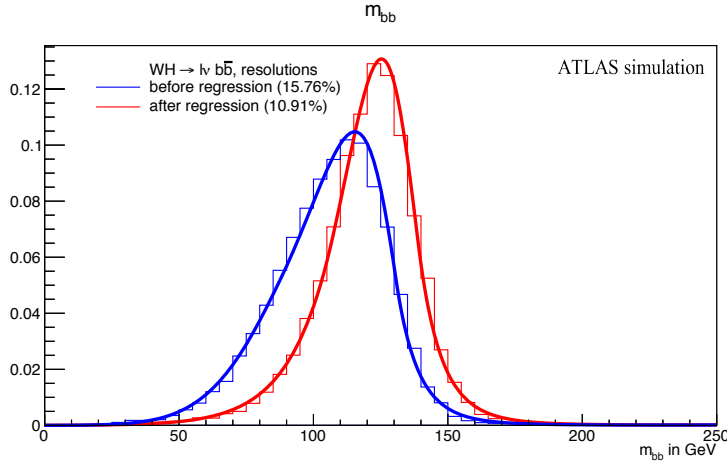


Figure 5.4: Shown is the $m_{b\bar{b}}$ distribution for a 2011 $W^\pm H \rightarrow l^\pm \nu b\bar{b}$, $m_H = 125$ GeV Monte Carlo sample for the reconstructed jets (blue) and the regression corrected jets (red) including the fit of a Bukin function to the histograms. The resolutions are 15.76% before and 10.91% after the regression.

This result may be compared to the CMS result for the $m_{b\bar{b}}$ resolution for the $W^\pm H \rightarrow l^\pm \nu b\bar{b}$ channel which also used a regression approach with BDTs to correct the b -jet energies. This correction was used as a direct orientation for this thesis. Figure 5.5 shows the public CMS regression result for a $W^\pm H \rightarrow l^\pm \nu b\bar{b}$ sample with a nominal Higgs mass of $m_H = 125$ GeV for the 2012 LHC period of data taking. CMS determined their $m_{b\bar{b}}$ resolution from the fit parameters mean (μ) and σ of a fit of a Gaussian function to the core of the distributions. The CMS $m_{b\bar{b}}$ resolution after the regression is [37]:

$$\text{CMS, after regression: } \frac{\sigma}{\mu} = \frac{(13.0 \pm 0.3) \text{ GeV}}{(123.8 \pm 0.2) \text{ GeV}} \hat{=} (10.5 \pm 0.2)\%$$

Although CMS is a different experiment with other features for their jet energy measurement and they use a different event and object selection (e.g. jets with a cone radius of $\Delta R = 0.5$) the resolution results after the regression are very similar to my results. This leads to the conclusion that the regression approach is in general suitable to correct calorimeter measurements for jet energies in high energy collider experiments. Additionally it may be concluded that the sources of inaccuracies in the jet energy measurement seem to be similar in both experiments and that my variable set covers the most important effects to get a good estimate for the true jet energy.

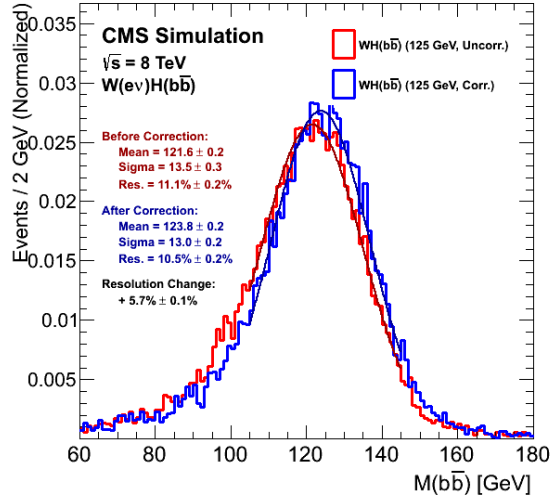


Figure 5.5: Shown is the CMS $m_{b\bar{b}}$ distribution for a 2012 $W^\pm H \rightarrow l^\pm \nu b\bar{b}$, $m_H = 125$ GeV Monte Carlo sample for the reconstructed jets (blue) and the regression corrected jets (red) including the fit of a Gaussian function to the core of the histograms. The resolutions are 11.1% before and 10.5% after the regression. [37]

5.9 Application to non- $W^\pm H \rightarrow l^\pm \nu b\bar{b}$ Samples

Now the regression correction is applied to b -jets in final states other than $W^\pm H \rightarrow l^\pm \nu b\bar{b}$: $ZH \rightarrow l^\pm l^- b\bar{b}$, W production associated with two b -jets and $t\bar{t}$ production since the top-quark decays to bW with a branching ratio of nearly 100%. These applications will prove that the regression is not tuned to the $W^\pm H \rightarrow l^\pm \nu b\bar{b}$ final state.

First the model is applied to the $ZH \rightarrow l^\pm l^- b\bar{b}$ channel. The object and event selection remains the same as for the $W^\pm H \rightarrow l^\pm \nu b\bar{b}$ just that exactly one leptonically decaying Z boson is required instead of a W boson. The result for the application to the $ZH \rightarrow l^\pm l^- b\bar{b}$, $m_H = 125$ GeV sample with a nominal Higgs mass of $m_H = 125$ GeV for the 2011 period of data taking is shown in fig. 5.6. For this channel a large improvement of the resolution of the $m_{b\bar{b}}$ distribution is observed, too. The results for the $m_{b\bar{b}}$ resolution σ_p/x_p with their statistical errors are derived from a Bukin fit to the histograms:

$$\begin{aligned} \text{before regression: } \frac{\sigma_p}{x_p} &= \frac{(19.12 \pm 0.76) \text{ GeV}}{(115.30 \pm 1.20) \text{ GeV}} \hat{=} (16.58 \pm 0.68)\% \\ \text{after regression: } \frac{\sigma_p}{x_p} &= \frac{(14.13 \pm 0.48) \text{ GeV}}{(125.48 \pm 0.80) \text{ GeV}} \hat{=} (11.26 \pm 0.39)\% \end{aligned}$$

The resolution after the regression is within its statistical error equivalent to the resolution for the $W^\pm H \rightarrow l^\pm \nu b\bar{b}$ channel after the regression. Also the improvement in resolution of approximately 32% is similar to the improvement which was achieved for the WH channel. Hence the underlying signature of the event besides the b -jets has a negligible influence on the $m_{b\bar{b}}$ resolution.

Unfortunately a training on the $ZH \rightarrow l^\pm l^- b\bar{b}$ using the two additional E_T^{mis} variables (see tab. 5.2) with an application to this channel afterwards was not possible due to lack of Monte Carlo statistic for the 2011 period of data taking. The amount of b -jets in the ZH samples is just 9283 even with the sample for $m_H = 125$ GeV included. This is too less for a reasonable training which means that the

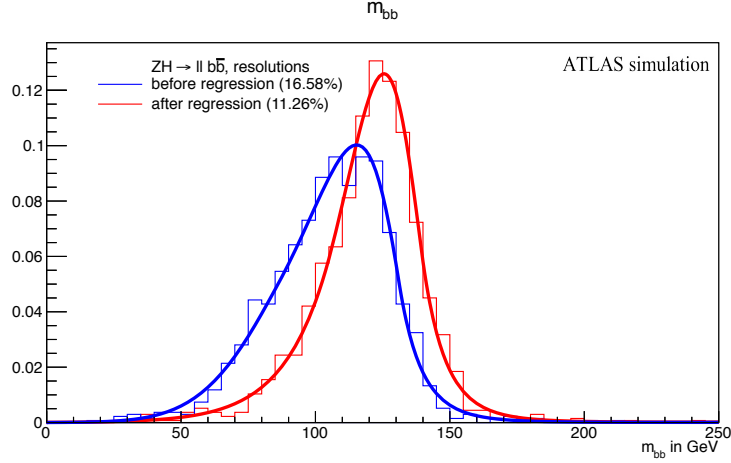


Figure 5.6: Shown is the $m_{b\bar{b}}$ distribution for a 2011 $ZH \rightarrow l^+ l^- b \bar{b}$, $m_H = 125$ GeV Monte Carlo sample for the reconstructed jets (blue) and the regression corrected jets (red) including the fit of a Bukin function to the histograms. The resolutions are 16.58% before and 11.26% after the regression.

BDTs cannot build a suitable model and become sensitive to statistical fluctuations. The overtraining is approximately 20% and therefore not tolerable. The training results are shown in the appendix A.2, fig. A.1.

An application of the trained model to the $ZH \rightarrow \nu \bar{\nu} b \bar{b}$ would also be interesting. The ongoing analyses in Bonn do not consider this channel. Therefore no Bonn D3PDs containing those Monte Carlo samples were available to test the regression.

The application of the trained model to the Wbb sample for the 2011 period of data taking is done with the same object and event selection as for the $W^\pm H \rightarrow l^\pm \bar{\nu} b \bar{b}$ channel since this channel has a similar signature. The application of the regression to this channel is important because the two b -quarks are not connected. They are produced separately and do not originate from a decay. Therefore the m_{bb} distribution is expected to be continuous. If the regression method is biased by the Higgs mass which the BDTs have seen indirectly in the training this will show in the m_{bb} spectrum of the Wbb sample. Figure 5.7 shows the m_{bb} distribution before and after the regression as well as for the truth b -quarks before FSR (green). After applying the regression the whole spectrum is shifted to higher masses since the regression corrects the underestimation of the jet energy. The spectrum on the partonic truth level is even more distributed to higher masses. In addition no artificial peaks arise in the signal region between 100 GeV and 150 GeV after the regression. This leads to the conclusion that the b -jet energy correction is neither tuned to nor bias by the $W^\pm H \rightarrow l^\pm \bar{\nu} b \bar{b}$ channel.

The application to the $t\bar{t}$ -pair production sample is made to examine whether the b -jet energy improvement via regression also improve other measurements containing b -jets such as the top mass measurement. It is also important to examine the effect of the regression on the $t\bar{t}$ sample to see what happens to the peak position. During the training the BDTs just have used samples with b -jets from a decaying particle (Higgs boson) with a maximum mass of $m_H = 140$ GeV whereas the top-quark has a significantly higher mass of $m_t = 173.07$ GeV [18]. In principle the regression should still be able to shift the peak closer to the nominal mass of the decaying particle, the top-quark, since it corrects the underestimation of the b -jet energy. The prediction model which was applied to the top sample was the same as it was for the application to Wbb production and $ZH \rightarrow l^+ l^- b \bar{b}$. The event selection changed to select $t\bar{t}$ pairs for which one W boson decays into a lepton and a neutrino and the other one into a pair

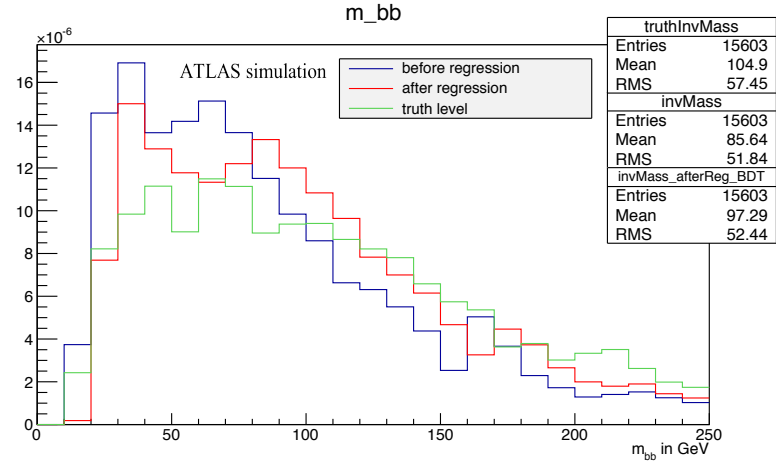


Figure 5.7: Shown is the m_{bb} distribution for a 2011 Wbb Monte Carlo sample for the reconstructed jets (blue), the regression corrected jets (red) and the truth b -partons (green). The regression shifts the whole distribution to higher masses but no peaks are generated.

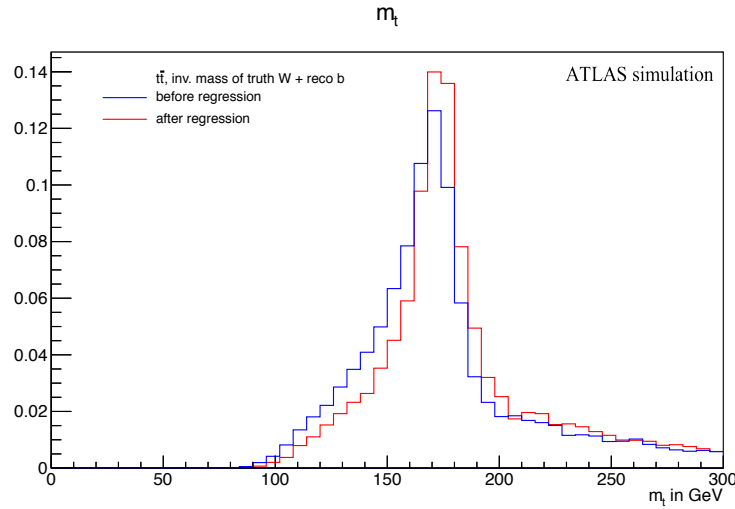


Figure 5.8: Shown is the top-mass distribution reconstructed from the decay $t \rightarrow Wb$ for a 2011 $t\bar{t}$ Monte Carlo sample for the truth W boson + the reconstructed b -jets (blue) and the truth W boson + the regression corrected jets (red).

of quarks. This requires two b -tagged jets, one leptonically decaying W boson and two additional jets (instead of one) from the other W boson decay. For the test of the regression on $t\bar{t}$ events the top-quark was reconstructed from the leptonically decaying W boson and the jet. Because the kinematics of the W boson are reconstructed from its decay products which are measured with certain errors the effect of the regression could be masked by the uncertainties on the measurement of the W boson kinematics. Therefore the decision was made to use the truth information of the W boson. This information was just available for the leptonically decaying W boson which is the reason for not reconstructing just one top quark. Since it is difficult to assign the right b -jet to the top-quark decay where the leptonically decaying W boson originates from, each of the two b -jets was combined with the truth W boson and the invariant mass was calculated. All wrong assignments create a combinatorial background which does

not follow an underlying structure. Therefore this background should be flat and equally distributed over the whole mass range. The resulting top mass distribution with the reconstructed b -jets and with the regression corrected b -jets is shown in fig. 5.8. Due to the highly asymmetric tails and the relatively narrow peak it was not possible to fit a function to this peak. For the same reason the mean and the RMS of the distributions lose their meaning for determining the peak resolution. Therefore only qualitative statements can be made about the effect of the regression. The figure shows that after applying the regression the whole peak is shifted to higher masses and not towards the Higgs mass region. The width of the peak seems to be narrower after applying the regression. Since the peak with the regression corrected b -jets is higher than the other one more events should contribute to the peak. The conclusion is that the b -jet energy correction behaves as expected and is able to correct the energy of b -jets besides the $W^\pm H \rightarrow l^\pm \nu b\bar{b}$ channel.

5.10 Application to Data

It is also important to see the influence of the regression on data. Therefore the regression was applied to the data from the 2011 period of data taking. To suppress the contribution from multi-jet backgrounds an additional cut to the mentioned cuts in section 5.5 was applied which requires the transverse mass m_T of the event to be larger than 40 GeV. The transverse mass is determined from the isolated lepton and the missing transverse energy in the event:

$$m_T = \sqrt{2p_T^{\text{lepton}} E_T^{\text{mis}} (1 - \cos(\phi^{\text{lepton}} - \phi^{E_T^{\text{mis}}}))} \quad (5.3)$$

The data are compared with the Monte Carlo samples for the different underlying physics processes. Since the Monte Carlo generators are not always able to model the underlying physics correctly for some samples additional scale factors were applied. These scale factors were derived from data for selections where it is certain that mainly one physics process is present. In this regions the Monte Carlo samples are multiplied with a simple factor such that they fit the data. The Monte Carlo generators which I used and their scale factors are the same as listed in appendix A.3, tab. A.5 and are motivated by the standard $V, H \rightarrow b\bar{b}$ analysis [34]. The multi-jet background (labeled as "QCD") is estimated from data by running the analysis with an inverted cut on the lepton isolation. The leptons will be isolated if they originate from a vector boson decay in $V, H \rightarrow b\bar{b}$ processes but not for multi-jet processes. This procedure is not accurate and makes it necessary to additionally scale the estimated multi-jet background to fit the data. It was fitted in the transverse mass spectrum after the m_T cut and the scale factors for the Monte Carlo samples were applied. The scale factor for the multi-jet background derived from the fit is 1.321. The results for the $m_{b\bar{b}}$ distribution and a comparison with the Monte Carlo samples for the underlying physics processes is shown before and after regression in fig. 5.9. It is just shown for the mass range from 50 GeV to 200 GeV since there are some modeling problems of the Monte Carlo samples in the lower and higher mass regions. It is noticeable that the scale factors from the standard $V, H \rightarrow b\bar{b}$ analysis are not optimal since there is an overshoot of data at low masses and an undershoot of data for high masses in both histograms. The statistics for the subset of 2011 data is not high enough to see big influences of the regression on the data. What may be observed is that the whole spectrum is shifted to higher masses as it is expected since the b -jet energy is underestimated for the reconstructed jets. It is also important to notice that after the regression no additional peaks appear in the background Monte Carlo samples. For the bins with Diboson (WZ, ZZ, WW) contributions a better agreement of data and Monte Carlo samples is observable after the regression correction is applied. An enhancement of data in the signal region can neither be seen before nor after the regression due to too less data

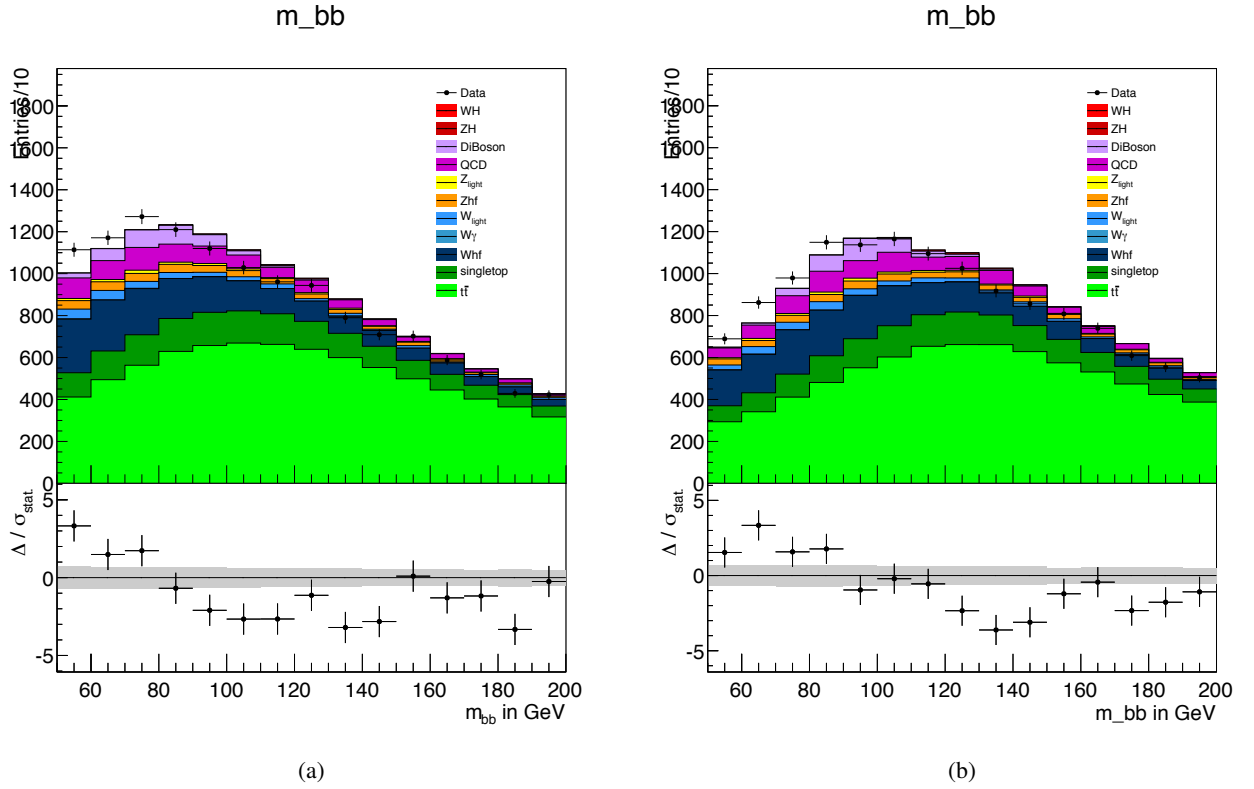


Figure 5.9: Shown is the $m_{b\bar{b}}$ distribution for 2011 data and Monte Carlo samples (a) before regression and (b) after the regression is applied.

statistics.

5.11 Comparison with other Jet Correction Methods

The problem of the inaccurate measurement of the jet energy is well known. Therefore many approaches exist to correct this effect. I tested two corrections myself with my selections on the samples I used — the standard correction which is used in the $V, H \rightarrow b\bar{b}$ channel and a similar one to this which was proposed by Vadim Kostyukhin from Bonn. Other proposals for corrections will also be discussed but could not be tested.

5.11.1 Comparison to the Standard $V, H \rightarrow b\bar{b}$ Correction

The standard jet correction which is used for the $V, H \rightarrow b\bar{b}$ channel accounts for two effects: 1. if the b -quark decays semileptonically into a muon a part of the momentum is carried away by the muon, 2. a general shift for the jet momentum vector which depends on the transverse momentum of the jet. In a first step the energy loss of the muon (if present) in the calorimeter is subtracted from the jet momentum vector. The momentum vector of the muon is added to the jet momentum vector afterwards. The muon has to be tight, has $p_T > 4$ GeV and the distance to the jet axis has to be smaller than $\Delta R = 0.4$. For the second step a correction factor is applied to the jet vector which depends on the transverse momentum of the jet. The correction factor C is calculated via $C = p_T^{\text{truth}}/p_T^{\text{reco}}$. For my implementation of this

correction I obtained the correction factors from a histogram which is provided by the $V, H \rightarrow b\bar{b}$ working group [38]. The histogram can be found in the appendix A.2, fig. A.2. This standard correction is named " $\mu + p_T^{reco}$ " correction in the literature [38]. A comparison of the regression and the $\mu + p_T^{reco}$ correction applied to a $W^\pm H \rightarrow l^\pm \nu b\bar{b}$ Monte Carlo sample for the 2011 LHC working period with a nominal Higgs mass of $m_H = 125$ GeV is shown in fig. 5.10. The $m_{b\bar{b}}$ resolution after the standard correction derived from the Bukin fit is:

$$\text{after standard correction: } \frac{\sigma_p}{x_p} = \frac{(15.72 \pm 0.04) \text{ GeV}}{(115.39 \pm 0.01) \text{ GeV}} \hat{=} (13.62 \pm 0.04)\%$$

This is an improvement of 13.5% in resolution with respect to the resolution with the uncorrected jets. Nevertheless it is still significantly worse than the regression approach. Although the $\mu + p_T^{reco}$ correction corrects for the energy loss coming from the escaping muon which is also a very important input to the regression as it will be shown below (sec. 5.12), a single scale factor C cannot cover the various effects which lead to the unprecise measurement of the jet energy. It is noticeable that the standard correction does not change the peak position and just the peak width. The reason is that this correction only uses the last stage of the truth information, the truth jets, to evaluate the correction factor C . On this stage there was already some deterioration of the jet energy due to final state radiation.

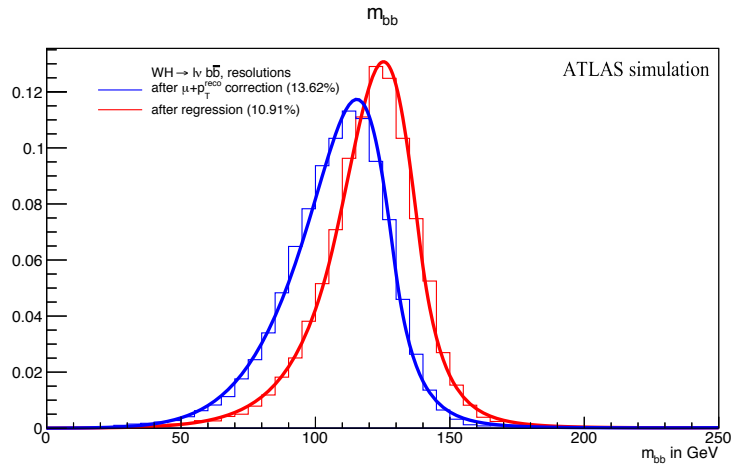


Figure 5.10: Shown is the $m_{b\bar{b}}$ distribution for 2011 $W^\pm H \rightarrow l^\pm \nu b\bar{b}$, $m_H = 125$ GeV Monte Carlo sample for the $\mu + p_T^{reco}$ corrected jets (blue) and the regression corrected jets (red) including the fit of a Bukin function to the histograms. The resolutions are 10.91% after the regression and 13.62% after the $\mu + p_T^{reco}$ correction.

Remark: During writing my thesis many new methods were considered for the jet energy correction for $V, H \rightarrow b\bar{b}$. Up to this time " $\mu + p_T^{reco}$ " was the standard correction. New developments tend to replace the p_T^{reco} correction by the GSC method (see section 5.11.3) or apply it additionally. Also other methods are under consideration.

5.11.2 Comparison to an Alteration of the Standard Correction

This correction was proposed as an alternative to the standard correction [39]. It accounts for the same two effects: semileptonic decays and a mismodeling in dependence of the jet p_T . The differences are that it also considers electrons in the jet and the p_T dependent correction uses a correction function instead of a single factor. All corrections are derived from the truth information on parton level. This "alternative

$\mu + p_T^{reco}$ correction is applied in three steps. The first step is equal to the standard correction: the energy loss of the muon in the calorimeter is subtracted from the jet momentum vector and then the muon momentum vector is added to the jet momentum vector. The second step applies the correction function [39]:

$$\mathbf{p}_{corr}^{jet} = \left(0.98 + 0.6e^{-0.031p_T[\text{GeV}]}\right) \mathbf{p}^{jet} \quad (5.4)$$

In the last step the jet momentum vector will be multiplied by a global scale factor $C = p_T^{truth}/p_T^{reco}$ of 1.025 if there is an electron in the jet and 1.038 if there is a muon in the jet. The electron has to be at least loose and has to have a transverse momentum of $p_T > 2.4 \text{ GeV}$. For the muon the same requirements as for the standard correction are valid [39].

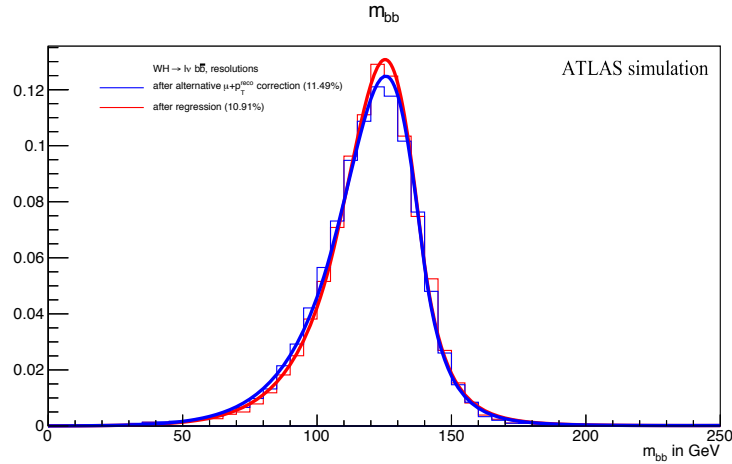


Figure 5.11: Shown is the $m_{b\bar{b}}$ distribution for 2011 $W^\pm H \rightarrow l^\pm \nu b\bar{b}$, $m_H = 125 \text{ GeV}$ Monte Carlo sample for the jets with the alternative $\mu + p_T^{reco}$ correction (blue) and the regression corrected jets (red) including the fit of a Bukin function to the histograms. The resolutions are 10.91% after the regression and 11.49% after the alternative $\mu + p_T^{reco}$ correction.

The effects on the $m_{b\bar{b}}$ resolution for $W^\pm H \rightarrow l^\pm \nu b\bar{b}$ sample for the 2011 period of data taking with $m_H = 125 \text{ GeV}$ is shown in fig. 5.11 in comparison to the distribution for the invariant mass of the two regression corrected jets. The figure shows that there is nearly no difference between those two corrections. The resolution derived from the Bukin fit is:

$$\text{after alternative } \mu + p_T^{reco} \text{ correction: } \frac{\sigma_p}{x_p} = \frac{(14.42 \pm 0.12) \text{ GeV}}{(125.52 \pm 0.20) \text{ GeV}} \hat{=} (11.49 \pm 0.10)\%$$

which is an improvement of 27% with respect to the distribution of the invariant mass of the uncorrected jets. Considering the resolutions derived from the fitted peak width and peak position this method is just 5% worse than the regression approach. This means that the correction function from the alternative $\mu + p_T^{reco}$ correction describes the effects, which lead to unprecise jet energy measurements, very well. Besides this version of the $\mu + p_T^{reco}$ correction is capable of shifting the peak position to the right position since it uses the partonic truth information to determine the correction function and scale factors. This is the advantage of this version compared to the standard one. Since the scale factors are determined for a slightly different event selection [39] as the one which was used here a small further improvement of the resolution could be possible with suitable scale factors. It may be concluded that the alternative standard correction as well as the regression are good approaches to identify sources for mismeasurements of the

jet energy and correct for them. Nevertheless the alternative standard correction seems to be more or less at its limit whereas the regression has still possibilities to be tuned (see also chapter 7).

5.11.3 Comparison to Other Methods

Other methods to correct the b -jet energy which are investigated for the $V, H \rightarrow b\bar{b}$ channel are the global sequential calibration (GSC) and a kinematic fit (KL). The GSC methods sequentially applies correction factors C . These correction factors are calculated from the average ratio between the reconstructed jet p_T^{reco} and the p_T^{truth} of the truth jet in dependence from different jet properties x :

$$C = \left(\left\langle \frac{p_T^{reco}}{p_T^{truth}}(x) \right\rangle \right)^{-1} \quad (5.5)$$

This is done for four different jet properties. The KL method uses a kinematic likelihood fit with twelve fit parameters which are properties of the leptons of the vector boson decay and the b -jets and three constraints on the parameters. These two methods are evaluated for the $ZH \rightarrow l^+l^-b\bar{b}$ channel for the 2012 LHC working period and applied to this channel in the combinations GSC+ μ + p_T^{reco} and GSC+ μ +KL for a nominal Higgs mass of $m_H = 125$ GeV. The $m_{b\bar{b}}$ resolution results are calculated for different combinations of cuts on the minimal transverse momentum of the jet with the lower p_T and regions for the transverse momentum of the Z boson. The resolutions vary from 11.0% and 8.81% for the GSC+ μ + p_T^{reco} correction and from 11% to 8.03% for the GSC+ μ +KL correction [40]. Although these corrections were tested on Monte Carlo samples for the 2012 period of data taking and with a slightly different event selection and a higher p_T cut on the b -jet with the higher energy there is some comparability with the regression for the $p_T^{2nd\,jet} > 20$ GeV and $p_T^Z < 90$ GeV region (see fig. 5.12). The regression also works with jets with a transverse momentum above 20 GeV and the lower p_T^Z region is the only region for the 2011 $ZH \rightarrow l^+l^-b\bar{b}$ sample for $m_H = 125$ GeV which has enough statistics to allow for a comparison. For the correction of the b -jets the regression model is used which was trained on the 2011 $W^\pm H \rightarrow l^\pm \nu b\bar{b}$ samples without the sample for $m_H = 125$ GeV.

In fig. 5.12 it is shown that all corrections reach a resolution of approximately 11%. Hence, as far as a comparison is possible, the regression, the GSC+ μ + p_T^{reco} correction and the GSC+ μ +KL correction perform quite similar. The plot from [40] in fig. 5.12a also shows the μ + p_T^{reco} correction with an additional calibration "EM+JES". EM just means that electromagnetic clusters were used to form the jets. This is also true for the jets which were used for the regression as well as for the implementation of the μ + p_T^{reco} correction in section 5.11.1. JES is an additional calibration factor which is derived from the relation between the calorimeter response and the true jet energy [41]. This calibration was not applied in section 5.11.1 since I also did not use them for any regression results. But it could be assumed that this single calibration factor could not fill the performance gap between the μ + p_T^{reco} correction and the regression. Overall these results imply that the regression also compares very well with newer approaches for improvement of the $m_{b\bar{b}}$ distribution.

5.12 Consistency Test of the Input Variables

The variables which are used for the training of the multivariate algorithm have a large impact on the performance. Therefore it is essential to test their influence on the training performance. This will be examined in this chapter. It is also important that the input variables are well described in the Monte Carlo samples. The agreement between data and Monte Carlo samples is shown in the appendix A.3

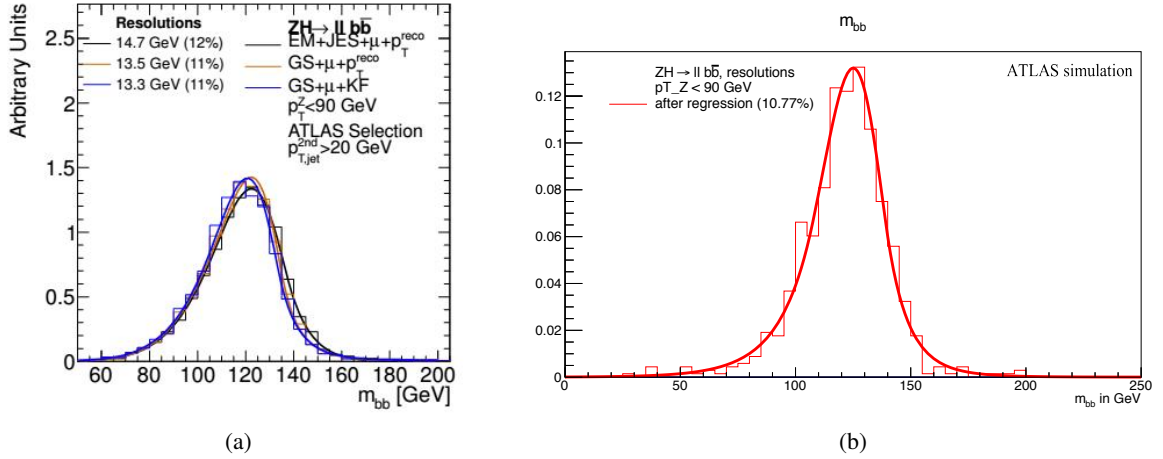


Figure 5.12: Shown is the $m_{b\bar{b}}$ distribution for (a) 2012 $ZH \rightarrow l^+ l^- b \bar{b}$ for $p_T^Z < 90$ GeV for the EM+JES+ $\mu + p_T^{reco}$ (black), GSC+ $\mu + p_T^{reco}$ (red) and GSC+ μ +KL correction (blue) [40] (b) 2011 $ZH \rightarrow l^+ l^- b \bar{b}$ for $p_T^Z < 90$ GeV for the regression correction. Bukin functions are fitted to all histograms. The resolutions are 12% after EM+JES+ $\mu + p_T^{reco}$ correction, 11% after the GSC+ $\mu + p_T^{reco}$ correction, 11% after the GSC+ μ +KL correction and 10.8% for the regression corrected jets.

which shows good agreement for most of the variables. Additionally some alternatives for the target and the input variables will be tested based on existing approaches from other groups and persons.

5.12.1 Importance of the Input Variables

This test was done to see the importance of the different input variables on the training performance. The *RMS* of the distribution of the difference between the transverse jet energy after the training and the true parton energy was used as a measure for the performance. A first hint on the influence of the input variables is given by the correlation ρ with the target variable. The correlation between two variables x and y with standard deviations σ_x and σ_y is given by:

$$\rho(x, y) = \frac{cov(x, y)}{\sigma_x \sigma_y} \quad \text{with the covariance: } cov(x, y) = \overline{xy} - \overline{x} \overline{y} \quad (5.6)$$

As a first approximation it can be assumed that a variable which is strongly correlated with the target has a large influence on the transverse energy of the jet and is therefore an important factor in the correction. The weakness of this measure is that it is a linear measure. Higher-order functional dependencies or non-functional relationships are not reflected very well by ρ . Contrary the BDTs recognize such relationships and may derive a good estimate for the transverse energy from those variables. One example is the sum of the transverse momenta of the muons in the jet which is a very important variable but its rank based on the correlation with the target is the fifteenth. Therefore I only used the ranking as a basic orientation for the order in which the input variables are tested. A table of the correlations is given in the appendix A.2, table A.3. For the test of the importance of the input variables I added the variables sequentially beginning with the one having the largest correlation to the target. Exceptionally I added the sum of the transverse momenta of the muons in the jet second since I noticed earlier that this variable has a large influence. After every training the $RMS(E_T^{\text{jet}} - E_T^{\text{true}})$ was evaluated on the test sample. Figures 5.13 and 5.13 show the performance of the training in dependence on the added input variables. Figure 5.13 shows the whole range for $RMS(E_T^{\text{jet}} - E_T^{\text{true}})$ and it shows that the *RMS* decreases by

BDT Performance

Figure 5.13: Shown is the performance of the regression measured by $RMS(E_T^{\text{jet}} - E_T^{\text{true}})$ after the training in dependence on the added input variables.

BDT Performance

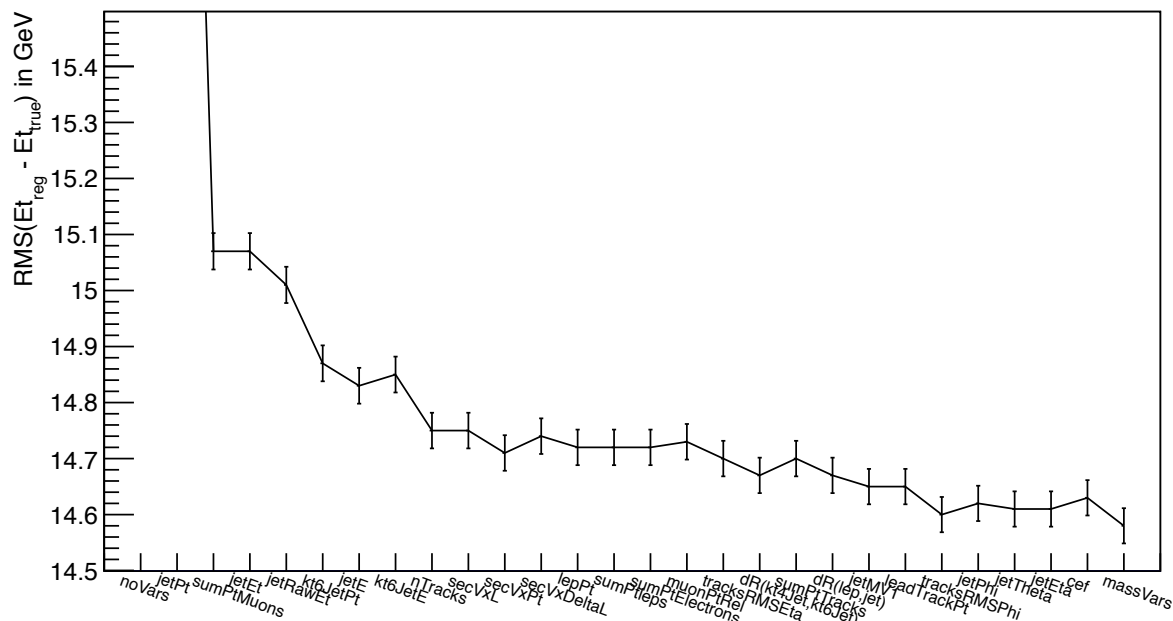


Figure 5.14: Shown is the performance of the regression measured by $RMS(E_T^{\text{jet}} - E_T^{\text{true}})$ after the training in dependence on the added input variables as a zoom into the region with smaller changes in RMS .

5.12.2 Test of an Alternative Target Variable

Some other regression approaches as well as many other jet corrections/calibrations use the information about the truth jets instead of the ones about the parton after the Higgs decay. The advantage of the truth jets is that they are less dependent on the Monte Carlo generator and the model it uses. The disadvantage is that the energy of them was already deteriorated due to the loss of final state radiation. This loss shows in the invariant mass spectrum of the two truth jets which is shown as an example for a $W^\pm H \rightarrow l^\pm \bar{\nu} b \bar{b}$ sample with a nominal Higgs mass of $m_H = 110 \text{ GeV}$ in fig. 5.15. This distribution's peak position is at the nominal Higgs mass. Nevertheless it has a long tail to smaller masses which originate from the loss of FSR. This will also manifest in the regression if the transverse momentum of the truth jets is chosen as the target. Most probable the right peak position of the $m_{b\bar{b}}$ will not be reconstructed with this target.

To verify this I trained BDTs with the same configuration as it was used before. Also the variables and the samples which were used remain the same. But for the target the transverse energy of the truth parton is replaced by the transverse energy of the truth jet. After the training the trained model is applied to the $W^\pm H \rightarrow t^\pm \bar{\nu} b \bar{b}$ samples for the 2011 period of data taking with $m_H = 125$ GeV like it was done for the old target variable. The result can be seen in fig. 5.16. The resolution of the $m_{b\bar{b}}$ distribution after the application of this new trained model is calculated from a Bukin fit to:

after regression with alternative target: $\frac{\sigma_p}{x_n} = \frac{(14.52 \pm 0.20) \text{ GeV}}{(110.28 \pm 0.14) \text{ GeV}} \hat{=} (13.17 \pm 0.18)\%$

This is significantly worse than the result with the old target. Remarkable is that this result just originates

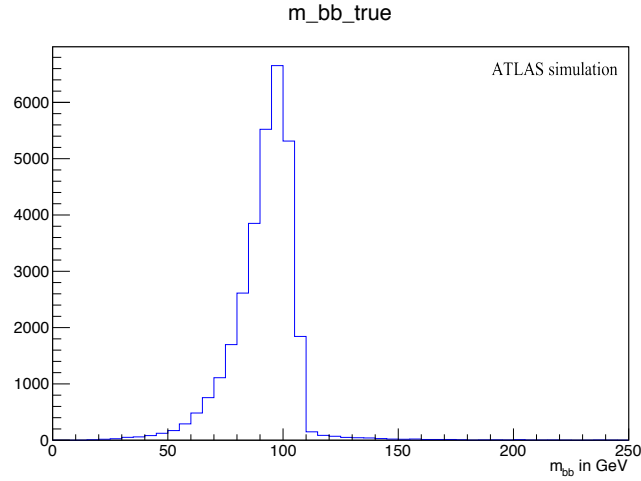


Figure 5.15: Shown is the invariant mass distribution of the truth jets exemplary for a 2011 $W^\pm H \rightarrow l^\pm \nu b \bar{b}$ samples with $m_H = 110$ GeV

from the worse peak position which is with (110.28 ± 0.14) GeV significantly underestimated. This behavior was expected. But the width of the peak $\sigma_p = (14.52 \pm 0.20)$ GeV compares relatively well with the width of (13.68 ± 0.11) GeV which is achieved with the parton E_T target.

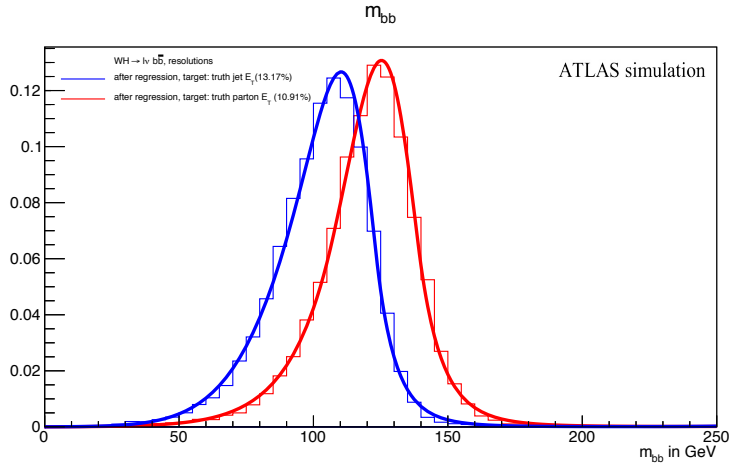


Figure 5.16: Shown is the $m_{b\bar{b}}$ distribution for 2011 $W^\pm H \rightarrow l^\pm \nu b \bar{b}$, $m_H = 125$ GeV Monte Carlo sample for the regression corrected jets with two different targets for the training: the E_T of the truth jets (blue) and the E_T of the true parton (red). A Bukin function was fitted to both histograms. The resolutions are 10.91% for the parton E_T target and 13.17% for the truth jet E_T target.

5.12.3 Test of Alternative Sets of Input Variables

The CDF and CMS experiment used multivariate regression to correct the b -jet energy and also performed several tests to find an optimal set of variables (listed in tab. 5.2). In this section the BDTs will be trained with the input variable sets of CDF and CMS. The target will be again the transverse energy of the truth parton and the configuration of the BDTs remains the same.

The CDF variables set contains 9 variables (see tab. 5.2). I used all of them with a small change: I used the decay length and its error calculated in the xyz -plane instead of the xy -plane since only the former one was available in the Bonn D3PDs. After the training with this input variables the trained model was applied to the $W^\pm H \rightarrow l^\pm \nu^{(\pm)} b \bar{b}$ sample with $m_H = 125$ GeV. The resulting $m_{b\bar{b}}$ distribution is shown in fig. 5.17 in comparison to the training with the full set of 30 variables. The resolution after the application of the model trained with the CDF input variable set is:

$$\text{after regression with CDF input variables: } \frac{\sigma_p}{x_p} = \frac{(16.04 \pm 0.29) \text{ GeV}}{(125.75 \pm 0.43) \text{ GeV}} \hat{=} (12.76 \pm 0.23)\%$$

This is worse than the result which is achieved with the full variable set. Nevertheless the peak position is approximated rather well but the width is more than 2 GeV bigger than the width after applying the model with the full variable set. One main reason could be that the CDF variable set does not cover semileptonic decays and therefore does not correct for the energy carried away by the muon. It was shown in the previous section that the transverse momentum of the muon is a very important input variable.

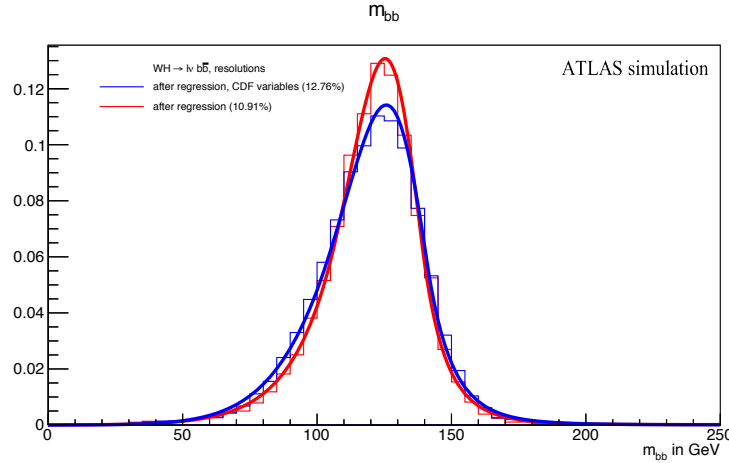


Figure 5.17: Shown is the $m_{b\bar{b}}$ distribution for 2011 $W^\pm H \rightarrow l^\pm \nu^{(\pm)} b \bar{b}$, $m_H = 125$ GeV Monte Carlo sample for the regression corrected jets with two different sets of input variables for the training: the 9 CDF variables (blue) and the 30 input variables of my approach (red). A Bukin function was fitted to both histograms. The resolutions are 10.91% for the full variable set and 12.76% for the CDF variable set.

The CMS variable set should contain 16 input parameters (see tab. 5.2). I could just use 14 since the jet energy uncertainty ("JEC") and the event density (" ρ_{25} ") were not available in the Bonn D3PDs. I also did a small alteration: instead of the number of jet constituents which cannot be measured in the ATLAS detector I used the number of tracks. Since CMS uses the raw p_T instead of the raw E_T I implemented this variable just for this test. For the standard training just raw E_T is used. Figure 5.18 shows the $m_{b\bar{b}}$ distribution after the model trained on those 14 CMS based variables is applied. The resolution derived from the Bukin fit is:

$$\text{after regression with CMS based input variables: } \frac{\sigma_p}{x_p} = \frac{(14.08 \pm 0.11) \text{ GeV}}{(125.50 \pm 0.03) \text{ GeV}} \hat{=} (11.22 \pm 0.09)\%$$

This resolution is within its errors nearly equivalent to the resolution of $(10.91 \pm 0.09)\%$ which is reached with a training with the full variable set. Possible explanations for the slightly worse resolution with

the CMS based variable set may be the lack of the two additional CMS variables or the availability of the out of cone leakage variables in the my set of 30 variables. Nevertheless this result emphasizes that some of the 30 variables may be dropped without any significant performance losses.

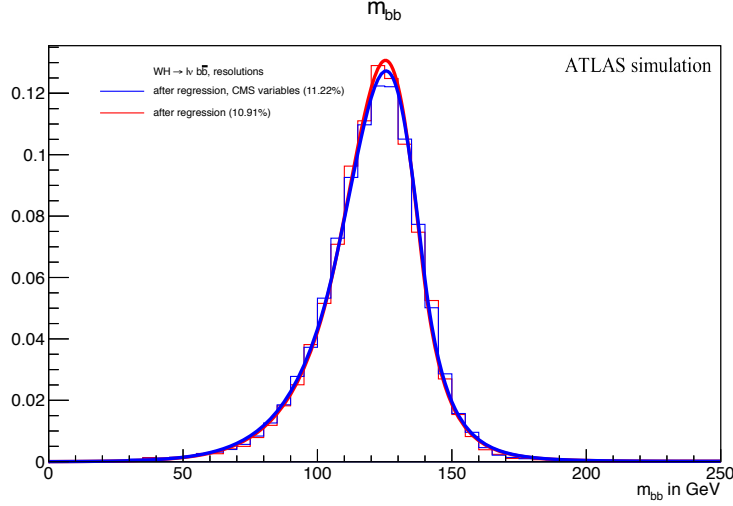


Figure 5.18: Shown is the $m_{b\bar{b}}$ distribution for 2011 $W^\pm H \rightarrow l^\pm \nu^{(-)} b\bar{b}$, $m_H = 125$ GeV Monte Carlo sample for the regression corrected jets with two different sets of input variables for the training: 14 variables based on the CMS variables (blue) and the 30 input variables of my approach (red). A Bukin function was fitted to both histograms. The resolutions are 10.91% for the full variable set and 12.22% for the CMS based variable set.

Overall this study with alternative variables in combination with the study with a different target variable showed the importance of the target and the input variables. The conclusion is that the reconstruction of the right peak position is mainly dependent on the target variable whereas the reduction of the peak width is mainly dependent on the choice of input variables.

Summary

In this thesis I introduced and tested successfully a new method for the improvement of the Higgs boson mass resolution in the channel $W^\pm H \rightarrow l^\pm \nu^{(\pm)} b \bar{b}$ by correcting the energy of b -jets. This method uses a machine learning algorithm (BDTs) to derive a prediction model for the b -jet energy based on 30 input variables which are related to the b -jet energy and account for the most important effects which lead to inaccuracies in the measurement of the b -jet energy. The prediction model is built with the information from Monte Carlo samples which contain simulated $W^\pm H \rightarrow l^\pm \nu^{(\pm)} b \bar{b}$ events. The target is the transverse energy of the b -quark directly after the Higgs boson decay and before any deterioration due to final state radiation. The energy was chosen as a target because it is the component of the jet four momentum vector which is measured with the largest inaccuracies and therefore has to be corrected. To avoid any dependence of the correction on a certain Higgs mass, the correction is jet-based and Monte Carlo samples with all available Higgs boson mass values are used to build the prediction model.

The regression achieves an relative improvement of approximately 16% for the transverse energy of the b -jets. Applying the regression to a $W^\pm H \rightarrow l^\pm \nu^{(\pm)} b \bar{b}$ Monte Carlo sample with a simulated Higgs boson mass of $m_H = 125$ GeV results in an improvement of approximately 31% for the relative di-jet mass resolution. A similar improvement is achieved for the $ZH \rightarrow l^+ l^- b \bar{b}$ channel. The regression narrows the peak width and also reconstructs the correct peak position which was underestimated before. Other methods for the correction of the b -jet four momentum vector, as proposed by the ATLAS group for the $V, H \rightarrow b \bar{b}$ analysis (HSG5) and from V. Kostyukhin, were also tested. Both methods add the momentum of an escaping muon from a semileptonic b -decay to the jet momentum and apply additional scale factors for the jet momentum. The results for all methods tested on a $W^\pm H \rightarrow l^\pm \nu^{(\pm)} b \bar{b}$ Monte Carlo sample with a simulated Higgs boson mass of $m_H = 125$ GeV are shown in table 6.1. The regression performs significantly better than the HSG5 recommendation and slightly better than the method from V. Kostyukhin.

Crosschecks which support the results of the regression correction and show its consistency were also presented in this thesis. The general applicability of the regression method to b -jets from any process was demonstrated for top-antitop-quark production where the regression was used to correct the b -jet which lead to a slight improvement of the top mass measurement. The application of the regression to the b -jets from a W boson with associated b -jets Monte Carlo sample proved that the regression was not tuned to a certain value for $m_{b\bar{b}}$ since no artificial peaks were induced in the $m_{b\bar{b}}$ distribution of this sample.

In additional tests I showed that the set of 30 input variables yields a very good performance of the

correction method	x_p in GeV	σ_p in GeV	resolution σ_p/x_p
reconstructed	115.40 ± 0.26	18.18 ± 0.16	$(15.76 \pm 0.14)\%$
HSG5 recommendation	115.39 ± 0.01	15.72 ± 0.04	$(13.62 \pm 0.04)\%$
V. Kostyukhin recommendation	125.52 ± 0.20	14.42 ± 0.12	$(11.49 \pm 0.10)\%$
regression	125.32 ± 0.20	13.68 ± 0.11	$(10.91 \pm 0.09)\%$

Table 6.1: Results for the resolution $m_{b\bar{b}}$ for different correction methods for the b -jets from $W^\pm H \rightarrow l^\pm \nu^{(-)} b\bar{b}$ decays with $m_H = 125$ GeV. The regression gives the best result for the Higgs mass resolution.

regression correction. Several input variables have a significant influence on the performance of the regression. These variables are: the sum of the transverse momenta of the muons in the jet, variables for jets with a cone radius of $\Delta R = 0.6$, the number of tracks, secondary vertex variables as well as variables sensitive to the jet shape. The other input variables also have an influence on the performance of the BDTs. None of the 30 input variables decreases the performance of the BDTs significantly.

Furthermore I showed that it is important to use the initial b -quark energy as a target instead of the energy of the jets on truth level. The former target is able to correct the underestimation of the b -jet energy and reconstructs the correct peak position in the $m_{b\bar{b}}$ distribution for the $W^\pm H \rightarrow l^\pm \nu^{(-)} b\bar{b}$ channel.

It can be concluded that the regression as described and tested in this thesis is a suitable method to improve the Higgs boson mass resolution in the $W^\pm H \rightarrow l^\pm \nu^{(-)} b\bar{b}$ decay channel. I also showed that the regression compares very well with other jet correction methods. Therefore it has the potential to become a standard correction for b -jets especially in the context of $V, H \rightarrow b\bar{b}$ analyses.

Outlook

This thesis has shown that the training based on 30 input variables for the regression leads to a large improvement in the Higgs mass resolution. Also the individual importance of the input variables for the results was shown. Still there might be some options for further studies. This chapter will give an overview of these options whereas the two main parts are the determination of a final set of input variables and the extension to later LHC runs and ATLAS analyses which both includes possible additional studies. These studies are partially of minor importance and where not done due to the time constraints of this thesis.

7.1 Determination of a Final Set of Input Variables

A possible task for the future is to revise the set of input variables. This would be important if the regression approach should become a part of standard ATLAS analyses. It was already shown that the set of 30 input variables covers the most important effects on the b -jet energy measurement. It was also shown that some of those variables have a high importance. These variables are: the sum of the transverse momenta of the muons in the jet, variables for jets with a cone radius of $\Delta R = 0.6$, the number of tracks, secondary vertex variables as well as variables sensitive to the jet shape. It could be considered to drop some of the other variables for practicability reasons since every variable comes with its own systematic uncertainty. These systematic uncertainties may be taken into account while determining a final set of input variables.

Besides the input variables also systematic uncertainties on the target variable may be studied. Since the invariant mass of the two b -quarks on the truth parton level is exactly the nominal Higgs mass this target has the big advantage that it is able to reconstruct the peak position of the $m_{b\bar{b}}$ distribution correctly. This leads to the conclusion that this target should be kept. A problem could arise because the parton information depends on the shower model of the used Monte Carlo generator. Therefore a test with different generators would be useful to study systematic effects.

It may be worth to review the modeling of the input variables in the training Monte Carlos for the full data set and with dedicated scale factors. If some variables show a large mismodeling it will be necessary to decide whether they should be kept. But the first comparison of data and Monte Carlos samples (see A.3) does imply that all variables are reasonably well modeled. The source for the absence of secondary vertex variables for 88% of the b -jets is not yet understood. To use the secondary variables does not harm the performance which was proven in 5.12 but more secondary vertex information are

desirable since the potential for further improvements by including those variables would be expected. The variables for the leptons in the jet are also only available for a subset of b -jets due to the branching ratio for semileptonic decays. Therefore a categorized training could be taken into consideration. This means that the training sample would be split into subsamples/categories via a categorization criterion like: the lepton variables are filled with values > 0 . In each of those categories an independent training is performed. This means that also the application has to be done category wise. This categorized training and application has the potential to improve the performance of the regression further. The drawback is that each of the subsamples have to contain enough events for a reasonable training. This is not the case for the available samples for the 2011 period of data taking. Another interesting additional test would be to apply the standard muon in jet correction which subtracts the muon energy loss in the calorimeter from the jet momentum and adds the muon momentum to the jet momentum afterwards before the regression. Since the escaping muon is an effect which affects the whole jet momentum vector the standard muon in jet correction might give a slightly better performance because it is a momentum vector based correction. In the regression just $\sum_{\text{muons}} p_T$ is used to derive an energy based correction.

Once the set of input variables is fixed the systematic uncertainties for the regression have to be studied and understood.

7.2 Extension of the Regression to the Full Data Set and Compatibility with ATLAS Analyses

The last step for the regression is the training for the 2012 period of data taking and the application of the regression to the full data set of 2011+2012 afterwards. With the new release of the Bonn D3PDs all information are available to train the regression for the 2012 data set. The regression may be used without big problem within the standard $V, H \rightarrow b\bar{b}$ since it relies on common object definitions and selections. To use the regression within the standard analysis the systematic uncertainty for the regression has to be known to testify the significance of the signal for the whole data set. An execution of the full analysis including an application of the regression to b -jets will show whether the regression improves the search for the Higgs boson in the $V, H \rightarrow b\bar{b}$ decay channel.

Appendix

A.1 Discussion of Different Measures for the $m_{b\bar{b}}$ Resolution

There is no common measure which defines the resolution of an invariant mass distribution. The important measures to make a statement about the invariant mass resolution is the width of the peak and the position of the peak. But also in the ATLAS/CMS literature many different measures for this are used. Therefore this chapter will introduce the most common ones and the disadvantages/advantages will be explained.

- **RMS/mean:**

A quick and simple way to get a value for the invariant mass resolution is the ratio between the *RMS* and the mean value of the distribution. The *RMS* is a measure for the spread of N measurements x_i . Note that the *RMS* is in this case equal to the standard deviation σ which is found in the literature. *RMS* is not the quadratic mean! This naming is due to naming conventions in ROOT which also established in high energy physics. The mean value $\bar{x} = \mu$ and *RMS* = σ are defined as [42]:

$$\mu = \frac{1}{N} \sum_{i=1} x_i \quad \sigma = \sqrt{\frac{1}{N} \sum_i (x_i - \mu)^2} \quad (\text{A.1})$$

and their errors are [42]:

$$\Delta\mu = \frac{\sigma}{\sqrt{N}} \quad \Delta\sigma = \frac{\sigma}{\sqrt{2N}} \quad (\text{A.2})$$

When building the fraction of *RMS*/ μ the error of this is according to Gaussian error propagation:

$$\Delta\left(\frac{RMS}{\mu}\right) = \sqrt{\left(\Delta\sigma \frac{1}{\mu}\right)^2 + \left(\Delta\mu \frac{\sigma}{\mu^2}\right)^2} \quad (\text{A.3})$$

The problem with this measures are that it loses its meaning for asymmetric distributions or distributions with long tails. The more asymmetric a distribution the more the mean differs from the peak position. The *RMS* is susceptible for long tails. It gets larger for long tails and loses its connection to the width of the actual peak. Nevertheless the *RMS*/mean is a good measure to get a quick estimate for the resolution. The other advantage is, that these value are not tunable, i.e. they cannot be calculate in a certain way to get the desired values. Therefore this measure was

used in this thesis especially in cases for whom the trends of the resolution were important.

- **RMS/median:**

This measure is quite similar to $RMS/mean$ but instead of the mean it uses the median of the distribution. The median is the value for which half of the data elements are below it and the other half above it. This has the advantage that the median should give a better estimate for the peak position. It is still susceptible to highly asymmetric distributions but not as sensitive as the mean. But there is a problem since all distributions which are used are binned. Therefore there are several data points with the same value. This makes the median dependent on the binning. Due to this and a not very large difference between median and mean for the considered distributions, the median will not be used in this thesis [42].

- **Core resolution of a Gaussian fit:**

Due to the statistical nature of particle decays overlaid with blurring of detector resolution the invariant mass distribution should have a Gaussian shape. Therefore the idea is to fit a Gaussian function to the data and read of the mean μ and the width σ from the fit results. For an ideal Gaussian shape the μ is the peak position and the σ is the standard deviation. The Gaussian function [42]:

$$f(x) = A_p e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (A.4)$$

has three parameters A , μ and σ which are determined in the fitting process such that the function fits the data best. Then the ratio σ/μ gives a measure for the resolution of the distribution. A_p is just the height of the peak and does not play a role for the resolution. The problem with this method is that the peak has no perfect Gaussian shape due to systematic effects. A Gauss fit often fails to describe the tails of the distribution and the asymmetry of the distribution. Since the core of the distribution, i.e. a narrow region around the peak, often has a quite good Gaussian shape many analyses limit the fit only to the core of the distribution. This gives reasonable results. Nevertheless the fit is highly dependent on the fit range meaning the definition of the core region which makes this method highly tunable. An independent measure for a good fit range does not exist. Normally the narrower the fit region the better the resolution. Due to this problem this method will neither be used in this thesis.

- **Resolution of a Bukin fit:**

This function was especially developed for high energy particle physics problems [36]. The Bukin function is based on a convolution of a Gaussian and an exponential function. It is dedicated to fit it to asymmetric functions with a peak by modeling the tails separately. I used the Bukin function as it is implemented in the RooFit package which is a fit interface for ROOT [43]:

$$f(x) = A_p \exp \left[\frac{\xi \sqrt{\xi^2 + 1} (x - x_1) \sqrt{2 \ln 2}}{\sigma_p (\sqrt{\xi^2 + 1} - \xi)^2 \ln (\sqrt{\xi^2 + 1} + \xi)} + \rho \left(\frac{x - x_i}{x_p - x_i} \right)^2 - \ln 2 \right] \quad (A.5)$$

$$x_{1,2} = x_p + \sigma_p \sqrt{2 \ln 2} \left(\frac{\xi}{\sqrt{\xi^2 + 1}} \mp 1 \right)$$

where $\rho = \rho_1$ and $x_i = x_1$ for $x < x_1$ and $\rho = \rho_2$ and $x_i = x_2$ for $x \geq x_2$. On the whole the Bukin

function depends on six free parameters A_p , x_p , σ_p , ρ_1 , ρ_2 and ξ which have to be fitted:

- A_p – peak height
- x_p – peak position
- σ_p – peak width: $\frac{FWHM}{2\sqrt{2\ln 2}} = \frac{FWHM}{2.35}$; $FWHM \hat{=}$ full width half maximum
- ρ_1 – parameter of the "left tail"
- ρ_2 – parameter of the "right tail"
- ξ – peak asymmetry parameter

The peak width σ_p is equivalent to the width σ of a Gaussian function. The resolution would then be given by the ratio of the peak width σ_p and the peak position x_p . The other parameters are not needed. But the asymmetry parameter may give a hint if also the asymmetry of the peak gets lessened due to the correction. Since the asymmetry is likely to change for the correction the Bukin fit seems to give a fairer comparison than $RMS/Mean$. Tests showed that the Bukin function is able to describe asymmetric shapes like the $m_{b\bar{b}}$ distribution very well if the start values for the free parameters are chosen carefully. The values for σ_p and x_p can just be tuned slightly by the starting conditions. The Bukin function is also used in the supporting note for $H \rightarrow b\bar{b}$ concerning the $m_{b\bar{b}}$ resolution [40]. Due to the above mentioned points this function was used as a standard measure in this thesis to compare the results of the regression.

A.2 Additional Figures and Tables for the Analysis

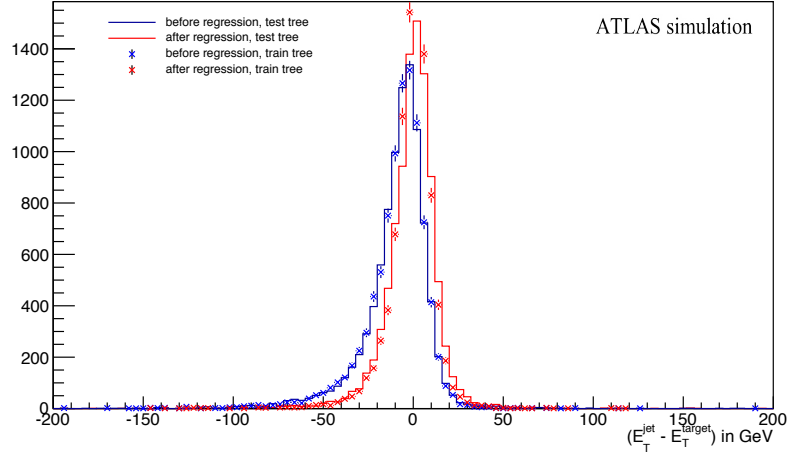


Figure A.1: Shown is the difference between the jet E_T and the truth parton E_T (target) for the reconstructed jets (blue) and the jets with the estimated E_T from the training (red) separated for the training (crosses) and the test (lines) sample. The training was done on $ZH \rightarrow l^+l^-b\bar{b}$ samples with all available Higgs masses for the 2011 run. The results for the training and the test tree differ a lot which is a sign for overtraining due to less events in the sample.

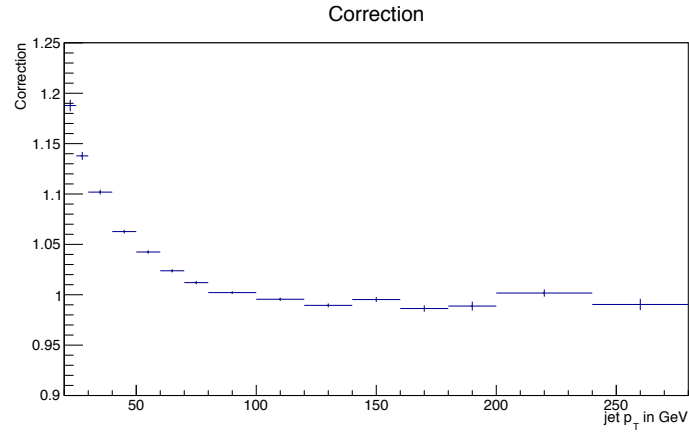


Figure A.2: Histogram to get the scale factors in dependence of the jet p_T for the standard correction " $\mu + p_T^{reco}$ " [38].

	$\mu(m_{b\bar{b}})$ in GeV	$\Delta\mu(m_{b\bar{b}})$ in GeV	$RMS(m_{b\bar{b}})$ in GeV	$\Delta RMS(m_{b\bar{b}})$ in GeV	N	$\frac{RMS(m_{b\bar{b}})}{\mu(m_{b\bar{b}})}$	$\Delta\left(\frac{RMS(m_{b\bar{b}})}{\mu(m_{b\bar{b}})}\right)$
reco η, ϕ , rest truth	125.52	0.06	9.24	0.04	24 436	0.0736	0.0003
reco m , rest truth	118.86	0.07	11.07	0.05	24 436	0.0931	0.0004
reco E_T , rest truth	113.23	0.11	16.93	0.08	24 436	0.1496	0.0007

Table A.1: Detailed evaluation of the $m_{b\bar{b}}$ resolution for truth b -quark vectors where truth η and truth ϕ or truth m or truth E_T were replaced by the corresponding reconstructed quantities.

	$\mu(\Delta E_T)$ in GeV	$\Delta\mu(\Delta E_T)$ in GeV	$RMS(\Delta E_T)$ in GeV	$\Delta RMS(\Delta E_T)$ in GeV	N
before regression, train tree	-8.397	0.052	17.21	0.04	107 527
before regression, test tree	-8.357	0.053	17.36	0.04	107 527
after regression, train tree	-1.765	0.044	14.42	0.03	107 527
after regression, test tree	-1.670	0.044	14.58	0.03	107 527

Table A.2: RMS and mean values for $\Delta E_T = E_T^{\text{jet}} - E_T^{\text{true}}$ before and after the regression separate for the training and the test tree.

rank	variable	correlation with target
1	jetPt	9.069×10^{-1}
2	jetEt	9.068×10^{-1}
3	jetRawEt	9.051×10^{-1}
4	kt6JetPt	8.570×10^{-1}
5	jetMass	7.230×10^{-1}
6	kt6JetMass	6.398×10^{-1}
7	jetMt	6.152×10^{-1}
8	jetEnergy	5.904×10^{-1}
9	kt6JetEnergy	5.536×10^{-1}
10	nTracks	4.901×10^{-1}
11	secVx_m	3.169×10^{-1}
12	secVx_decayLength	2.926×10^{-1}
13	secVx_pt	2.723×10^{-1}
14	secVx_decayLengthErr	2.216×10^{-1}
15	sumPt_muonsInJet	1.871×10^{-1}
16	pt_lepInJet	1.304×10^{-1}
17	sumPt_lepsInJet	1.238×10^{-1}
18	sumPt_electronsInJet	8.390×10^{-2}
19	ptRel_muonInJet	8.183×10^{-2}
20	tracksRMSEta	7.127×10^{-2}
21	dRkt4Jetkt6Jet	6.693×10^{-2}
22	dR_jetLepInJet	3.550×10^{-2}
23	sumPt_tracks	3.373×10^{-2}
24	jetMV1	2.723×10^{-2}
25	pt_leadTrack	1.515×10^{-2}
26	tracksRMSPhi	1.073×10^{-2}
27	jetPhi	8.114×10^{-3}
28	jetTheta	2.462×10^{-3}
29	jetEta	2.222×10^{-3}
30	cef	1.650×10^{-4}

Table A.3: Ranking of the input variables for the regression, based on the correlation ρ with the target variable the truth b -quark E_T .

added variable	$RMS(\Delta E_T)$ in GeV	$\Delta RMS(\Delta E_T)$ in GeV	N
	17.36	0.04	107 527
jet p_T	17.12	0.04	107 527
$\sum_{\text{muons}} p_T$	15.07	0.03	107 527
jet E_T	15.07	0.03	107 527
jet raw E_T	15.01	0.03	107 527
anti-kt6 jet p_T	14.87	0.03	107 527
jet E	14.83	0.03	107 527
anti-kt6 jet E	14.85	0.03	107 527
$N(\text{tracks})$	14.75	0.03	107 527
sec vertex L_{3D}	14.75	0.03	107 527
sec. vertex p_T	14.71	0.03	107 527
sec. vertex ΔL_{3D}	14.74	0.03	107 527
lepton in jet p_T	14.72	0.03	107 527
$\sum_{\text{leptons}} p_T$	14.72	0.03	107 527
$\sum_{\text{electrons}} p_T$	14.72	0.03	107 527
muon in jet p_T^{rel}	14.73	0.03	107 527
$RMS(\eta(\text{tracks}))$	14.70	0.03	107 527
$\Delta R(\text{anti-kt4 jet, anti-kt6 jet})$	14.67	0.03	107 527
$\Delta R(\text{jet, lepton in jet})$	14.70	0.03	107 527
$\sum_{\text{tracks}} p_T$	14.67	0.03	107 527
jet MV1	14.65	0.03	107 527
leading track p_T	14.65	0.03	107 527
$RMS(\phi(\text{tracks}))$	14.60	0.03	107 527
jet ϕ	14.62	0.03	107 527
jet θ	14.61	0.03	107 527
jet η	14.61	0.03	107 527
cef	14.63	0.03	107 527
jet m , anti-kt6 jet m , jet m_T , sec. vetex m	14.58	0.03	107 527

Table A.4: Detailed results for the test of the importance of the input variables on the regression result for a single b -jet. Shown are the results for $RMS(\Delta E_T) = RMS(E_T^{\text{jet}} - E_T^{\text{true}})$ in dependence on the used input variables which were added sequentially.

physics process	generator	scale factor
WH125	Pythia	1
ZH125	Pythia	1
Z+light	AlpgenJimmy	0.87
Z+hf	AlpgenJimmy	1.26
W+light	AlpgenJimmy	1.21
W+hf	Alpgen (for c) PowHeg (for b)	1.11
W+ γ	AlpgenJimmy	
Diboson	Herwig	1
single top	AcerMC	1
$t\bar{t}$	MC@NLO	1.15
multi-jet ("QCD")	estimated from data	1.17

Table A.5: List of the used Monte Carlo samples with their generators and scale factors and the scale factor for the multi-jet contribution which was calculated from a fit to data in the m_T^W distribution.

A.3 Modeling of the Input Variables for the Training

The input variables play a big role for the trainings performance. Since they are derived from Monte Carlo samples it is necessary to check if those variables are well described in the Monte Carlos. The way to do this is to compare the Monte Carlos with the actual data recorded in the detector. For this not only the WH Monte Carlo samples are needed but also all other Monte Carlos for physics processes which are backgrounds for this channel. The Monte Carlo generators which are used for the different processes are motivated by the standard analysis [34]. For some of the Monte Carlo samples additional scale factors are necessary since it was found that they do not model the desired process exactly. The scale factors I used are also motivated by the standard analysis. In the standard analysis the jets in the W/Z +jets samples were additionally flavor tagged based on the truth information in the Monte Carlo samples [34]. This was not possible for this thesis. Nevertheless the scale factors should be sufficient on average and sufficient to recognize extremely mismodeled input variables. The background from multi-jet production cannot be well modeled by Monte Carlo generators. Therefore it is estimated from data by performing the whole $V, H \rightarrow b\bar{b}$ analysis with an inverted cut on the lepton isolation. Since this is only an estimate the contribution from multi-jet processes has to be fitted to data afterwards. I fitted the multi-jet contribution in the transverse mass spectrum of the W boson (m_T^W) because the multi-jet events are accumulated in the region of small m_T^W and are the dominant contribution in this region. The scale factor which was found by the fit is 1.17. A table of the Monte Carlo generators and the scale factors I used for the different background contributions are given in A.5. Figures A.4, A.5, A.6, A.7, A.8 and A.9 show the comparison of data and Monte Carlo samples for all input variables which are used in the regression. In general there seems to be a small mismodeling in all kinematic variables since there are overshoots of data for small values and undershoots for larger values. Besides from this all variables are reasonably well modeled so that they are suitable input variables for the regression. A slightly bigger deviation is observable for variables connected to a mass. Nevertheless those variables do not have to be excluded from the training since the deviation between data and Monte Carlo samples is acceptable.

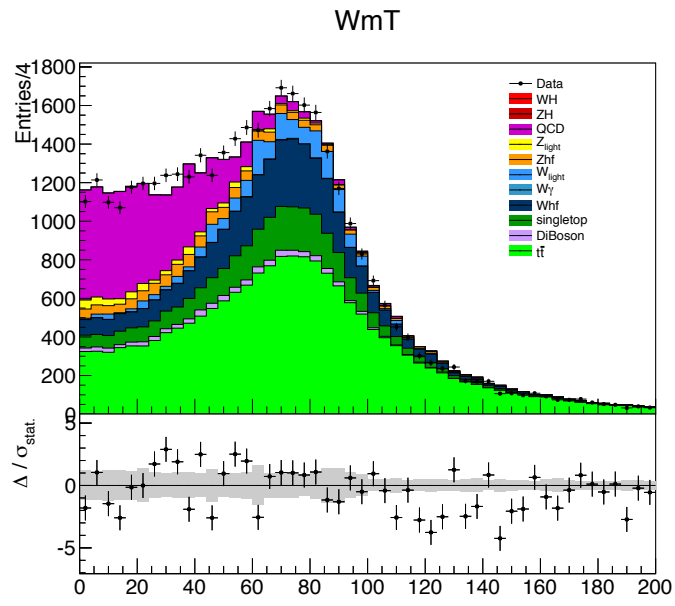


Figure A.3: Shown is the transverse mass distribution of the W boson to which the multi-jet contribution was fitted.

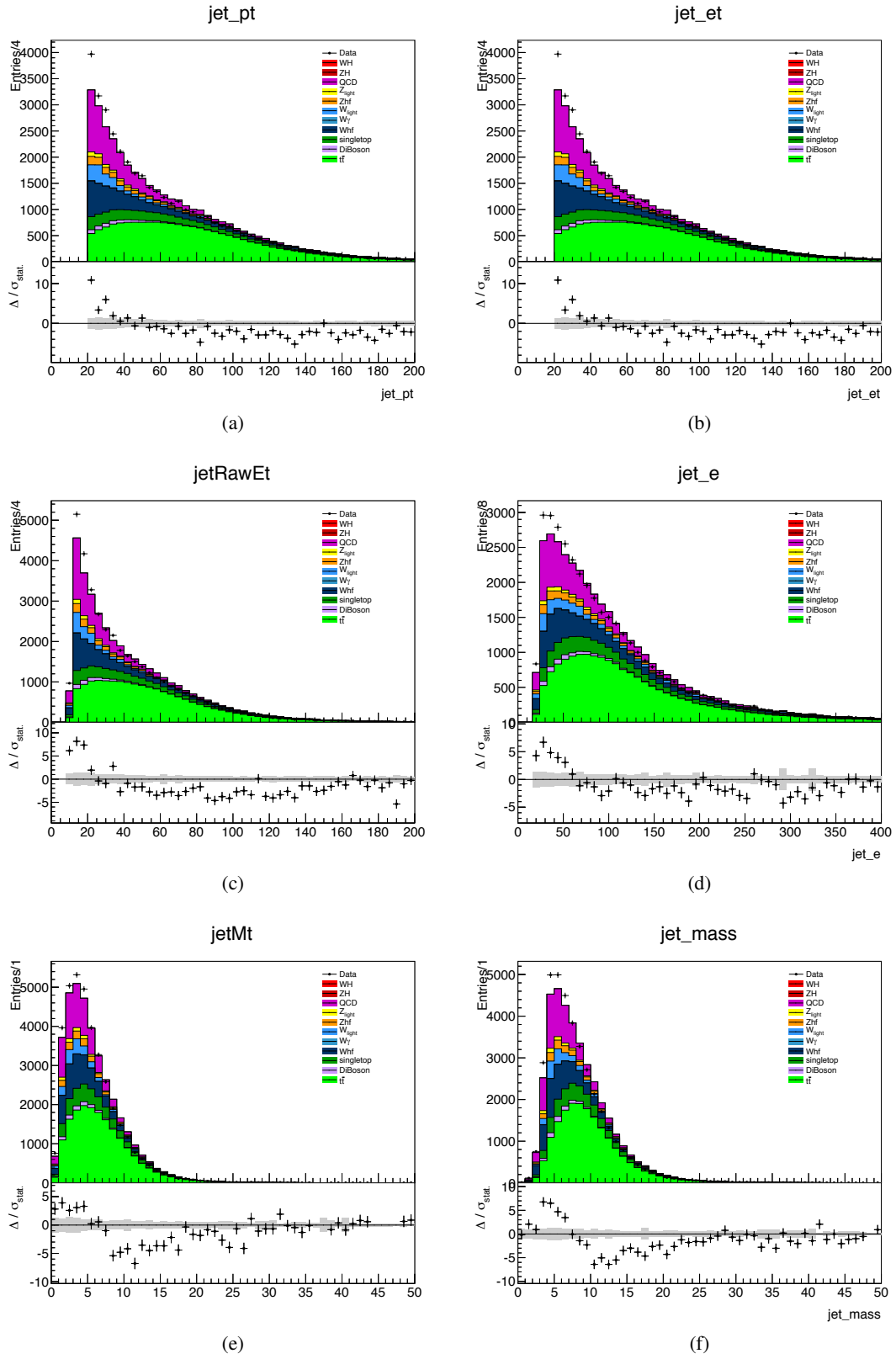


Figure A.4: (a) $\text{jet } p_T$, (b) $\text{jet } E_T$, (c) $\text{jet raw } E_T$, (d) $\text{jet } E$, (e) $\text{jet } m_T$, (f) $\text{jet } m$

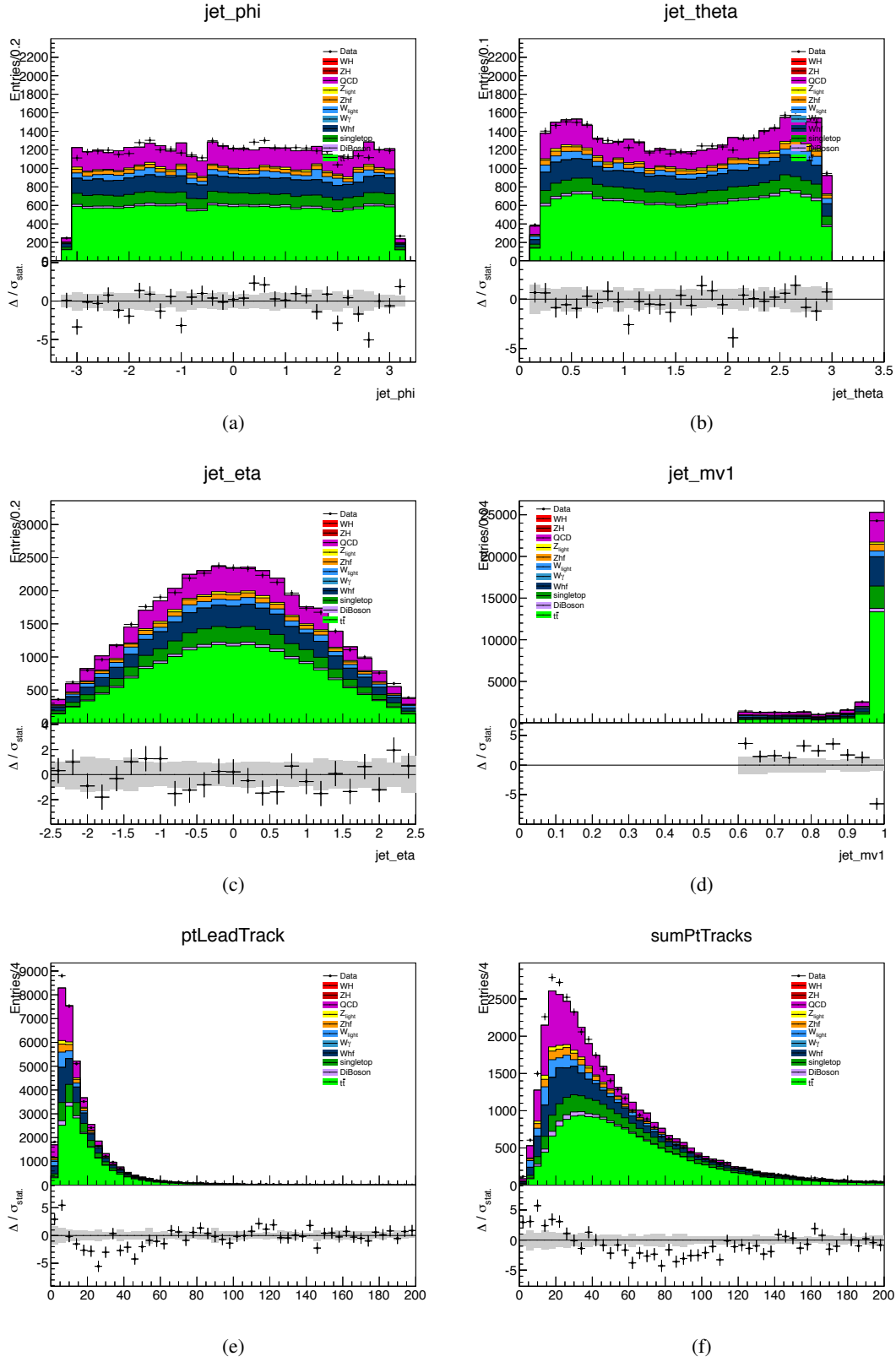


Figure A.5: (a) jet ϕ , (b) jet θ , (c) jet η , (d) jet MV1, (e) lead. track p_T , (f) $\sum_{\text{tracks}} p_T$

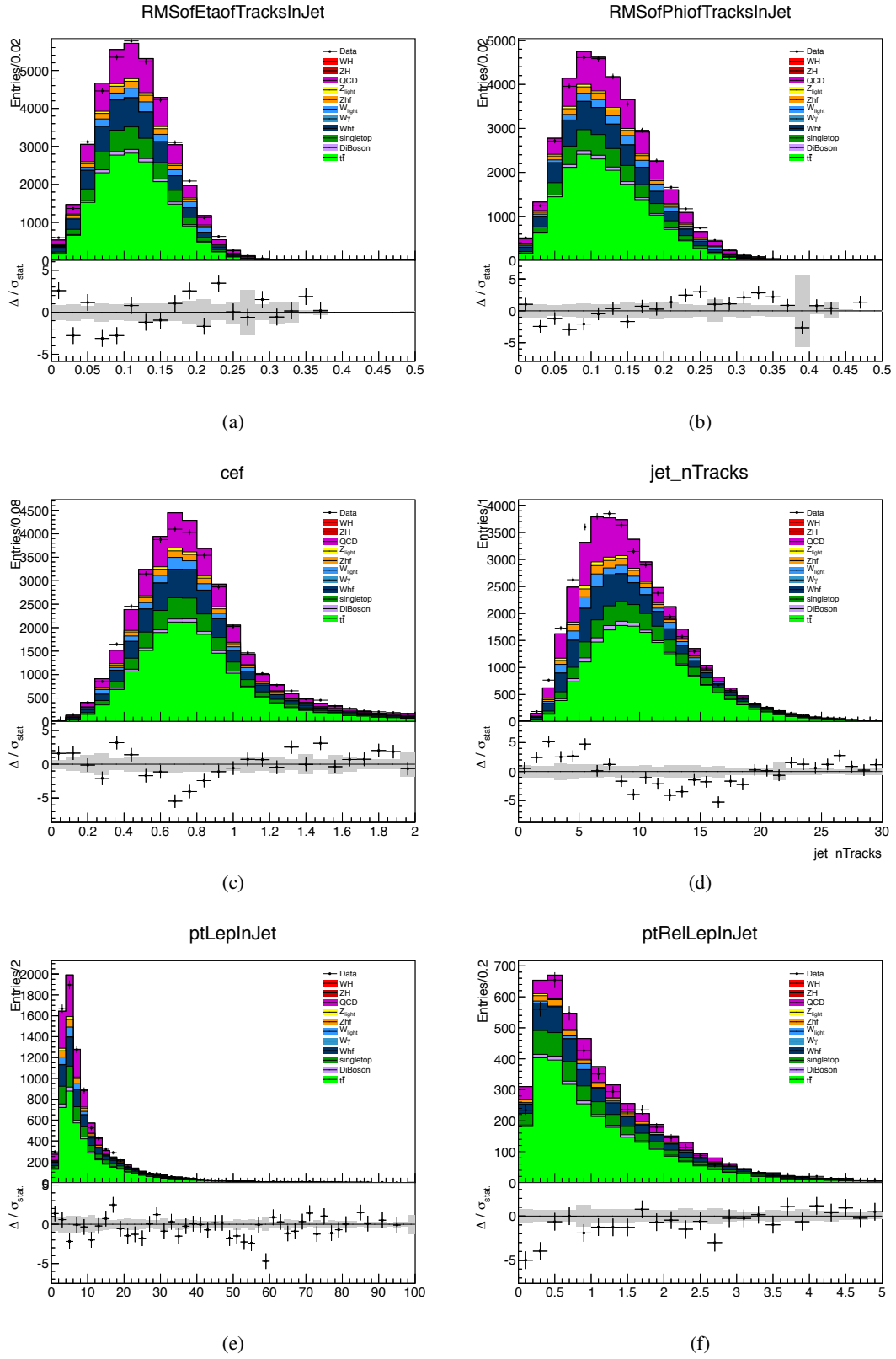


Figure A.6: (a) $RMS(\eta(\text{tracks}))$, (b) $RMS(\phi(\text{tracks}))$, (c) cef , (d) $N(\text{tracks})$, (e) lepton in jet p_T , (f) muon in jet p_T rel

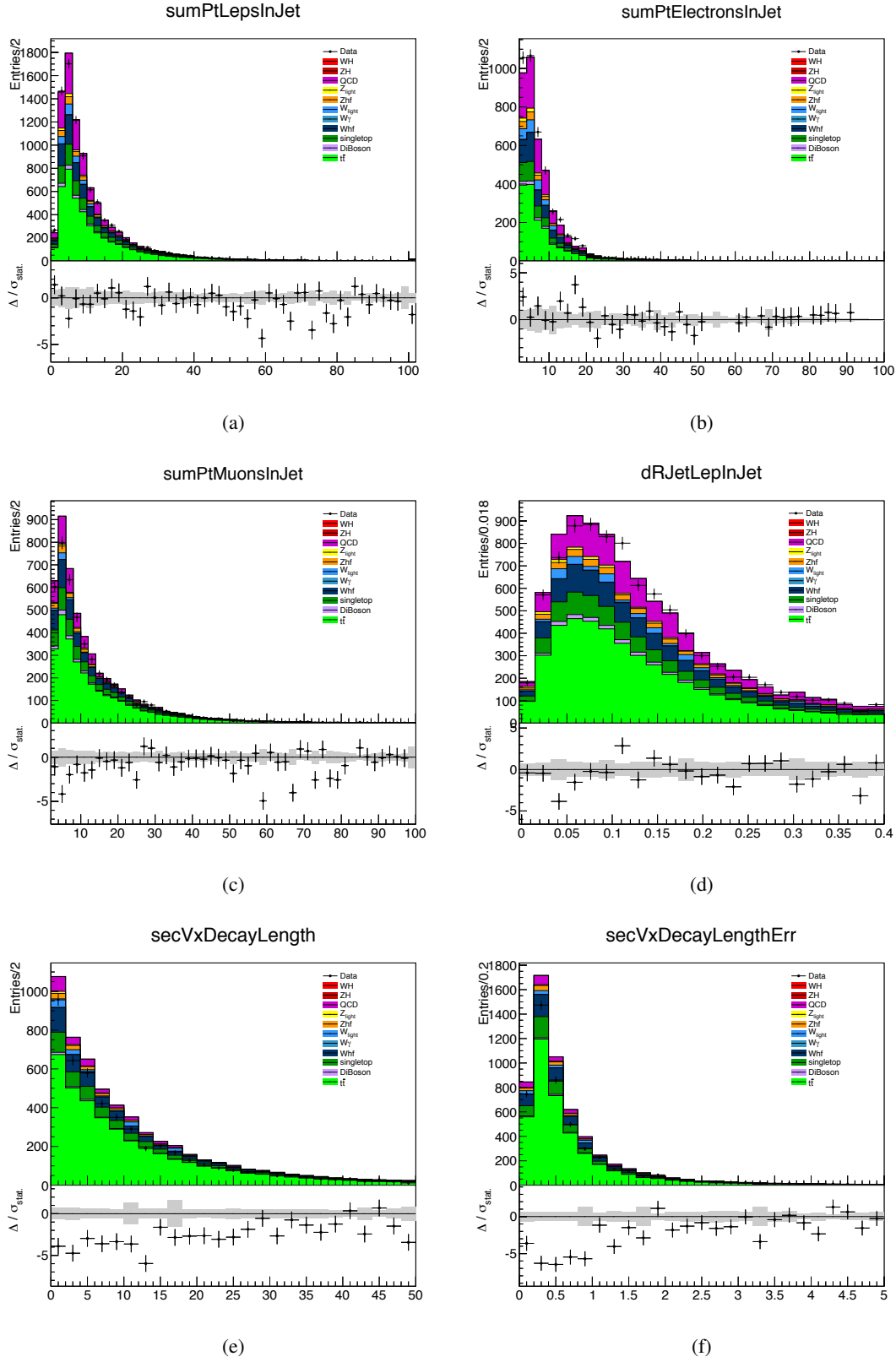


Figure A.7: (a) $\sum p_T$ (leptons), (b) $\sum p_T$ (electrons), (c) $\sum p_T$ (muons), (d) $\Delta R(\text{jet}, \text{lep})$, (e) secondary vertex L_{3D} , (f) secondary vertex ΔL_{3D}

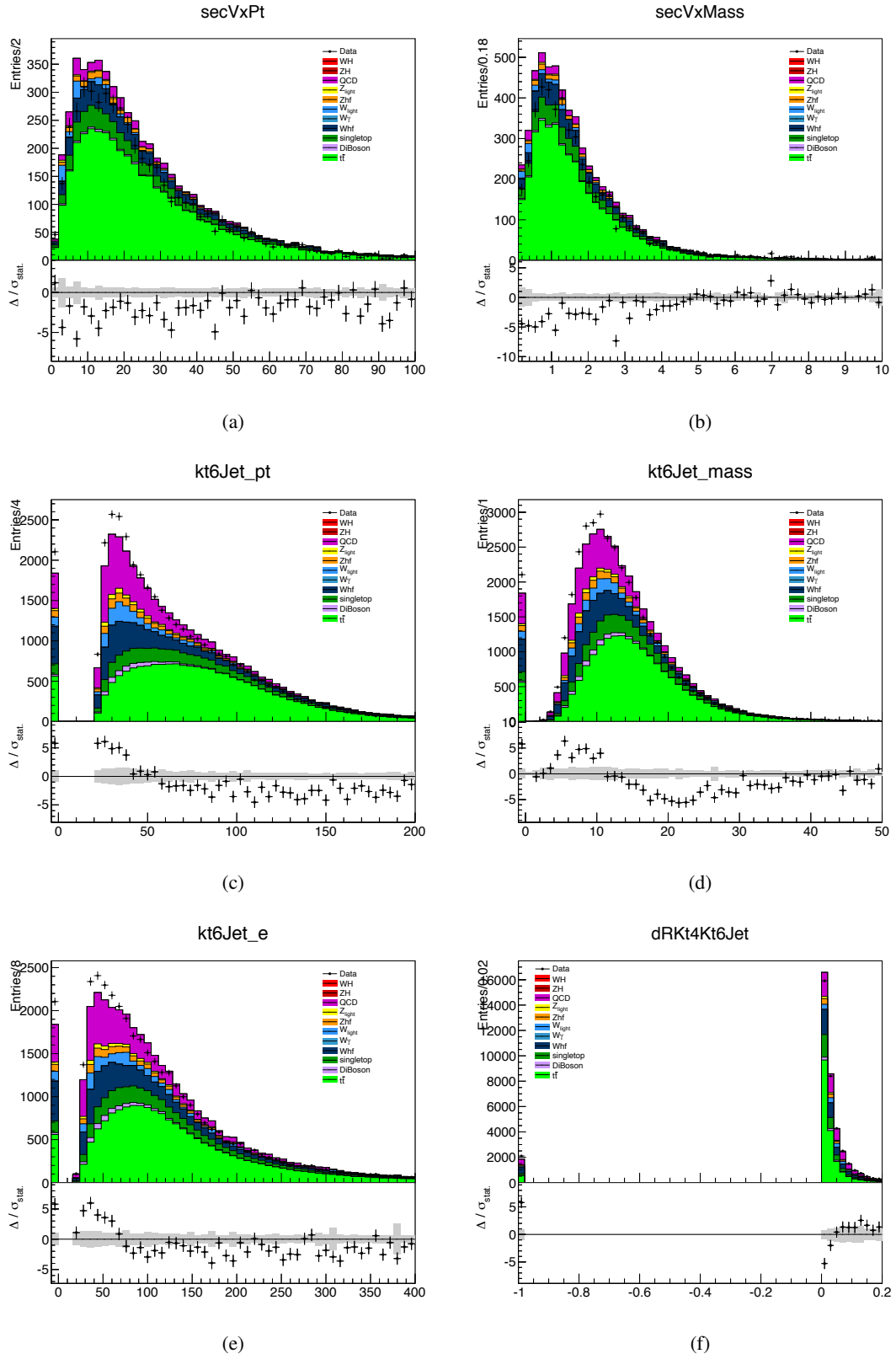


Figure A.8: (a) secondary vertex p_T , (b) secondary vertex m , (c) jet with $R = 0.6$ p_T , (d) jet with $R = 0.6$ m , (e) jet with $R = 0.6$ E , (f) $\Delta R(\text{kt4 jet}, \text{kt6 jet})$

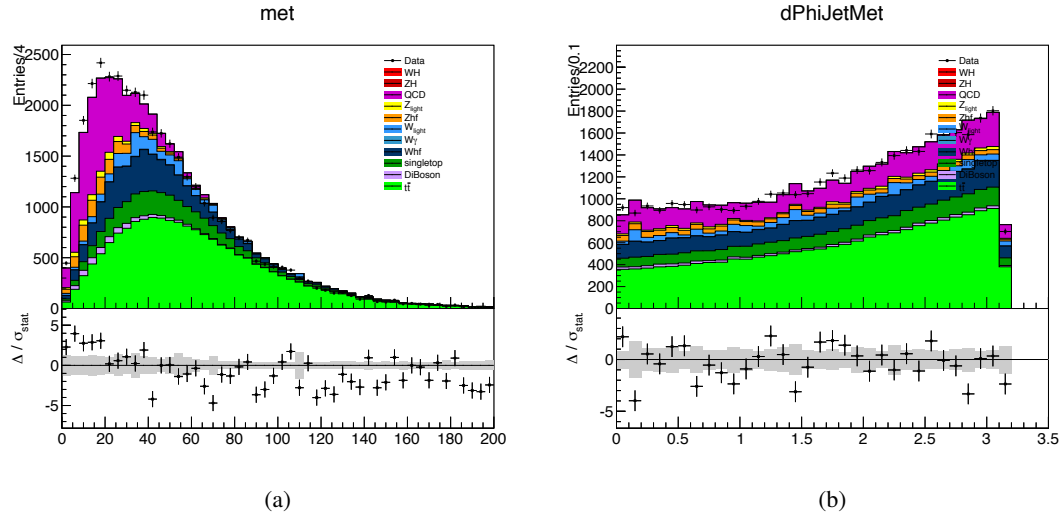


Figure A.9: (a) E_T^{mis} , (b) $\Delta\phi(\text{jet}, E_T^{mis})$

Bibliography

- [1] V. D. Barger and R. J. N. Phillips, *Collider Physics*, 1st ed., Westview Press, 1996.
- [2] I. Brock and T. Schörner-Sadenius, eds., *Physics at the Terascale*, 1st ed., Wiley-VCH, 2011.
- [3] L. Evans and P. Bryant, “The CERN Large Hadron Collider: Accelerator and Experiments”, *Journal of instrumentation* 3.S08001 (2008).
- [4] J. M. Campbell, J. W. Huston and W. J. Stirling, “Hard Interactions of Quarks and Gluons: a Primer for LHC Physics”, *Rept. Prog. Phys.* 70 (2007).
- [5] C. Lefevre, *LHC: the guide*, CERN-Brochure-2009-003-Eng, 2009, URL: <http://cds.cern.ch/record/1165534>.
- [6] ATLAS Collaboration, “The ATLAS Experiment at the CERN Large Hadron Collider”, *Journal of instrumentation* 3.S08003 (2008).
- [7] ATLAS Collaboration, *ATLAS Experiment - ATLAS Photos*, URL: <http://www.atlas.ch/photos/>.
- [8] ATLAS Collaboration, “Expected Performance of the ATLAS Experiment - Detector, Trigger and Physics”, 2009, arXiv: 0901.0512v4.
- [9] R. K. Ellis, W. J. Stirling and B. R. Webber, *QCD and Collider Physics*, 1st ed., Cambridge University Press, 1996.
- [10] ATLAS Collaboration, “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC”, *Phys. Lett. B* 716.1 (2012).
- [11] CMS Collaboration, “Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC”, *Phys. Lett. B* 716.1 (2012).
- [12] J. F. Gunion et al., *The Higgs Hunter’s Guide*, 1st ed., Perseus, 2000.
- [13] Novaes, S. F., “Standard Model: An Introduction”, 2000, arXiv: hep-ph/0001283.
- [14] Beringer, J. et al. (Particle Data Group), “The Review of Particle Physics, Higgs Bosons: Theory and Searches”, *Phys. Rev. D* 86 (2012) 010001, URL: <http://pdg.lbl.gov>.
- [15] S. Dittmaier et al., “Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables”, 2011, arXiv: 1101.0593v3.
- [16] R. Cahn and G. Goldhaber, *The Experimental Foundations of Particle Physics*, 2nd ed., Cambridge University Press, 2009.

- [17] M. Cacciari and G. P. Salam, “The anti- k_t jet clustering algorithm”, 2008, arXiv: 0802.1189v2.
- [18] Beringer, J. et al. (Particle Data Group), “The Review of Particle Physics, Particle listings”, *Phys. Rev. D* 86 (2012) 010001, URL: <http://pdg.lbl.gov>.
- [19] ATLAS Collaboration, “ b -jet tagging calibration on c -jets containing D^{*+} mesons”, ATLAS-CONF-2012-039, CERN, Mar. 2012, URL: <http://cds.cern.ch/record/1435193>.
- [20] ATLAS Collaboration, “Commissioning of the ATLAS high-performance b -tagging algorithms in the 7 TeV collision data”, ATLAS-CONF-2011-102, CERN, July 2011, URL: <https://cds.cern.ch/record/1369219>.
- [21] T. Sjöstrand, “Monte Carlo Generators”, 2006, arXiv: hep-ph/0611247.
- [22] B. R. Webber, “Fragmentation and Hadronization”, 1999, arXiv: hep-ph/9912292.
- [23] I. Antcheva and O. Couet, eds., *ROOT Users Guide 5.21*, 2008, URL: <http://root.cern.ch/drupal/content/users-guide#UG>.
- [24] P. I. Bonn, *Overkill Twiki page - main Overkill manual*, URL: https://wiki.physik.uni-bonn.de/atlas/private/index.php/Overkill_Twiki_page.
- [25] ATLAS Collaboration, “Search for the $b\bar{b}$ decay of the Standard Model Higgs boson in associated (W/Z)H production with the ATLAS detector”, ATL-CONF-2013-079, CERN, July 2013, URL: <http://cds.cern.ch/record/1563235>.
- [26] G. James et al., *An Introduction to Statistical Learning*, 1st ed., Springer, 2013.
- [27] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, 2nd ed., Springer, 2009.
- [28] A. Hoecker et al., *TMVA 4 - Toolkit for Multivariate Data Analysis with ROOT, Users Guide*, CERN-OPEN-2007-007, 2009, arXiv: physics/0703039.
- [29] T. Aaltonen et al., “Improved b -jet Energy Correction for $H \rightarrow b\bar{b}$ Searches at CDF”, 2011, arXiv: 1107.3026.
- [30] CMS Collaboration, “Search for the standard model Higgs boson produced in association with W or Z bosons, and decaying to bottom quarks (ICHEP 2012)”, CMS-PAS-HIG-12-019, CERN, July 2012, URL: <https://cds.cern.ch/record/1460692>.
- [31] C. Callenberg, “Rekonstruktion der Higgs-Masse im Kanal $WH, H \rightarrow b\bar{b}$ mit multivariaten Methoden”, Universität Bonn, Aug. 2012.
- [32] T. Sjöstrand, S. Mrenna and P. Skands, *A Brief Introduction to PYTHIA 8.1*, CERN-LCGAPP-2007-04, 2007, arXiv: 0710.3820.
- [33] D. Lopes-Pegna, “Search for the SM Higgs Boson decaying to $b\bar{b}$ at CMS”, High Energy Physics Seminar at University of Bonn and private communication.
- [34] ATLAS Collaboration, “Search for the Standard Model Higgs boson in associated production with a vector boson and decaying to bottom quarks with the ATLAS detector”, ATL-COM-PHYS-2013-465, CERN, July 2013, URL: <https://cds.cern.ch/record/1543067>.
- [35] J. Therhaag and E. von Toerne, “Discussions with two of the TMVA developers about TMVA features”, private communication.
- [36] A. Bukin, “Fitting function for asymmetric peaks”, 2007, arXiv: 0711.4449v2.

-
- [37] CMS Collaboration, *Search for the standard model Higgs boson produced in association with W or Z bosons, and decaying to bottom quarks*,
URL: <https://twiki.cern.ch/twiki/bin/view/CMSPublic/Hig13012TWiki>.
 - [38] A. Buzatu et al., “Invariant Mass Studies for the $H \rightarrow b\bar{b}$ Measurements for HCP”,
ATL-COM-PHYS-2012-1451, CERN, Nov. 2012,
URL: <https://cds.cern.ch/record/1482360>.
 - [39] V. Kostyukhin, *First bb -mass correction*,
Bonn Standard Model Meeting, 18. October 2012 and private communication,
URL: <https://indico.cern.ch/getFile.py/access?contribId=1&resId=1&materialId=slides&confId=213579>.
 - [40] Aad, G. et al., “Invariant Mass Studies for the $H \rightarrow b\bar{b}$ Measurements for LHCP”,
ATL-COM-PHYS-2013-449, CERN, Aug. 2013,
URL: <https://cds.cern.ch/record/1541965>.
 - [41] ATLAS Collaboration,
“Jet energy measurement with the ATLAS detector in proton-proton collisions at $\sqrt{s} = 7\text{TeV}$ ”,
2011, arXiv: 1112.6426.
 - [42] R. J. Barlow, *Statistics, A Guide to the Use of Statistical Methods in the Physical Sciences*,
1st ed., John Wiley & Sons Ltd., 1997.
 - [43] A. Bevan and F. Wilson, eds., *AFit User Guide*, 2010,
URL: <http://pprc.qmul.ac.uk/~bevan/afit/>.

List of Figures

2.1	Schematic proton-proton collision	5
2.2	LHC accelerator chain	6
2.3	Cross sections for proton-(anti)proton collisions	7
2.4	Structure of the ATLAS detector and particle signatures	9
2.5	Higgs decay channels in dependence of m_H	12
2.6	Higgs production processes at the LHC, $W^\pm H \rightarrow l^\pm \nu b \bar{b}$	13
3.1	Fragmentation of jets	18
4.1	Schematic picture of a decision tree for regression	25
5.1	Truth $m_{b\bar{b}}$, releases 17.0.6.4.7 (2011) and 17.2.7	29
5.2	$m_{b\bar{b}}$ for different mixtures of truth and reconstructed quantities in the b vectors	30
5.3	WH , $E_T^{\text{jet}} - E_T^{\text{true}}$ before and after training separate for training and test tree	37
5.4	$WH125$, $m_{b\bar{b}}$ before and after the regression	39
5.5	CMS $WH125$ results: $m_{b\bar{b}}$ before and after the regression	40
5.6	$ZH125$, $m_{b\bar{b}}$ before and after the regression	41
5.7	Wbb , $m_{b\bar{b}}$ before and after the regression and on truth level	42
5.8	$t\bar{t}$, $m_t^{\text{lep.}}$ before and after the regression	42
5.9	$m_{b\bar{b}}$ for 2011 data and Monte Carlo samples	44
5.10	$WH125$, $m_{b\bar{b}}$ after the regression and after the $\mu + p_T^{\text{reco}}$ correction	45
5.11	$WH125$, $m_{b\bar{b}}$ after the regression and after the alternative $\mu + p_T^{\text{reco}}$ correction	46
5.12	$ZH125$, $m_{b\bar{b}}$ comparison for the regression and the GSC and KL corrections	48
5.13	Performance of the regression in dependence on the added input variables	49
5.14	Performance of the regression in dependence on the added input variables, zoom	50
5.15	$WH110$, invariant mass distribution of the truth jets	51
5.16	$WH125$, $m_{b\bar{b}}$ after the regression and after the regression with the truth jet E_T as a target	51
5.17	$WH125$, $m_{b\bar{b}}$ after the regression and after the regression with the CDF input variables	52
5.18	$WH125$, $m_{b\bar{b}}$ after the regression and after the regression with the CMS input variables	53
A.1	ZH , $E_T^{\text{jet}} - E_T^{\text{true}}$ before and after training separate for training and test tree	62
A.2	Histogram for the p_T^{reco} correction	62
A.3	Transverse W mass	67
A.4	Input variables 1	68
A.5	Input variables 2	69
A.6	Input variables 3	70

List of Figures

A.7	Input variables 4	71
A.8	Input variables 5	72
A.9	Input variables 6	73

List of Tables

3.1	Selections for the standard $V, H \rightarrow b\bar{b}$ analysis	22
5.1	List of available Monte Carlo samples and their features	29
5.2	List of input variables for the regression	33
5.3	Selections for the regression	35
5.4	Objects associated to the jets	37
6.1	$W^\pm H \rightarrow l^\pm \bar{\nu} b\bar{b}$, $m_H = 125$ GeV, results for the $m_{b\bar{b}}$ resolution	56
A.1	$m_{b\bar{b}}$ for different mixtures of truth and reconstructed quantities in the b vectors	63
A.2	$E_T^{\text{jet}} - E_T^{\text{true}}$ before and after regression separate for training and test tree	63
A.3	Ranking of input variables for the regression	64
A.4	Results for the test of the importance of the input variables	65
A.5	Used Monte Carlo samples and their scale factors	66