Fast preselections for cross-section measurements and Higgs search at the ATLAS experiment

Diplomarbeit

vorgelegt der

Fakultät für Physik

der

Ludwig-Maximilians-Universität München

von

Christoph Bußenius

München, den 4. Juni 2010

Erstgutachterin: Prof. Dr. Dorothee Schaile
Zweitgutachter: Dr. Hans von der Schmitt

Abstract

The Large Hadron Collider (LHC) is the most modern particle accelerator and is in operation since 2009. Its detector experiments have great goals, e. g. to detect the Higgs boson. The "ATLAS" detector, which is one of the multi-purpose detectors at LHC, is already recording a huge amount of data. Many of the processes that are being analyzed have very small cross-sections and strong background, requiring sophisticated selection algorithms and several data processing iterations.

The data is indexed by so-called TAGs, which give a very short summary of the recorded measurements. A restricted set of physical quantities is available in TAGs. This diploma thesis discusses how these quantities can be used for physics analyses.

Using TAGs, the cross-section of the $Z \to \mu^- \mu^+$ decay will be measured. This requires an event selection for dimuon events and the determination of the integrated luminosity $\int \mathscr{L}\,\mathrm{d}t$.

For the analysis of the Higgs decay $H \to W^+W^- \to \ell\nu\bar{\ell}\bar{\nu}$, TAGs can be used to build an event preselection. Preselections for this process use criteria such as the missing transverse energy and the muon isolation, which are only provided by TAGs in a limited form. The implications for the analysis will be discussed.

Zusammenfassung

Der Large Hadron Collider (LHC) ist der modernste Teilchenbeschleuniger der Welt und wurde 2009 in Betrieb genommen. Seine Detektorexperimente haben hohe Ziele, wie die Entdeckung des Higgs-Bosons. Der „ATLAS“-Detektor ist einer der Universaldetektoren am LHC und hat schon gigantische Mengen an Daten aufgenommen. Viele der analysierten Prozesse haben sehr kleine Wirkungsquerschnitte und starken Untergrund. Dadurch werden ausgefeilte Selektionsalgorithmen und Datenverarbeitung in mehreren Durchläufen notwendig.

Die Daten werden durch sogenannte TAGs indexiert. Diese enthalten eine kurze Zusammenfassung der aufgenommenen Messungen, wobei nur eine beschränkte Menge an physikalischen Größen verfügbar ist. Die vorliegende Diplomarbeit untersucht, wie diese Größen für Physikanalysen verwendet werden können.

Mithilfe von TAGs wird der Wirkungsquerschnitt des Zerfalls $Z \to \mu^- \mu^+$ gemessen werden. Dazu ist eine Ereignisselektion für Zwei-Myonen-Ereignisse notwendig und die Bestimmung der integrierten Luminosität $\int \mathscr{L}\,\mathrm{d}t$.

Für die Analyse des Higgszerfalls $H \to W^+W^- \to \ell\nu\bar{\ell}\bar{\nu}$ kann man mit TAGs eine Vorselektion von Ereignissen erstellen. Vorselektionen für diesen Prozess benutzen Kriterien wie die fehlende Transversalenergie und die Isolation von Myonen. Diese Information ist in TAGs nur in einer begrenzten Form enthalten. Die Auswirkungen dessen auf die Analyse werden diskutiert werden.

Contents

Chapter 1

Introduction

1.1 ATLAS

ATLAS[1] is one of the particle detectors that are situated at the LHC. ATLAS and CMS[2] are the two largest of the six detectors. The detector is the heart of the ATLAS experiment, in which scientists from 173 universities and other institutions in 37 countries participate [1].

1.1.1 LHC

The Large Hadron Collider (LHC) at CERN is the most modern particle accelerator as of 2010. It is designed to collide protons at center-of-mass energies $\sqrt{s} = 14$ TeV and lead ions at 2.76 TeV per nucleon. Currently, LHC can collide protons at 7 TeV.

LHC is located below the ground of Switzerland and France in the tunnel that formerly contained the Large Electron-Positron (LEP) collider (cf figure 1.1). LEP had been running until 2000 with center-of-mass energies ranging from 88 to 209 GeV.

LEP could not be operated at much higher energies because of the increasing energy loss due to synchrotron radiation. The power of the radiation declines with the fourth power of the mass of the accelerated particles, thus it poses a much smaller problem for proton beams. Each proton loses 6.7 keV per rotation to synchrotron radiation at the LHC design energy [2].

The protons at LHC, which are extracted from hydrogen gas, are accelerated in bunches of $\sim 10^{11}$. The tunnel contains pipes for two separate beams, one in each direction.

[1] **A** **T**oroidal **L**HC **A**pparatus

[2] **C**ompact **M**uon **S**olenoid

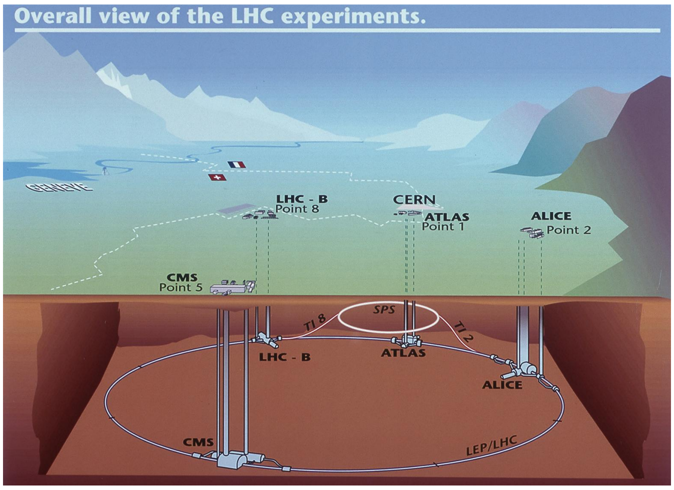


Figure 1.1: The LHC tunnel with the main experiments [3]

1.1.2 Physics at LHC, the Higgs particle

The *Standard Model* has been very successful as a theoretical foundation of particle physics and has been verified by many experiments: it predicted the existence of W and Z bosons, gluons, top and charm quarks before they have been discovered. There is only one particle in the Standard Model that has not been observed in experiments yet, which is the Higgs boson H.

Experiments with LEP and Tevatron[3] have failed to observe the Higgs particle so far. LEP has excluded Higgs masses below 114.4 GeV, and Tevatron has excluded the range 160–170 GeV, both with 95% confidence level. LHC is trying to find the Higgs particle in a wider mass range using the multi-purpose detectors ATLAS and CMS.

Further research includes precision measurements of Standard Model parameters like the mass of the top quark, and the search for SUSY (supersymmetry).

Also, there are some smaller and more specialized experiments at LHC: The LHCb[4] experiment is dedicated to the physics of B mesons, especially the study of CP violation. ALICE[5] will study the quark-gluon plasma using heavy ion collisions.

[3] Tevatron is a $p\bar{p}$ accelerator at Fermilab in Batavia, Illinois, USA. It is currently operating at $\sqrt{s} = 1.96$ TeV.

[4] **LHC b**eauty

[5] **A** **L**arge **I**on **C**ollider **E**xperiment

1.1.3 Luminosity measurement

The luminosity is an important quantity in the accelerator. The design luminosity of LHC is 10^{33} cm^{-2}s^{-1} at $\sqrt{s} = 14$ TeV. Precise knowledge of the luminosity is important for cross-section measurements (cf chapter 2) and many physics analyses.

The luminosity can be determined from beam parameters as

$$\mathscr{L} = \frac{n_1 n_2 f}{4\pi\sigma_x\sigma_y} \tag{1.1}$$

where n_1, n_2 are the number of protons in two colliding bunches, f is the revolution frequency, and σ_x, σ_y are the transverse bunch widths.

At ATLAS, the luminosity is monitored using the LUCID[6] subdetector. The underlying principle here is that the number of interactions in a bunch-crossing is proportional to the number of detected particles. This monitoring can only measure relative changes in luminosity. It must be calibrated using absolute luminosity measurements, which is done by the ALFA[7] subdetector. However, ALFA can only be used under special preconditions, i. e. special calibration runs of LHC are necessary [8, chap. 13].

1.1.4 Detector components

Figure 1.4 provides a schematic view of the detector. ATLAS is of a cylindrical shape with 44 m length and a radius of 11 m. At the axis is the proton beam pipe that is part of LHC. As the rings of LHC lie below the ground, so does the detector.

The very center of ATLAS is the interaction point where the proton beams collide. Most components are layered around the beam pipe like an onion. As can be seen in the figure, the detector is divided into a barrel region and two end-caps.

There are three principal subdetector systems. The *Inner Detector* is closest to the interaction point; the *Calorimetry System* is built around the Inner Detector; the *Muon Spectrometer* takes up most of the volume of ATLAS, being the thickest and outermost shell.

The following sections will describe each of the subdetectors. Before that, the coordinate system is defined, and the Magnet Systems are described.

Coordinate system

Accurate description of the detector and physics analyses requires that all participating physicists agree on a coordinate system. The beam direction (a tangent to

[6] **L**uminosity **M**easurement Using **C**herenkov **I**ntegrating **D**etector

[7] **A**bsolute **L**uminosity **F**or **ATLAS**

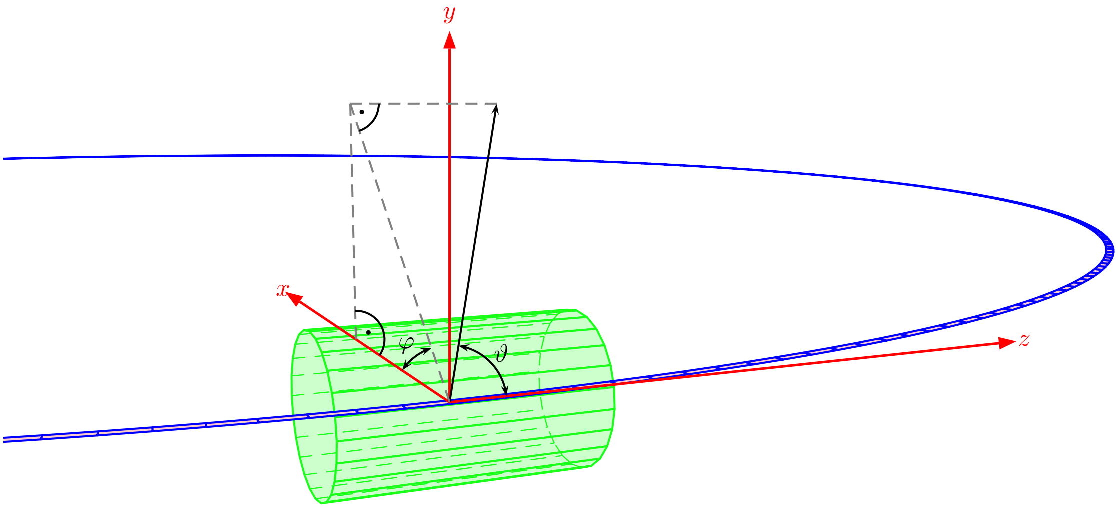


Figure 1.2: LHC and ATLAS with the coordinate system. The scale of the ring is 10 times smaller than the scale of the detector.

the LHC ring) is used as the z axis with the interaction point as the coordinate system origin. In spherical coordinates, ϑ is the polar angle from the beam axis and φ is the azimuthal angle with $\varphi = 0$ pointing toward the ring center (cf figure 1.2).

Another coordinate, η, the *pseudorapidity,* is often used as a substitute for ϑ. It is defined as

$$\eta = -\ln \tan \frac{\vartheta}{2}. \tag{1.2}$$

The ϑ range from 0 to π corresponds to an η range from ∞ to $-\infty$. However, $|\eta|$ approaches infinity only in the region very close to the z axis where $\vartheta = 0$ or $\vartheta = \pi$. As the z axis is the beam direction, the detector coverage is poor for these ϑ regions, and they are dominated by background. Usually one restricts the range to

$$7.7^\circ < \vartheta < 172.3^\circ \tag{1.3}$$

which corresponds to

$$2.7 > \eta > -2.7. \tag{1.4}$$

Figure 1.3 illustrates the relation between ϑ and η.

Note that $\eta = 0$ is the transverse plane (i. e. the xy plane), and that η is symmetric to this plane.

From equation 1.2 it follows that if ϑ is the polar angle corresponding to a momentum vector $\vec{p}$, then

$$\eta = \frac{1}{2} \ln \frac{|\vec{p}| + p_z}{|\vec{p}| - p_z}. \tag{1.5}$$

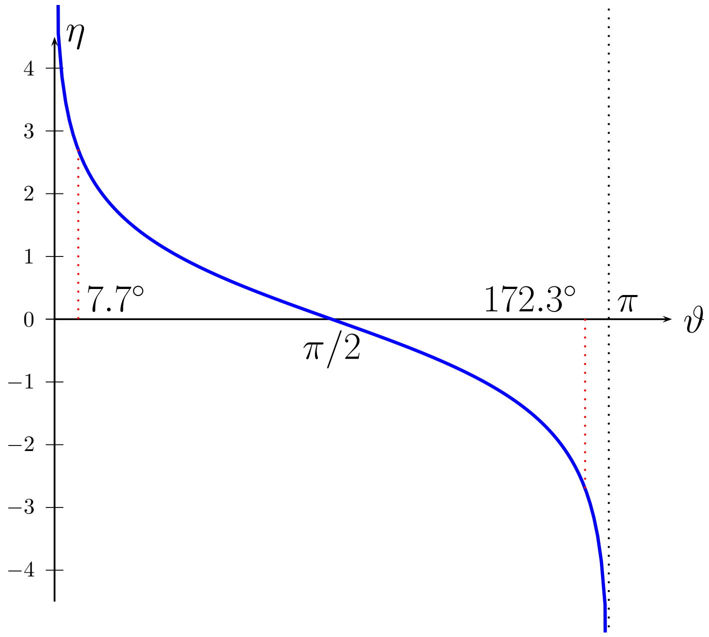


Figure 1.3: ϑ and η

In high-energy approximation, η is equal to the *rapidity* [4], which is defined as

$$Y = \frac{1}{2} \ln \frac{E + p_z}{E - p_z} \tag{1.6}$$

with ΔY being a Lorentz-invariant quantity.

The quantity $\Delta R := \sqrt{\Delta\varphi^2 + \Delta\eta^2}$ is used to specify distances between particles or jets and for opening angles of cones.

Magnet system

The path of a charged particle will be bent if it moves through a magnetic field. If the field strength and the charge are known, the radius of curvature may be used to calculate the momentum component transverse to the field. Stronger fields imply smaller curvature radii, which may be measured with higher accuracy. There are two strong magnetic fields in the ATLAS detector:

The Inner Detector is surrounded by the field of the **central solenoid.** It is provided by solenoidal superconducting magnets that operate at a temperature of 4.5 K. The cooling is provided by the same cryostat that is used for the calorimeter. The field strength varies between 2 T at the interaction point and 0.5 T at the ends farthest from the transverse plain. The inhomogeneity is due to the dimensions of the solenoid, which is 80 cm shorter than the Inner Detector.

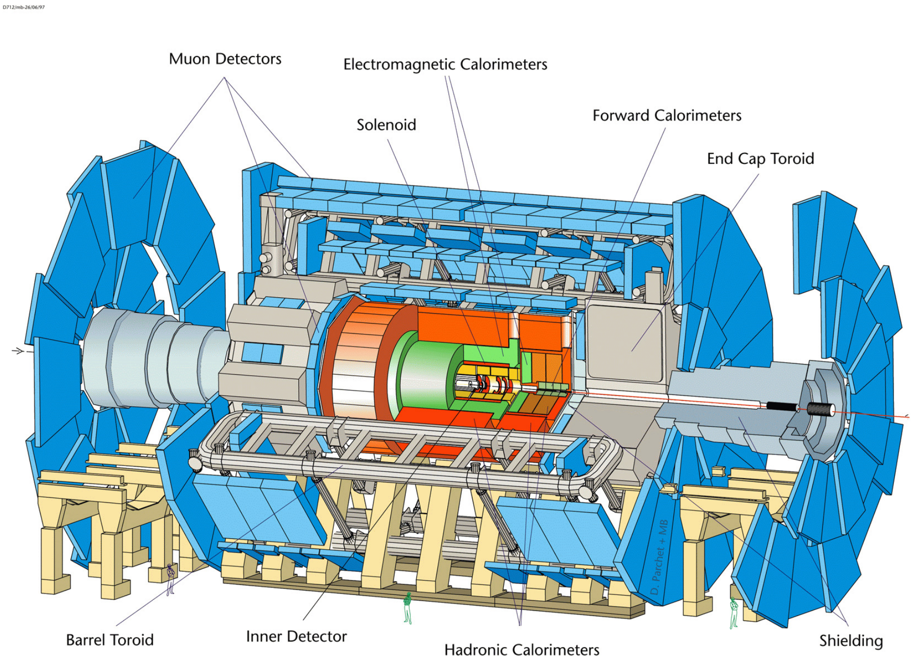


Figure 1.4: The ATLAS detector [8]

The Muon System is provided with a magnetic field by one toroid in the barrel region and two in the end-caps. Each toroid consists of eight coils. This **toroid magnet system** generates a magnetic field with 0.5 T strength on average and toroidal field lines. To minimize multiple scattering effects, the coils are surrounded by air rather than having an iron core.

Inner Detector

The task of the Inner Detector is precise measurement of particle trajectories *(tracks),* which are used for the determination of momenta, charges and impact parameters. Also, the direction of flight allows for matching the tracks with measurements of other detector components.

High track resolution is imperative to cope with $\sim$ 1000 particles that are produced every 25 ns in the $|\eta| < 2.5$ region covered by the Inner Detector. This is realized by a combination of three subdetectors. The inner part is the *Pixel Detector,* which is built of three layers of 50 µm × 300 µm silicon pixels, each providing one track point. The second part consists of 80 µm × 300 cm silicon strip detectors. The last part is built of straw tubes that determine about 36 more track points. In good conditions, the Inner Detector can resolve the transverse momentum of a muon with $p = p_\mathrm{T} = 20$ GeV to 1.5% [8, p. 60].

Calorimetry System

Calorimeters measure the energy of particles by absorbing them. ATLAS uses two calorimeters, which envelop the Inner Detector. The inner shell is the **electromagnetic calorimeter** and absorbs electrons and photons in lead absorber plates. The EM calorimeter uses a cryostat to keep it sufficiently cool. Liquid argon is used to detect the showers that result from particle interactions with matter. In the **hadronic calorimeter**, iron absorbs particles that interact through the strong force. Showers in the hadronic calorimeter are detected using scintillators.

For the determination of the missing energy, a large η-coverage of the calorimeters is required. The parts of the calorimetry system located in the barrel and end-cap regions cover $|\eta|$ ranges up to 3.2. For the range $3.1 < |\eta| < 4.9$, a special forward calorimeter is used.

Muon Spectrometer

Apart from neutrinos, the only particles that reach the Muon System are muons with energy $\gtrsim 6$ GeV.

Muons are produced in many interesting processes at ATLAS. Therefore much effort has been put into building a sophisticated Muon Spectrometer. Precise measuring of position and momentum (via the curvature of muon tracks) is realized using **Monitored Drift Tube** (MDT) chambers. These are positioned in several layers in the barrel region and in the end-cap region. The tubes are filled with a mixture of Argon and CO_2. In the region near the beam axis, strong background is expected; therefore **Cathode Strip Chambers** (CSC) are used in the innermost layer.

The Muon System also houses trigger elements. **Resistive Plate Chambers** (RPC) are used in the barrel region. Muons passing through these chambers lead to a fast discharge, which serves as a trigger. In the end-caps, **Thin Gap Chambers** (TGC) are used as trigger elements.

The Muon System covers an η range from -2.7 to 2.7.

Triggers

Protons in LHC will circulate with a frequency of 40 MHz. At the planned luminosity, the rate of proton interactions (*events*) will be ~ 1 GHz, yet the rate of interesting events is much smaller.

The data recorded amounts to the order of 1 MB for one event. Obviously, it is neither possible nor desirable to store every single event. This leads to the need for triggers to make fast decisions whether or not to record an event. The ATLAS trigger system is divided into three levels:

Level 1 triggers are hardware-based and built into the Calorimeter and Muon System. They have a latency of only 4 µs. Level 1 decides which events might be interesting and defines *regions of interest* for those events. Regions of interest specify in which part of the detector possibly interesting objects may be found for the event. This information is passed on to the second level at a rate of 75 kHz, which allows for higher latency in the next trigger level.

Level 2 triggers are software-based and reduce the event rate to 1 kHz.

The final level is called *Event Filter.* It is software-based and uses a computer farm near the detector. Its target rate is 200 Hz.

Each level contains triggers for a number of possibly interesting conditions. The *trigger menu* defines what triggers are available. Examples are

- A trigger that checks for a muon with $p_{\mathrm{T}} > 6$ GeV. This trigger is called `L2_mu6` for level two and `EF_mu6` for the event filter.
- A trigger that checks for *two* muons with $p_{\mathrm{T}} > 6$ GeV, called `L2_2mu6` or `EF_2mu6`.

Some conditions may be interesting but still occur at higher rates than needed. In those cases, triggers are *prescaled,* i. e. only a fraction (prescale factor) of the triggered events is recorded.

1.1.5 Reconstruction

Only a number of particles produced at ATLAS are stable enough to have a chance to leave or even reach the detector, as opposed to decaying in the beam pipe. These stable end products are electrons, muons, photons, charged pions and kaons, and neutrinos. The latter may only be detected indirectly by means of missing momentum and energy.

Because of the longevity and good detectability of electrons and muons at ATLAS, these two kinds of particles (and their antiparticles) are of great importance for physics analyses.

Apart from leptons[8], an important phenomenon that can be detected is known as *jets.* Partons (quarks and gluons) cannot be separated into free particles due to their color confinement. In an attempt to do so, the energy density between them will increase until it suffices to create new partons. These combine into baryons and mesons, which is known as hadronization. The new particles form boosted ensembles known as jets.

[8] The term *leptons* will be used for muons and electrons

It should be noted that both jets and missing energy are difficult to measure. Many variables have to be considered in order to calculate the total energy, and to decide how to assign individual particles to jets. Software for this task kept evolving after the technical devices have been put in place, and there are many versions of that software.

Most physics analyses use the above-mentioned end products and the associated quantities (e. g. p_{T} of a jet) as input. However, they are not directly measured in the detector. A series of software algorithms is used to reconstruct the events.

Electrons

The standard algorithm for electron reconstruction is seeded from the electromagnetic calorimeters. It starts from the clusters reconstructed in the calorimeters and tries to assign a track from the Inner Detector to it [6].

Several standard cuts are done for electrons during reconstruction. The cuts are divided into groups, known as *loose, medium* and *tight* cuts.

Loose cuts provide excellent identification efficiency, i. e. few true electrons evade these cuts. However the background rejection is low, so many loose electrons are false reconstructions. Tight cuts on the other hand provide the highest rejection of background.

- Loose cuts are solely based on calorimeter information, e. g. the hadronic leakage, which is the ratio of the reconstructed energy in the hadronic calorimeter to that in the electromagnetic calorimeter.

- Medium cuts use track variables too, like the number of hits in the pixel detector. This is mainly used to reduce the $\pi^0 \Rightarrow \gamma\gamma$ background.

- Tight cuts use all available particle-identification tools, like a cut on the difference between the real track to the track that is extrapolated from the cluster.

Electrons are stored in a collection known as Electron AOD Container. Each electron stored therein is flagged according to which series of cuts it has passed. Electrons with the “loose” flag are also called “loose electrons”; likewise for medium and tight.

Note that the definition of these terms may change from one ATLAS software release to another.

Figure 1.5 may be used to get an overview on the relative occurrences of the electrons available for the different definitions.

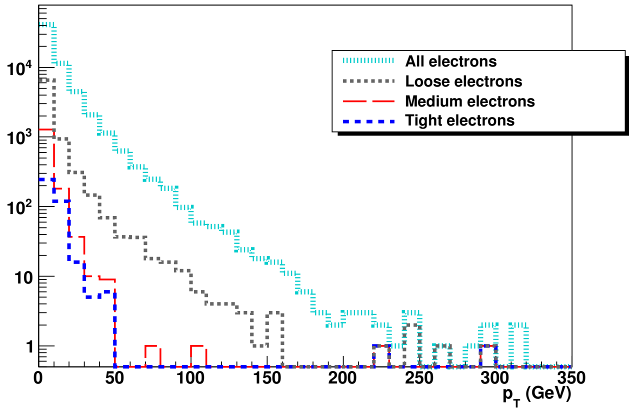


Figure 1.5: Distribution of p_{T} for reconstructed electrons of different tightness (FDR2 run 52290)

Muons

As muons are the only particles detected in the Muon System, they are easy to identify. For the reconstruction of their properties, the measurements from the Muon System may be *combined* with those of the Inner Detector by matching the muons with a nearby track, which improves the precision of the momentum measurement for muons with 30 GeV $\lesssim p_{\mathrm{T}} \lesssim$ 200 GeV [6, p. 163].

On the other hand, *standalone* muons are reconstructed from Muon Spectrometer measurements alone. They have higher p_{T} acceptance and are not restricted to the $|\eta| < 2.5$ boundary of the Inner Detector. Standalone muons may also include muons that are produced in the calorimeter rather than at the interaction point. This is not true for combined reconstruction because those fake muons do not have any tracks in the Inner Detector.

Two families of algorithms are available for the reconstruction of muons, known as *STACO*[9] and *MuID*. They differ in the methods used for combined and standalone reconstruction [6, 9].

Isolation variables

Leptons produced in a decay of W or Z bosons occur isolated from hadronic activity. This is a property that is often used to separate these leptons from those produced in other processes. The isolation can be formalized by a number of variables that are determined during reconstruction. In [19], detailed studies of isolation variables for muons can be found.

[9] **Sta**tistical **Co**mbination

Calorimeter isolation: The quantity $E_{T,\text{cone}}$ denotes the energy that is registered by the calorimeter cells within a cone around the respective particle, decreased by the energy of the particle itself. Several cone sizes are used, e. g. for cones with $\Delta R = 0.03$, the variable is named $E_{T,\text{cone},30}$.

Track isolation: The Inner Detector provides another isolation variable. The track isolation of a particle is the sum of the transverse momenta of all tracks within a cone around the particle, not counting the own track of the particle.

Jets

The most important detector component for jet reconstruction are the calorimeters.

Many algorithms have been developed for the combination of particles to jets [17]. One of the widely used ones is the ATLAS *Cone* algorithm, which aims to maximize energy in a geometric cone. Depending on the analysis, narrow or wide jets may be preferred, which use cone sizes of $\Delta R = 0.4$ and 0.7, respectively. The algorithm is iterative, i. e. when a candidate cone is identified, its momentum vector is calculated and the cone is redrawn around the new center. The first candidate cones are initiated by *seeding* the algorithm with calorimeter objects. Several kinds of objects may be used, e. g. calorimeter towers or topological clusters. Towers are formed by projecting a grid of $\eta \times \varphi = 0.1 \times 0.1$ onto the calorimeter cells. The concept of topological clusters has been created as an attempt to reconstruct three-dimensional energy depositions in the calorimeters.

The identification of jets originating from b quarks ("b-jets") is important for many analyses, for instance to veto the $t\bar{t}$ background. Due to the relatively long lifetime of hadrons containing a b quark, b-jets typically give rise to secondary vertices a few millimeters away from the primary production vertex. So-called *b-tagging* techniques [18] try identify these vertices by assigning a b-tagging weight to the corresponding jets. One such technique is "SV0", which fits the secondary vertex and returns the significance of the decay length of the secondary vertex. The efficiency of b-tagging is difficult to estimate from simulations and will have to be determined using real collision data.

Missing transverse energy

The missing transverse energy $\not{E}_\mathrm{T}$ is calculated during reconstruction as the negative vector sum of all measured and estimated momenta in the calorimeter [6]. A precise calculation has to take into account the resolution of the detector, sources of noise, and the fact that the detector has limited coverage and spatial expansion.

As the last step of the $\not{E}_{\mathrm{T}}$ reconstruction, a refined calibration is done by associating the calorimeter cells with the reconstructed high-p_{T} objects (which are electrons, photons, muons, hadronically decaying τ-leptons, b-jets and light jets) that they are assigned to. The refined missing energy is designated as `MET_RefFinal`.

For the muon contribution, combined muons are used for $|\eta| < 2.5$. Standalone muons are of lower quality for this task because of the poor p_{T} resolution. Nevertheless, they are used for $2.5 < |\eta| < 2.7$ where no combined muons are available.

Albeit fixed in current versions of the ATLAS software, there was a bug that led to the inclusion of standalone muons regardless of η, which substantially affected the $\not{E}_{\mathrm{T}}$ resolution. To correct this error, the Higgs working group used a special algorithm to correct the muon term [30]. The corrected version of the missing energy is known as `MET_RefFinal_corrected`.

1.2 ATLAS Computing System

Each year, the LHC experiments produce data in the order of 10 Petabytes. A large computing infrastructure and the ATLAS software suite have been created to cope with this large amount of information.

The data formats used are defined in [10] and will be summarized in the following:

1.2.1 Data formats

RAW

After passing the Event Filter, the events will be output in a format known as *RAW,* which contains basically the readout information of the subdetector components, such as hits in the tracking detectors or calorimeter cells. RAW data are about 1.6 MB per event, bundled in files up to 2 GB. The reconstruction process (described in section 1.1.5) can use either real detector data or simulated Monte Carlo data as its input.

ESD

The resulting data format is known as *ESD* (Event Summary Data). Apart from the results of the reconstruction process, ESD still contains enough information to access the original RAW data for re-reconstruction purposes. ESD occupies about 500 kB per event.

AOD

AOD (Analysis Object Data) is a reduced representation that is derived from ESD. Its target size is 100 kB per event. The AOD format is the base for physics analysis. It contains information about all reconstructed objects, such as leptons, jets, tracks, vertices, missing energy. Most of the information is included in several versions, e. g. from different reconstruction algorithms. The *Athena* software framework is a powerful tool to develop analyses based on AOD data. It allows for access to the reconstructed variables as well as the inclusion of other algorithms. Athena is also suited for running on remote Grid elements.

TAG

With 100 kB per event, AOD files are rather large. Sometimes, only basic properties of the events in a run need to be assessed. *TAG* files are designed for this task. Their size is $\sim$ 1 kB per event (for collision data) as they contain only a very short summary *(metadata)* of events. They are produced during reconstruction along with ESD and AOD files, although they contain only information that may be gathered from AOD alone. The TAG information is not only available in files but also in the *TAG database.* The most convenient way to access the TAG database is “ELSSI” (Event Level Selection Service Interface), which is a web site that facilitates composing and submitting queries to the TAG database in the Standard Query Language (SQL). The user is provided with counts, histograms of basic quantities (such as p_{T}, $\not{E}_{\mathrm{T}}$), or downloads in form of TAG files.

TAG queries can reduce datasets to a subset that can be used for further analysis. For this task, a useful feature of the Athena framework is *back navigation:* After determining relevant events by using either TAG files or the database, AOD or ESD information is accessed for these events only. This is possible because the events in TAG files store information about the names (represented as strings of hexadecimal digits called GUID, **G**rid **u**nique **id**entifiers) of the original data streams. Given a TAG sample (such as event selections produced by ELSSI), Athena is able to locate the events in AOD or ESD files. The grid software Ganga (described below) also supports this, i. e. it can use a TAG selection in order to ensure that software runs on locations where the original data for the selected events is available.

1.2.2 Distributed analysis

The enormous storage and processing requirements of the large data flow cannot be accommodated by a single site. The WLCG (Worldwide LHC Computing Grid) provides distributed storage facilities and the necessary CPU power for studies of real and simulated data.

A hierarchical structure is used for data distribution, called a *Tier* structure. It implements redundant data storage, minimizing single points of failure. As a first step, the data is recorded at a **Tier-0** center at CERN. It is then distributed to 10 **Tier-1** centers worldwide. They store primarily RAW and ESD data. The German center is GridKa, which is located at KIT (Karlsruhe Institute of Technology). **Tier-2** centers are dedicated to user-specific analyses and store mainly AODs and DPDs (which are smaller files built from AODs). The lowest level in the hierarchy is **Tier-3**, which consists of individual workstations and small computer clusters.

To minimize copying of large datasets and therefore the bandwidth use of wide-area networks, a Grid policy requires that jobs be sent to data, i. e. distributed analysis programs run only on those computer systems that have the required input data available in nearby computer centers, as opposed to running on arbitrary grid elements, which would require downloading datasets before the job can start.

In order to run computations on the Grid, a *middleware* is used. This is a software that decides which computing elements to assign jobs to, taking care of load balancing and user-defined requirements on e. g. dataset availability or software versions. The middleware submits jobs for computation by contacting servers known as WMS (Workload Management Systems).

The most convenient way to access the middleware is to use a frontend such as "Ganga". Ganga and the middleware assist in various tasks such as extending the submitted code with functionality to access input data and make the output available to the user.

1.2.3 Monte Carlo studies

Particle interactions usually have many degrees of freedom. They are governed by quantum mechanical laws and can be regarded as random experiments. In many cases, the experimental verification of a theory requires comparing the statistical distributions of the outcomes. Monte Carlo (MC) methods are an important technique to understand the processes.

The first step to produce simulated data is the event generation. In this step, four-vectors for a specified primary interaction are produced and the full chain of decays and hadronization is generated. Several programs are used for this task, e. g. Pythia [11], MC@NLO [12], Alpgen [13], Herwig [14].

The next step is the simulation of the interaction between the generated particles and the detector. This is done by GEANT4 [15], a software platform for the simulation of the passage of particles through matter.

The ATLAS project uses simulated ESD and AOD datasets for Monte Carlo studies. These are also used as a test case for the ATLAS software suite and analysis programs.

FDR

FDR (Full Dress Rehearsal) [20, 21] was an effort in 2008 to test the full data processing chain, including reconstruction and distribution through the Tier system. For this purpose, data samples have been prepared from a mix of several simulated physics processes. The goal was to produce a realistic physic mixture, with a primary focus on useful trigger rates. FDR took place in three phases (0, 1 and 2). The most accurate data has been produced in phase 2 ("FDR2").

Similar to real data, the FDR data are available in different *streams*. Events are classified into one or more streams according to which triggers have fired. For analyses based on end states containing muons, the so-called Muon stream can be used.

1.3 TAG files in detail

1.3.1 Contents

Tables A.1 and A.2 list the contents of the TAG files in version 14 of the ATLAS software. It should be noted that the specification is subject to constant change. In version 15, some variables have been renamed or redefined as summarized in table A.3. TAG files that have been produced a while ago—like those for FDR2—use older versions of the specification.

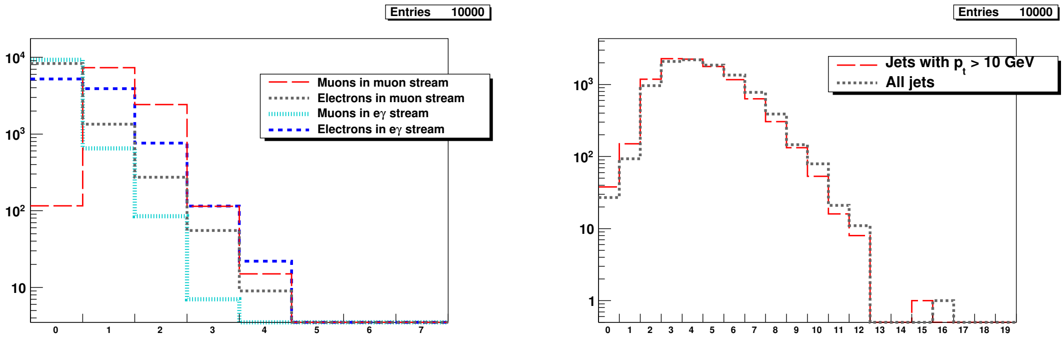
Also, some variables do not contain any useful values yet. They are mere placeholders for a feature whose implementation is anticipated. For instance, the variable known as "bunch-by-bunch luminosity" is always 0. Section 2.5 describes how the (integrated) luminosity for TAG events can be calculated.

From the names of the variables it can be seen that the entries do not contain data structures of any kind, only integer and float variables. Objects (like muons) are represented as a series of numbered variables like `LooseMuonPt1` through `LooseMuonPt4`. For events that contain less than 4 muons, the unneeded variables will contain all zero values.

"Loose" particles

In section 1.1.5 on page 11 the expression "loose" was defined for electrons. In general, loose is a term for collections of particles that have undergone some identification filters with emphasis more on high selection efficiency rather than background or fake rejection. This is well suited for TAGs, which try to provide all interesting particles for a large number of analyses.

The actual definition of what loose particles are in this context is quite specific to TAGs: **Loose electrons** are electrons that have been classified as loose



(a) Loose muons and loose electrons (FDR2 run 52283)

(b) Jets (FDR2 run 52290). The dotted histogram uses information that is not available in TAGs.

Figure 1.6: The number of objects per event in FDR2

according to section 1.1.5 **and** have $p_{\mathrm{T}} > 7.0$ GeV [26]. **Loose muons** are muons with $p_{\mathrm{T}} > 6.0$ GeV that have been classified by the STACO algorithm as either 'standalone', 'combined' or 'low p_{T} reconstructed' [27]. The `LooseMuonTightness` variables indicate the type of classification.

Number of particles

The variables `NLooseElectron, NLooseMuon, NJet` etc. contain the original number of those particles, i. e. that many jets were in the AOD that was used to create the TAG. However there are only six jets in the TAG file, namely the jets with the largest p_{T}. This is sufficient for nearly all possible analyses. Still, the `NJets` variable may be greater than 6, as can be seen in figure 1.6b.

The same is true for electrons and muons, though after the "loose" selections (including the p_{T} cuts that the TAG creation uses), hardly any dataset contains events with more than four objects per event, like in figure 1.6a.

Charge

For charged particles, the charge is encoded as the sign of the `Pt` variable. Ordinarily, p_{T} is always non-negative by its definition, however in TAG files, all `LooseElectronPt` values are negative for electrons, while they are positive for positrons.

Triggers

Each of the 32 `L2PassedTrigMask` and 32 `EFPassedTrigMask` variables contains 32 bits, which are numbered continuously. The names of the triggers can be turned into numbers by consulting the *Chain Tag Map* [28].

For instance, `EF_2mu6` is an EF trigger that looks for events with 2 muons that have $p_{\mathrm{T}} > 6$ GeV. According to the Chain Tag Map, it has the number 135, which is $4 \times 32 + 7$. Events where this trigger has fired will have the 8th bit in `EFPassedTrigMask4` set.

Chapter 2

Cross-section measurements

2.1 Cross-section

The *cross-section* σ is used as a measure for the probability of an interaction between particles. It is the area of a mental target; for the case of a contact interaction of point-like particles hitting extended particles, the cross-section is given by the area of the target particle.

The unit for σ is *barn,* where $1\mathrm{b} = 10^{-24}\ \mathrm{cm}^2$. Measurable cross-sections at LHC range from 1 fb to 100 µb.

The *luminosity* is used as a measure for the intensity of the collider. Its units are $[\mathrm{cm}^{-2}\mathrm{s}^{-1}]$ or $[\mathrm{b}^{-1}\mathrm{s}^{-1}]$. The *integrated luminosity* $\int\!\mathscr{L}\,\mathrm{d}t$ is used to describe how significant an amount of data is. It describes the number of expected events per cross-section of a given process.

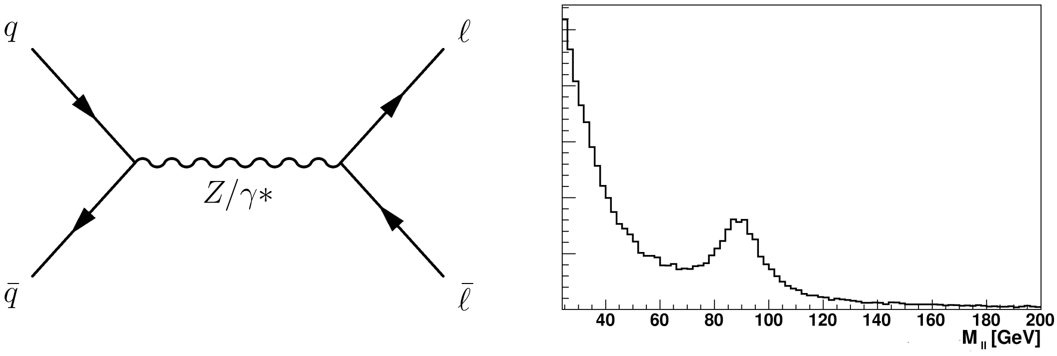
In order to calculate a cross-section, the following equation may be used:

$$\sigma = \frac{N}{\int\!\mathscr{L}\,\mathrm{d}t} \tag{2.1}$$

where N is the number of events that occurred for the process whose cross-section is calculated. To obtain N, one counts events that match certain selection criteria. The number of counted events are denoted N_{sel}, which is not equal to N due to some correction factors:

$$N = \frac{N_{\mathrm{sel}} - N_{\mathrm{sel}}^{\mathrm{BG}}}{\varepsilon} \tag{2.2}$$

Here, N_{sel} is the number of events that are found by the selection. In general, they also include a portion of events that are not genuine instances of the investigated process but have been selected due to similar detector signatures of other processes. $N_{\mathrm{sel}}^{\mathrm{BG}}$ is the number of these *background* events.



(a) Feynman diagram

(b) Dilepton mass [19]

Figure 2.1: The Drell-Yan process

Also, there will be a fraction $1 - \varepsilon$ of interactions missing in the selection. Usually one denotes ε as the product of the trigger efficiency $\varepsilon_{\text{trig}}$ and the selection efficiency ε_{sel}.

ε_{sel} may be studied using Monte Carlo experiments. In those, a particle process and its detection is simulated and ε_{sel} and $\varepsilon_{\text{trig}}$ can be determined.

In this chapter, decays of the Z boson have been used to study how cross-section measurements can be done with TAG files. The selection criteria that are used to obtain N_{sel} will be introduced in section 2.3. Section 2.5 will explain how $\int \mathscr{L}\,\mathrm{d}t$ was obtained.

2.2 Drell-Yan process

The production of lepton pairs in hadron-hadron scattering via the creation of a Z boson or a virtual photon from a quark-antiquark annihilation is known as *Drell-Yan process* (figure 2.1a). The leptons that come from a Z boson decay can be recognized by their mass peak at $\sim$ 91 GeV (figure 2.1b), which is the Z mass.

This interaction was first suggested in 1970 and is very well understood meanwhile. It is widely used at ATLAS as a calibration process. In this chapter, the Drell-Yan Z decay into two muons will be used to demonstrate how cross-sections may be determined using TAG selections.

2.3 Cuts

These cuts were used for the event selection. See also [16].

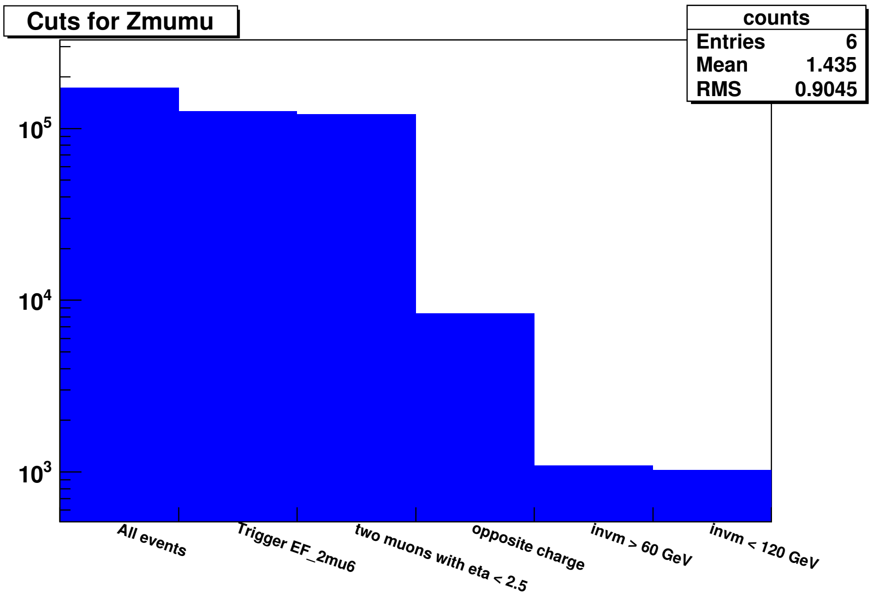


Figure 2.2: Cut flow for Z selection

- Trigger: A two-muon 6 GeV trigger was used. It will trigger whenever it detects an event with 2 muons that have $p_{\mathrm{T}} > 6$ GeV. In principle, 20 GeV triggers are available as well and could be used for Z decays. However, for the FDR2 data, luminosity information was only available for 6 GeV triggers.

- Only events with at least **two loose muons** are looked at. For events with more than two muons, every combination of two muons is considered for the following cuts. The event is counted if at least one muon pair exists that fulfills the criteria.

- Oppositely charged muons are of course required.

- $|\eta| < 2.5$ for both muons since this is the range of the Inner Detector. Although the Muon System covers the range to 2.7, muons that are beyond the Inner Detector have not been used.

- 60 GeV $< M_{12} <$ 120 GeV (where M_{12} is the invariant mass of the two muons) is a requirement that was used to have a 30 GeV acceptance window around the Z mass.

2.4 Determining the quantities from the TAG variables

All variables used in the cuts may be determined from TAG files. The charge, p_T, η can be retrieved directly. The invariant mass is calculated from the TAG variables as outlined in table 2.1. Because these computations are often needed in particle physics, the ROOT framework provides functionality[1] to obtain these results when p_T, η and φ are given.

2.5 Luminosity calculation

2.5.1 Luminosity blocks

To facilitate the handling of large amounts of data, it is divided into so-called *luminosity blocks* (LB). These are intervals of the order of 1 minute length [6, 7]. For the FDR2 data, the integrated luminosity of one block is 6 nb^{-1}, while as of April 2010, the actual luminosity blocks are of the order of 0.01 μb^{-1}.

When determining a cross-section, the luminosity must be obtained for all LBs that were searched for events. This includes those blocks where no events have been found, illustrated in figure 2.3 for $\mu\mu$ events. If a preselection is made, one must store the luminosities of all LBs that were present in the original data set. For this reason, skimmed[2] TAG files produced by the ELSSI frontend to the TAG database also contain information about LBs that are not present in the skim.

2.5.2 Conditions Database

The Conditions Database (CondDB), which is part of the ATLAS database project, contains the integrated luminosity of the luminosity blocks in ATLAS runs. Infor-

ϑ	$2\arctan\exp(-\eta)$	(from 1.2 on page 6)
p	$\frac{p_\mathrm{T}}{\sin\vartheta}$	
α_{12}	$\arccos\left[\sin\vartheta_1\sin\vartheta_2\cos(\varphi_2-\varphi_1)+\cos\vartheta_1\cos\vartheta_2\right]$	
M_{12}	$\sqrt{2p_1p_2\,(1-\cos\alpha_{12})}$	

Table 2.1: Calculation of other parameters for muons and electrons: Only $p_\mathrm{T}, \eta, \varphi$ are given. The expression for α_{12}, the angle between two particles, may be obtained by using the scalar product.

[1] In particular, the `SetPtEtaPhi` method in the `TLorentzVector` class

[2] *Skimming:* Removing events from a collection based on selection criteria

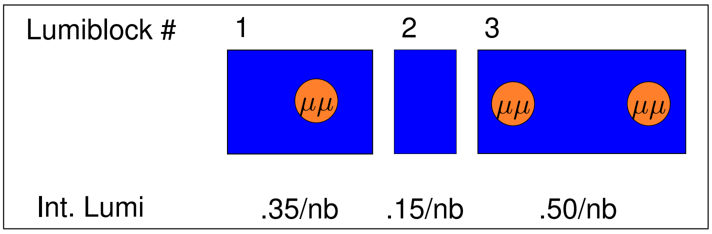


Figure 2.3: Luminosity blocks (example)

mation is available for each trigger that was defined during the run. The CondDB may also be used to get (simulated) integrated luminosities for the FDR data. For obtaining the information, a software tool called `LumiCalc` may be used; for a given trigger it retrieves the integrated luminosity by run number and LB number.

2.6 Datasets used

For this analysis, data from the FDR project have been chosen. Collision data were not available by the start of this diploma thesis, and as of now (May 2010), the number of observed Z particles is of the order of 1, therefore simulated data had to be used and the FDR sample was an obvious choice since it resembles realistic data.

2.7 Cross-section of the $Z \to \mu^-\mu^+$ decay

64585 events from the muon stream of FDR2 run 52283 were used for this calculation. This amounts to 111 luminosity blocks. From the conditions database, the total integrated luminosity $\int \mathscr{L}\,\mathrm{d}t = 331.2\ \mathrm{nb}^{-1}$ was obtained. Using the selection criteria described in section 2.3, 553 selected events have been found.

Without further analysis of efficiencies and background rejection, this corresponds to a cross-section of

$$\sigma = \frac{N}{\int \mathscr{L}\,\mathrm{d}t} \approx 1.67 \pm 0.07\ \mathrm{nb} \tag{2.3}$$

2.8 Summary

The $Z \to \mu^-\mu^+$ process at ATLAS has a cross-section of $BR \times \sigma \approx 1.497$ nb [6]. While this value is close to the result presented above, it is not within the error margin.

Number of events	64585	
Number of found $Z \to \mu^-\mu^+$ events	553	
Integrated luminosity	331.2	nb^{-1}
Size of TAG files	30	MB
Size of AOD files	8	GB

Table 2.2: Amount of data used for this analysis

Due to the large cross-section of the Drell-Yan process for low masses (cf figure 2.1b), the invariant mass cut at 60 GeV has included a non-negligible portion of background events, which explains that the determined value was higher than the cross-section of $Z \to \mu^-\mu^+$.

There are also several efficiency considerations to be done. For instance, the excluded η range needs to be accounted for. A more precise discussion of the event selection for $Z \to \mu^-\mu^+$ can be found in [16].

Another point to notice here is that the result was obtained using FDR data and the weighting used for this mixture has not been publicized in detail.

The TAG files that were used amount to $\sim$ 30 MB. This size of data does not need setting up a distributed analysis. The data has been downloaded and then processed on a single machine. The actual analysis only took $\sim$ 1 minute (corresponding to a frequency of $\sim$ 1 kHz), a time that could still be shortened by optimizing the analysis code. Using the AODs (8 GB) can take a time in the order of an hour, also depending on the load of the Grid (if it is used). Even though the Grid provides fast computations using parallelization, a time of one minute is still well below the overhead time that one usually has to wait while a computing job is queued (this can be up to an hour).

As mentioned earlier, TAG files do not provide the full repertoire of reconstruction quantities. Therefore, for a precise calculation of a cross-section, one has to resort to other data formats. Still, this study has shown that TAG files or the TAG database can be used as a powerful tool to get quick results for a cross-section measurement.

There are a number of scenarios that depend on quick results. While devising an analysis, one often needs to make decisions like choosing a good value for the invariant mass cut. Using TAGs, one can rapidly create a histogram like the one presented in figure 2.1b and find out the number of events above and below a certain threshold. This can be repeated an arbitrary number of times, so one can use the results of one computation as a basis for the following computations, e. g. while choosing a good combination of cuts for $M_{\ell\ell}$ and η.

The fact that TAG can be stored locally also simplifies displaying individual events. For instance, if at some point in the analysis it is discovered that a number of events produce unanticipated results, the short but comprehensive event summary provided by TAG variables is often instrumental to find the explanation.

Scaling to larger datasets

At the LHC design luminosity $\mathscr{L} = 10^{33}\ \mathrm{cm}^{-2}\mathrm{s}^{-1}$, the data recorded during the course of a year can amount to $\int\!\mathscr{L}\,\mathrm{d}t = 10\ \mathrm{fb}^{-1}$. Assuming a muon stream composition like in FDR2, this corresponds to an event count of

$$N = \frac{10\mathrm{fb}^{-1}}{331\mathrm{nb}^{-1}} \times 64585 \approx 2 \times 10^{9}. \tag{2.4}$$

The AOD size for collision data is $\sim$ 100 kB, thus 200 TB will be assumed as the size of the muon stream AODs for one year. Using a Grid site with 1000 CPUs, the data could be processed in a matter of days, assuming realistic analysis event rates of 20 Hz per CPU. This of course requires low Grid utilization by other jobs, so less performance is to be expected if several physicists process large amounts of data simultaneously. Requiring the same analysis time, TAG data can be processed by only 20 CPUs (assuming a rate of 1000 Hz like presented above). The required data size is only 2 TB. It is remarkable that this is a size that fits on a contemporary hard disk. While full analysis of these TAGs can take some time, these files could be used to predict results based on portions of the data, or to display properties of individual events as described above.

More complex analyses than the one presented in this chapter can also greatly benefit from TAG-based preselections, which will be the topic of the next chapter.

Chapter 3

TAGs for a complex analysis

3.1 Search for the Higgs particle

At LHC, Higgs particles can be produced in four ways: Gluon-gluon fusion, vector boson fusion (fusion of two W or Z bosons), associated production with a W or Z boson, and associated production with a $t\bar{t}$ pair.

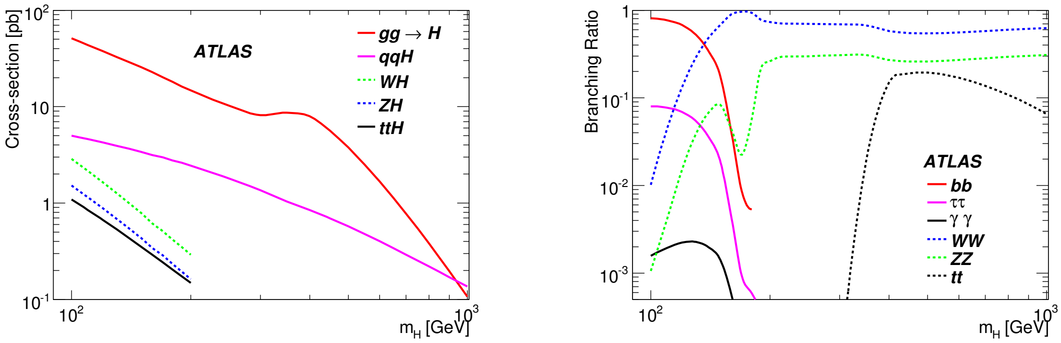
The cross-sections for the Higgs production channels depend on both the center-of-mass energy and the unknown Higgs mass m_{H}. Figure 3.1a shows the cross-sections as functions of m_{H} at $\sqrt{s} = 14$ TeV.

Figure 3.1b shows the dependence of the branching ratios of some Higgs decay modes on the Higgs mass. For large Higgs masses, one expects to observe more decays into heavier particles (Z, W, t), while for smaller Higgs masses, lighter particles are produced. For $m_{\mathrm{H}} \gtrsim 140$ GeV, the decay into two W bosons is predominant, which makes it an important channel for the Higgs search at LHC.

3.2 Higgs decay into WW

The Higgs decay channel $H \rightarrow W^+W^- \rightarrow \ell\nu\bar{\ell}\bar{\nu}$ has been studied by the *Higgs Search Group 3* (HSG3). The decay channel as shown in figure 3.2 produces two muons and two neutrinos. Similar decay chains produce two electrons (e^+e^-) or one electron and one muon ($e^+\mu^-$ or $e^-\mu^+$).

Several background processes can produce similar detector signatures. Figure 3.3 illustrates the importance of background processes that produce two leptons.



(a) The cross-section for the five Higgs production channels

(b) Branching ratio for some Higgs decay channels

Figure 3.1: Cross-section and branching ratios as a function of the Higgs mass, at 14 TeV [6, p. 1204].

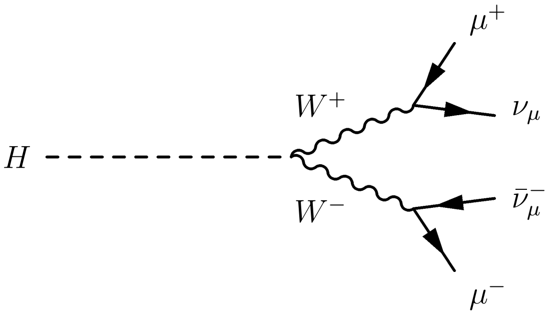


Figure 3.2: Higgs decay into two neutrinos and two leptons (here: muons)

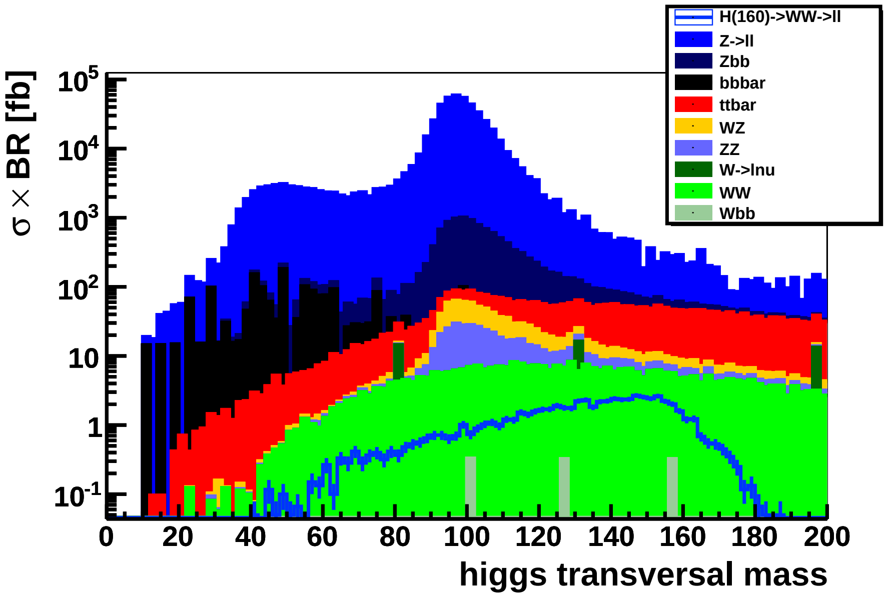


Figure 3.3: This histogram shows the distribution of the Higgs transverse mass—as defined in section 3.7—after cut Ia (the lepton selection). The contributions of the various background processes are stacked on top of each other, while the blue line shows the actual signal process. This diagram was created using Monte Carlo simulations of the processes shown [5].

3.3 Cuts

A series of cuts has been proposed by HSG3 to filter out the background using various criteria that show the differences between leptons and jets from $H \to WW$ and background processes [29, 31].

The full listing of the cuts is included in appendix B. Basically, the cuts can be divided into these parts:

Lepton and jet selection: For the reconstructed electrons, muons and jets, commonly used variables (such as p_{T}, $E_{T,\mathrm{cone}}$) are used to decide which objects are to be included in the further cuts.

For leptons, isolation variables are used to determine whether the leptons are produced close to other particles, which is not assumed for the signal process.

Overlap removals serve to recognize multiple reconstructions provoked by the same particle [6].

Common cuts (Higgs candidate preselection): The first cut (Ia in appendix B) requires that exactly two leptons have been found. Then some cuts are made on properties of the two leptons. The invariant mass is used in cut Ic to remove muons in the mass range of the Z particle, which are commonly found in backgrounds like $Z \to \mu^-\mu^+$. Cut Id uses the missing energy to remove backgrounds with no or less energetic neutrinos.

The missing transverse energy is obtained from the corrected version of the `MET_RefFinal` variable (cf section 1.1.5).

Cuts depending on the number of jets: The final selection is obtained by a series of cuts that depend on the number of jets. Cuts for a 0-, 1- and 2-jet bin have been defined.

3.4 Cross-sections and event processing

The $H \to W^+W^- \to \ell\nu\bar{\ell}\bar{\nu}$ process that was used for this analysis has a cross-section of 468 fb (cf section 3.6.2). This is much smaller than the cross-sections of background processes (200 pb for the $t\bar{t}$ background, 1.5 nb for the Z background).

The sample cross-section (number of events in the sample per unit $\int\mathscr{L}\,\mathrm{d}t$) for the FDR2 muon stream is about 500 nb. Searching such a stream for $H \to W^+W^- \to \ell\nu\bar{\ell}\bar{\nu}$ candidate events will only find one background event in 500 events and one signal event in 10^6 events.

In most cases the few variables in the TAG format are not sufficient for a sophisticated analysis. Usually such analyses need much more information, requiring AODs (or even ESDs).

To get statistically significant results, often a large amount of data has to be processed. Due to the rather large size of AOD events, this can take a long time and requires access to vast storage arrays. Often this is not possible on a single local machine (because it would take too long and would require downloading too much data), instead the task has to be split and distributed on the Grid.

3.5 Preselections

Instead of processing each AOD event it would be much more efficient to *preselect* the interesting events, e. g. for an analysis of the $\mu^-\mu^+$ channel, only those events with two highly isolated muons and 0–2 tag jets are needed.

3.5.1 TAG preselections

If TAGs are used to find events with these properties, one can retrieve the AOD variables only for the interesting events and do the final analysis with those.

When a complex analysis like $H \to W^+W^- \to \ell\nu\bar{\ell}\bar{\nu}$ is devised, careful studies lead to the decision which of the many variables are best suited for a given analysis.

An important constraint for the design of the TAG files was its size. For this reason, they cannot provide as many choices as AODs. Because TAG files are intended to be used for all kinds of analyses, a design decision like this can never be perfect for all applications. The aim of this chapter is to assess how to make best use of the few variables, for it is not straightforward to map a complex AOD based selection on the information available in TAGs.

Nevertheless, one should design preselections in such a way that the final analysis has the same results as without this preselection.

3.5.2 Designing inclusive preselections

An obvious way to ensure same results is to design the preselection *inclusive,* i. e. to avoid the case that the full event selection will find events that the preselection does not.

Getting too many events is not so much of a problem. After doing the preselection, it can be expected that the number of events to consider was narrowed substantially. Therefore one can easily collect the full AOD information for the remaining events and do the complicated cuts. If the preselection is inclusive, one

	rejected by AOD cut	survives AOD cut
rejected by TAG cut	a_{00}	$a_{10} \overset{!}{=} 0$
survives TAG cut	$a_{01} \overset{!}{=} \min$	a_{11}

Table 3.1: Inclusiveness

will end up with the same events as if one had done the complicated cuts without the preselection, which would require more resources and takes longer.

Table 3.1 represents this in a schematic way. The matrix contains four values a_{ji} that stand for the number of events that fail or pass the exact (AOD) or inexact (TAG) cuts. Demanding that the TAG preselection be inclusive, the upper right number must be zero.

The number a_{01} stands for the number of events that get sorted out during the postprocessing. The trivial preselection (which selects all events) would have this number equal to the total number of events. However, for a sophisticated preselection, it is desired that this number be as small as possible to minimize the number of events for post-processing.

a_{11} is the number of events that are used for the final analysis.

3.6 Comparison of AOD and TAG

3.6.1 TAG cut categories

There are different categories of variables when comparing TAG and AOD:

Direct correspondence: The most obvious candidate cuts are those using variables with exact correspondences in the TAG files. One such variable is transverse momentum of electrons.

So if a cut is done on two leptons with

$$p_{\mathrm{T}} > 15\ \mathrm{GeV} \tag{3.1}$$

it is a trivial task to design a TAG preselection, because the p_{T} variable for electrons and muons can be used directly.

Close correspondence: An example for such a cut would be

$$E_{T,\mathrm{cone},50} < 10\ \mathrm{GeV} \tag{3.2}$$

for a muon that was reconstructed using the STACO family of algorithms.

This variable is not contained in the TAG files; they only contain the respective variable for 0.4-cones.

$E_{T,\mathrm{cone},50}$ cannot be computed from $E_{T,\mathrm{cone},40}$ or the other way round if only TAG information is available. To do so, one would need to know about the energy of all calorimeter cells in the area between the 50- and the 40-cone, but the calorimeter cells are not contained in the TAG files. While it is possible to guess about the calorimeter isolation using e. g. the jet information that is contained in the TAG files, this can give no exact answers.

Due to the fact that the smaller cones are contained in the larger cones, the inequality

$$E_{T,\mathrm{cone},40} < E_{T,\mathrm{cone},50} \tag{3.3}$$

is always true. Therefore, if the cut is adjusted:

$$E_{T,\mathrm{cone},40} < 10 \text{ GeV} \tag{3.4}$$

it will yield all the events from the previous cut, and some more, because 3.2 implies 3.4 due to the transitivity of "<". Therefore 3.4 is a classic example for an inclusive preselection.

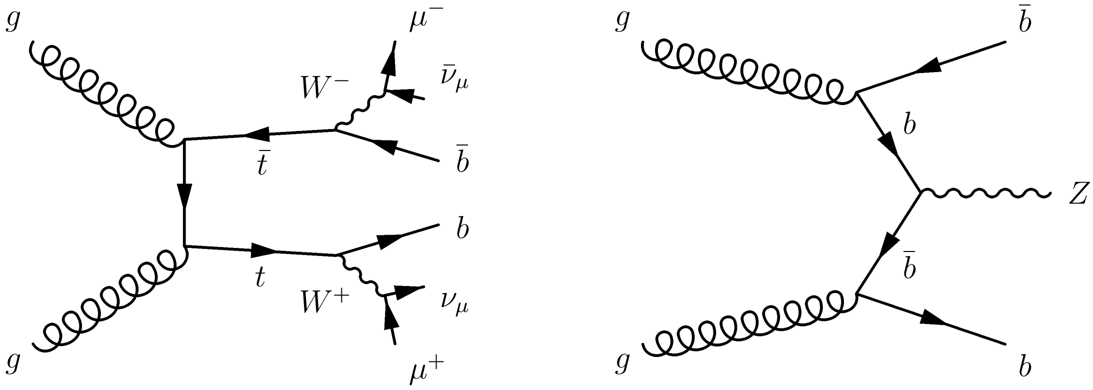
No correspondence: Some cuts cannot be adjusted in such a simple way. The worst case is a cut on a variable that is missing in the TAG files altogether. But there are other problematic situations: For instance, if the analysis requires a different jet definition than the one used for TAG production, some cuts on jet properties may produce results that can hardly be reconciled with the original analysis if missing events are not tolerated.

3.6.2 Signal and background samples

In order to study the cuts in detail, Monte Carlo samples for the signal process and some backgrounds were used. All samples have been retrieved both as AOD and TAG files. Therefore it was possible to directly compare TAG and AOD cuts and compute the numbers in table 3.1.

Selection of signal samples

For the $H \to W^+W^- \to \ell\nu\bar{\ell}\bar{\nu}$ process, the Monte Carlo run 106533 was used. It assumes a Higgs mass of 170 GeV and uses Higgs bosons that are produced by gluon-gluon fusion at a center-of-mass energy of 10 TeV. The cross-section is 467.87 fb, the integrated luminosity is 23.3 fb^{-1} [32].



(a) QCD $t\bar{t}$ production

(b) QCD $Zb\bar{b}$ production

Figure 3.4: Example Feynman diagrams illustrating the background processes that have been examined

Selection of background samples

The focus of this study is on methods that use very few available physical quantities as a preselection for a complex analysis. Particularly of interest are problems that involve variables with a close correspondence in TAG files, since this gives insight into the nature of these quantities and their suitability for physics analyses. For $H \to W^+W^- \to \ell\nu\bar{\ell}\bar{\nu}$, close correspondences are to be expected for $E_{T,\mathrm{cone}}$ and $\not\!\!E_{\mathrm{T}}$. Therefore, specific backgrounds have been selected which are reduced by cuts on these quantities. The chosen backgrounds are also very dominant (cf figure 3.3), i. e. a preselection that reduces these backgrounds is very valuable.

- The $t\bar{t}$ decay as displayed in figure 3.4a is characterized by much hadronic activity, thus the muons can be expected to be less isolated than the ones that are searched for. This makes this background a means to study the muon isolation.

  To study this background, the Monte Carlo run 105200 has been used. It has been generated using MC@NLO with a sample cross-section of 202.86 pb. The integrated luminosity is 492 pb^{-1}.

- The $Zb\bar{b}$ production as shown in figure 3.4b can be separated from the $H \to W^+W^- \to \ell\nu\bar{\ell}\bar{\nu}$ process by the distinction that it does not produce high energetic neutrinos. Therefore this background can be used to study the missing transverse energy, which is usually used as an indicator for neutrinos.

For $Zb\bar{b}$, the Monte Carlo run 109300 has been used. It was generated using Alpgen and has a sample cross-section of 12.22 pb. The integrated luminosity is 815 pb^{-1}.

- The Drell-Yan decay, which was already used in chapter 2, is a background to the signal process as well. It has a very large cross-section, however its leptons have masses (cf figure 2.1b) that are much smaller than the Higgs mass used in the signal sample, and they always have the same flavor.

  Two Monte Carlo samples have been used for this background: For low mass muons (10 GeV $< M_{\ell\ell} <$ 60 GeV), Pythia run 106051 with a sample cross-section of 684 pb and $\int \mathscr{L}\,\mathrm{d}t = 14.51\ \mathrm{pb}^{-1}$; for higher energetic muons (dominated by $Z \to \mu^- \mu^+$), Pythia run 106000 with 1098 pb and $\int \mathscr{L}\,\mathrm{d}t = 9.106\ \mathrm{pb}^{-1}$ has been used.

3.6.3 Inclusive TAG cuts

In this section, the cuts that were introduced in 3.3 will be analysed regarding how a preselection can be defined using TAG variables, while specifically trying to keep the preselection inclusive.

Overview

The following variables may be directly obtained from TAGs:

- Muons: p_T, η, the "combined" flag
- Electrons: p_T, η, tightness
- ΔR between leptons
- Invariant mass of a lepton system

Note that the tightness and p_T cuts that are used during TAG production (cf section 1.3) are also implied by the cuts in this analysis, so no needed objects are missing.

The following variables are used in the cuts but have no correspondence in TAGs:

- Muons:
  - χ^2 (from the combined muon matching in the STACO algorithm)
  - Impact parameter significance

    - Track isolation (not implemented in the version that produced the sample files)

- Electrons:
    - Isolation variables
    - Impact parameter significance
    - Reconstruction algorithm

All other variables are more complex to handle and will be discussed in the following subsections.

Interdependencies of cut variables

The cuts have quite a complex structure. There are several dependencies between the selection criteria for muons, electrons and jets. In particular, because of the overlap removal,

- the electron selection depends on the muon selection because of the overlap removal with muons,
- the jet selection depends on both the muon and electron selection (cuts j3 and j4),
- the electron selection is influenced by the jet selection (cut e4 requires medium electrons for the 2-jet channel).

As a result, one cannot easily omit any of the selection criteria without influencing the other cuts. Even if only the muon channel is to be analysed, the electron selection criteria are still influential for the jet selection.

Requirement of exactly 2 leptons

The first preselection cut (Ia) requires that exactly 2 leptons be present.

This complicates the implementation of the lepton selection. Unless it can be realized without uncertainties, it is possible that more leptons are tagged than in a fully-implemented selection. In particular, there might be events that have two leptons and pass all of the genuine cuts even though they have more than two leptons according to the cuts that can be implemented with TAG variables.

Therefore, for an inclusive preselection, cut Ia must be adjusted to allow *at least* two leptons. This raises questions for the later cuts that depend on the assumption that only two leptons are considered. The invariant mass used by some cuts is a property of the two-lepton system. A possible preselection for those cuts

MC Sample	All events	Exactly two tag leptons	More than two
$H \to W^+W^- \to \ell\nu\bar{\ell}\bar{\nu}$	4677	1859	0
$t\bar{t}$	2026772	89097	183
$Zb\bar{b}$	121536	30158	24
Drell-Yan (low masses)	145479	8077	0
$Z \to \mu^-\mu^+$	10982000	4515798	0

Table 3.2: Number of events with more than 2 leptons (scaled to $\int\mathscr{L}\,\mathrm{d}t = 10\ \mathrm{fb}^{-1}$)

would be to to accept all events where the invariant mass of any lepton-antilepton pair meets the cut requirement.

From table 3.2 it can be seen that only a very small fraction of the events actually contain more than two leptons. Therefore the adjustment of Ia to include these events does not have a great influence on the preselection.

In principle, by changing the preselection to allow more muons, the number of electrons might shrink because of the overlap removal. However, figure 3.5 shows that overlap removal does not have a notable influence on the selection for the samples at hand. Therefore the overlap removal cuts will not be used for the TAG preselection.

As will be seen later, the jet tagging cannot be implemented in this preselection very well. This indirectly influences the electron selection: Medium electrons should be used for the 2-jet bin, tight electrons otherwise. To avoid losing events, medium electrons need to be used in all cases. This has a slight influence on the signal sample, which is also shown in 3.5. Although $Zb\bar{b}$ shows a difference too, after the invariant mass cut Id, the results are the same with all medium electrons.

Lepton isolation

For electrons (cut e8) and muons (cut m7) it is required that

$$\frac{E_{T,\mathrm{cone,30}}}{p_\mathrm{T}} < 0.2 \quad \text{and} \quad E_{T,\mathrm{cone,30}} < 10\ \mathrm{GeV}. \tag{3.5}$$

TAG files include the $E_{T,\mathrm{cone,40}}$ variable for muons, while they do not provide isolation information of any kind for electrons.

By inequality 3.3 on page 33 it was suggested that $E_{T,\mathrm{cone,50}}$ may be replaced by $E_{T,\mathrm{cone,40}}$ without losing particles. However, here the $E_{T,\mathrm{cone,30}}$ variable is needed, which cannot be related to $E_{T,\mathrm{cone,40}}$ in a similar fashion.

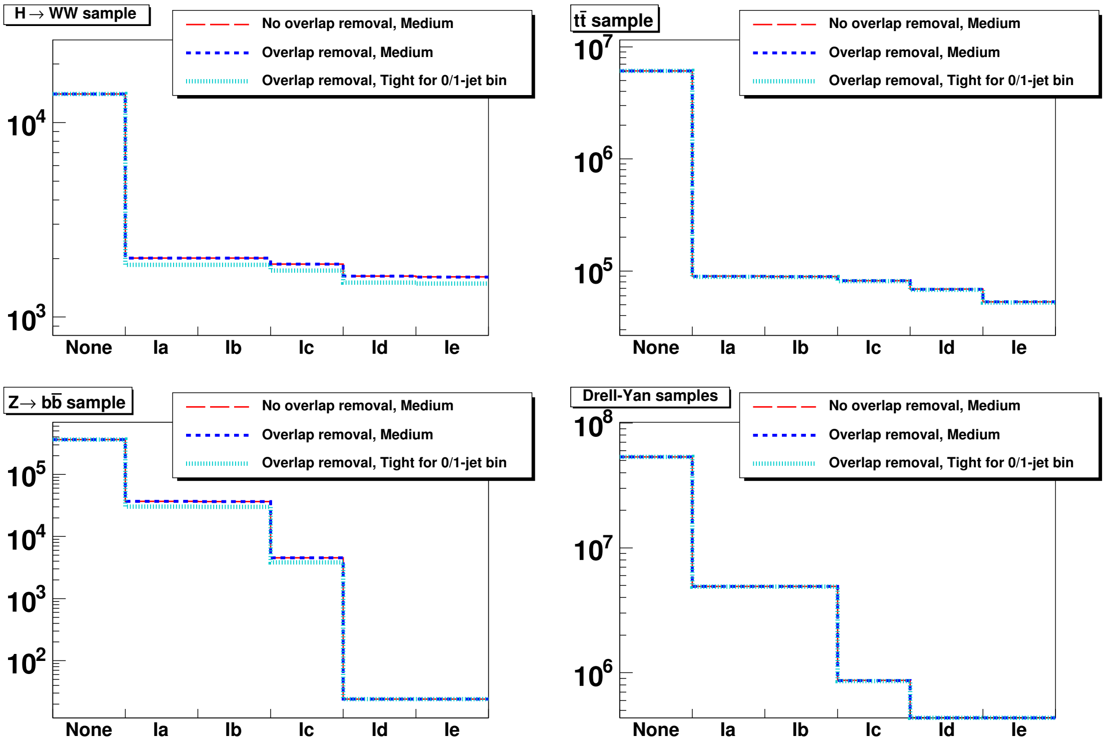


Figure 3.5: Influence of the overlap removal and the readjustment of the electron tightness. The x axis shows the Higgs preselection cut flow, labeled with the names of the cuts as explained in appendix B. Histograms are scaled to $\int \mathscr{L}\,\mathrm{d}t = 10\ \mathrm{fb}^{-1}$.

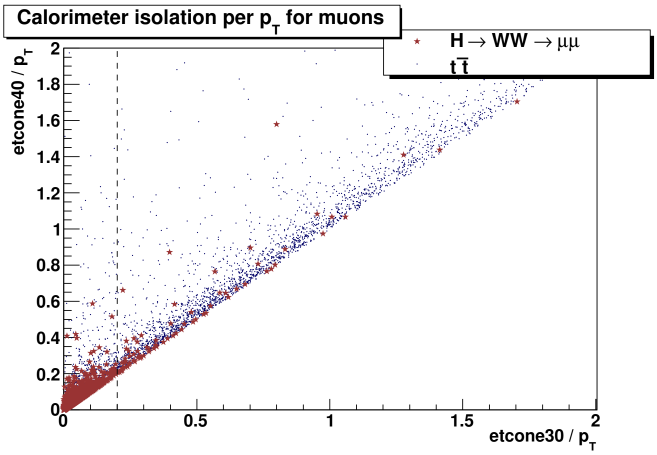


Figure 3.6: Correlation between $\frac{E_{T,\mathrm{cone},30}}{p_\mathrm{T}}$ and $\frac{E_{T,\mathrm{cone},40}}{p_\mathrm{T}}$ for muons. For this plot, events with two combined muons with $p_\mathrm{T} > 15$ GeV have been selected.

Figure 3.6 displays the $E_{T,\mathrm{cone},30}$ versus the $E_{T,\mathrm{cone},40}$ variable. The larger part of the graph is concentrated near the identity line, yet there are many points that have much greater $E_{T,\mathrm{cone},40}$ than $E_{T,\mathrm{cone},30}$, indicating high calorimeter response at the edge of a $\Delta R = 0.4$ cone. Especially for the $t\bar{t}$ events it is impossible to contain the points on the left-hand side of the dashed line based on $E_{T,\mathrm{cone},40}$ cuts.

The result of omitting the lepton isolation cuts can be seen in figure 3.7. The histogram has been separated into the ee, $e\mu$ and $\mu\mu$ channels according to the flavors of the leptons in the pair that is makes it through the selection.

Especially the muon isolation is helpful for reducing the $t\bar{t}$ background in the $\mu\mu$ and $e\mu$ channels. The figure also shows that for the important $t\bar{t}$ background, the calorimeter isolation alone can be quite instrumental even if the track isolation is not used.

Other lepton selection cuts

The restrictions on χ^2 and the impact parameter significance as well as the reconstruction algorithm ("author") had to be omitted because TAGs do not provide any corresponding variables. Figures 3.8 and 3.9 show that the influence on the selection is much smaller than that of the isolation cuts.

Jet definition

The analysis is designed for jets reconstructed from topological clusters ($\Delta R = 0.4$). The TAG files at hand have been produced with a software version that used jets reconstructed from calorimeter towers ($\Delta R = 0.7$). Figure 3.10 shows

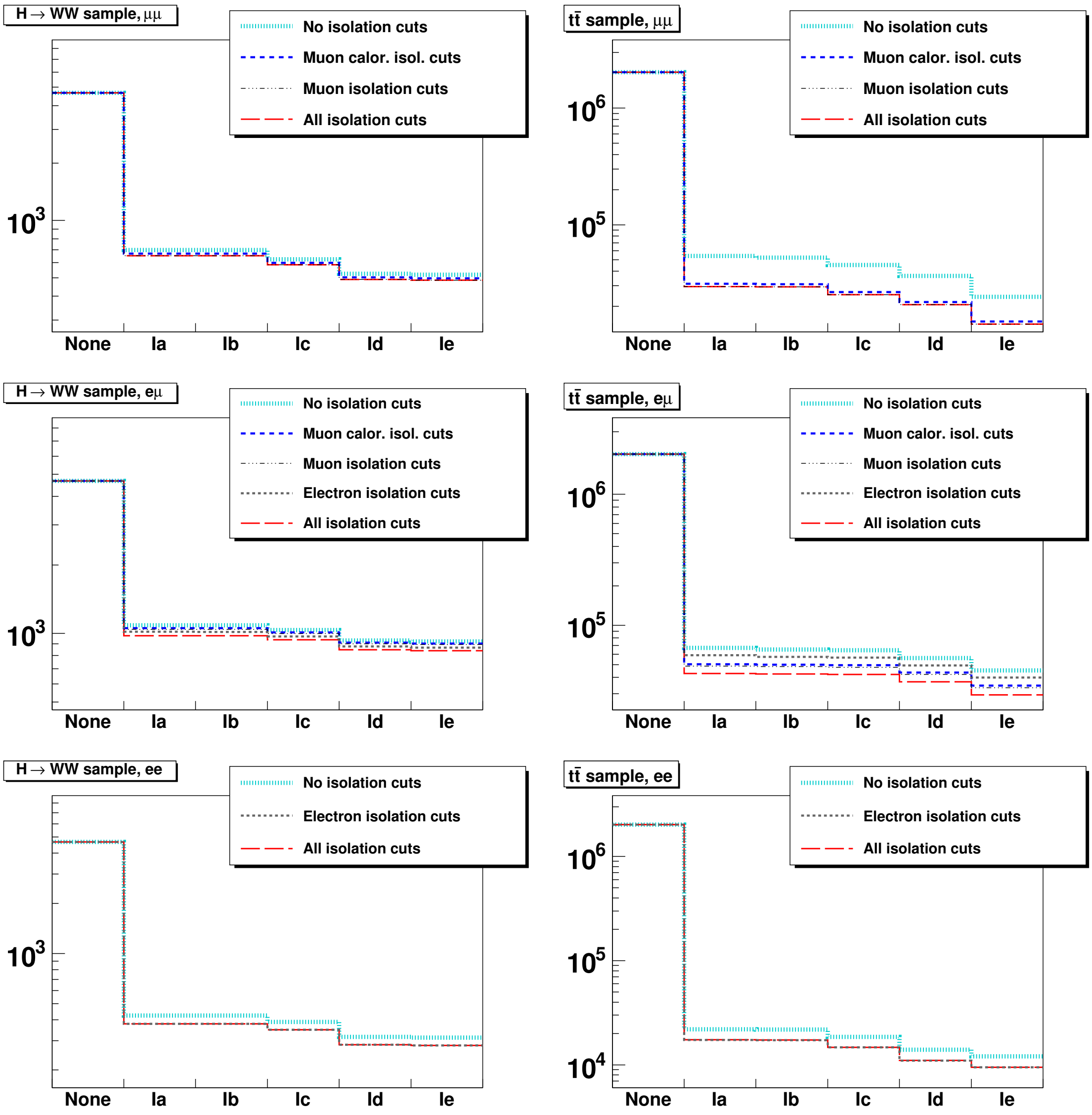


Figure 3.7: The influence of the track and calorimeter isolation on the event selection. The x axis shows the Higgs preselection cut flow, labeled with the names of the cuts as explained in appendix B. Histograms are scaled to $\int \mathscr{L}\, \mathrm{d}t = 10\ \mathrm{fb}^{-1}$.

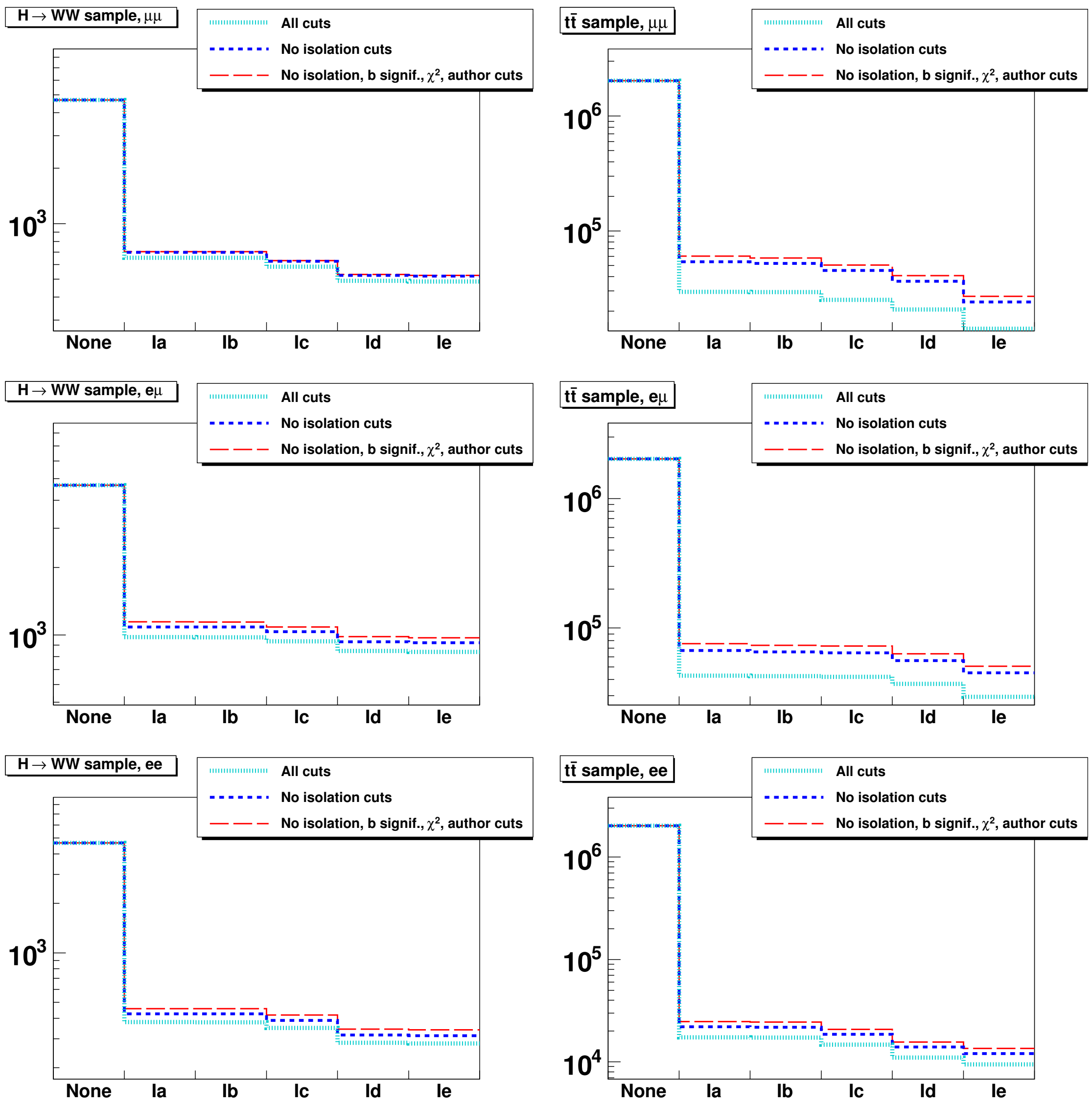


Figure 3.8: The influence of specific lepton selection cuts on the event selection ($H \to W^+W^- \to \ell\nu\bar{\ell}\bar{\nu}$ and $t\bar{t}$). The x axis shows the Higgs preselection cut flow, labeled with the names of the cuts as explained in appendix B. Histograms are scaled to $\int \mathscr{L}\,\mathrm{d}t = 10\ \mathrm{fb}^{-1}$.

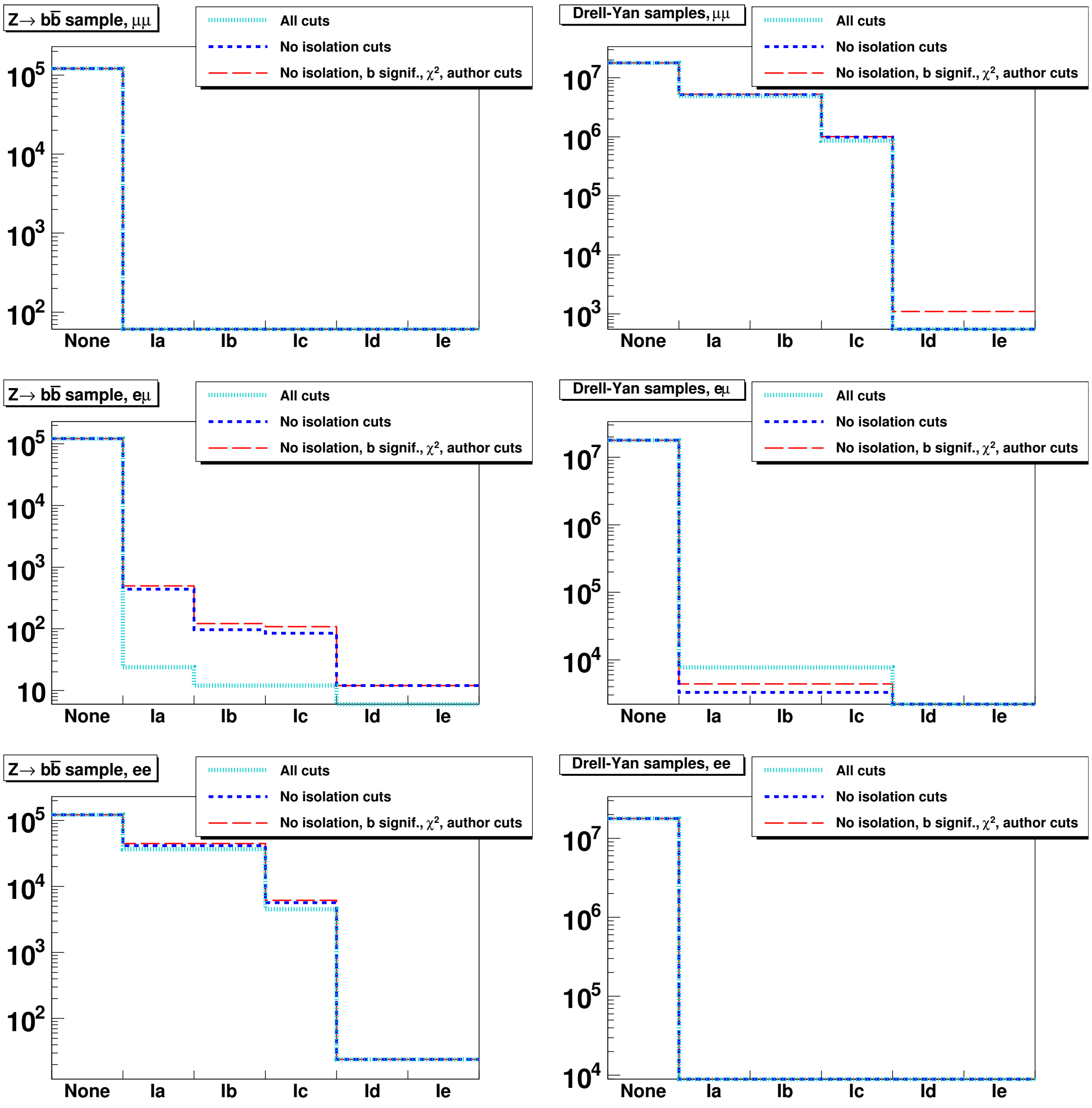


Figure 3.9: The influence of specific lepton selection cuts on the event selection ($Zb\bar{b}$ and Drell-Yan). The x axis shows the Higgs preselection cut flow, labeled with the names of the cuts as explained in appendix B. Histograms are scaled to $\int \mathscr{L}\,\mathrm{d}t = 10~\mathrm{fb}^{-1}$.

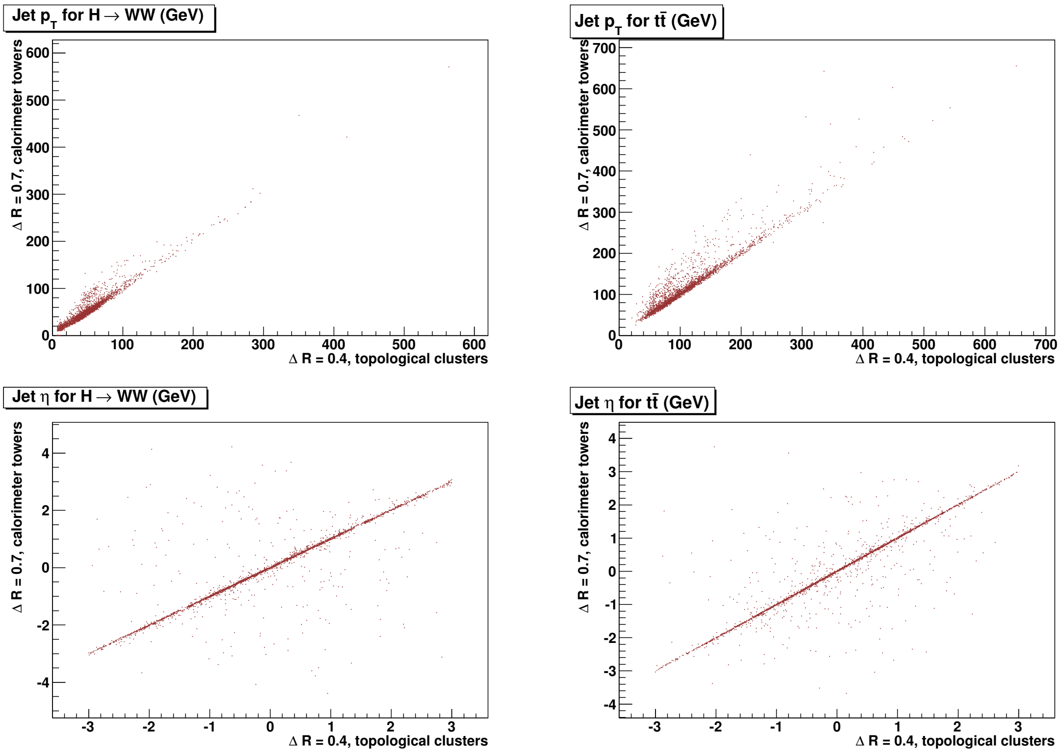


Figure 3.10: p_{T} and η of the highest-p_{T} jet after cut Ie. Variables for the jets contained in TAGs are plotted on the x axes while the y axes represent the corresponding variables from jets that are only in AODs. η values are only plotted for jets that have $p_{\mathrm{T}} > 15$ GeV according to both definitions.

the difference between the two jet collections for the p_{T} and η distributions. Note that for those points that are far from the identity line, the differing values need not be from the same jet, however it is not possible with TAGs to tell if that is the case.

The result is similar to the $E_{T,\mathrm{cone}}$ case in that it does not seem possible to create inclusive preselections for cuts based on jet properties.

Missing energy

Figure 3.11 illustrates how one can adjust the $\not\!\!E_{\mathrm{T}}$ cut to TAG variables. The original cut,

$$\not\!\!E_{\mathrm{T}} > 30 \text{ GeV} \qquad \text{(for same flavor)} \tag{3.6}$$

was designed for the corrected and refined $\not\!\!E_{\mathrm{T}}$ variable, `MET_RefFinal_corrected` (cf section 1.1.5 on page 13). This cut is represented by the vertical black line in

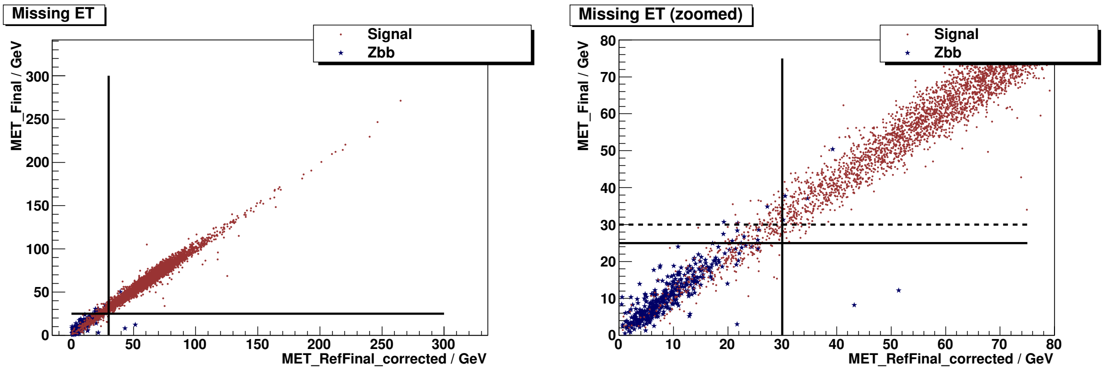


Figure 3.11: Missing E_{T} cut. Note that the $Zb\bar{b}$ background sample contained as many events as the signal sample, so there would be much more background events if both samples had the same integrated luminosity. This plot was made after cut Ic (i. e. at the point where the next step would be the $\not\!\!E_{\mathrm{T}}$ cut).

the figure. It can be seen that most of the $Zb\bar{b}$ background stays on the left-hand side of the line, while the right-hand side contains a significant portion of the signal events.

TAG provides the simple `MET_Final` variable as a measure for $\not\!\!E_{\mathrm{T}}$. The dashed black line shows how the same cut would look like with this variable. While the separation of the background is arguably not too bad either, the problem is that this cut would miss those events that are in the lower right square.

This can be solved by adjusting the limit from 30 to 25 GeV. For $e\mu$, the limit was adjusted from 40 to 34 GeV.

The additional events originating from this adjustment are very few compared to the large amount of background in the lower-left square. The $Zb\bar{b}$ sample contained two events after the AOD cut that evaded the adjusted TAG cut (lower-right square). However these stray events were filtered out by later AOD cuts, so the preselection is in fact inclusive.

The resulting preselection

In summary, an inclusive preselection can be realized with TAG variables as follows:

- Require at least two leptons with the following criteria:
  - For muons: Combined, $p_{\mathrm{T}} > 15$ GeV, $|\eta| < 2.5$
  - For electrons: Medium, $p_{\mathrm{T}} > 15$ GeV, $|\eta| < 2.47$ excluding $1.37 < |\eta| < 1.52$

- $E_{T,\text{cone}} > 34$ GeV ($e\mu$) or $E_{T,\text{cone}} > 25$ GeV (at least one ee or $\mu\mu$ pair)
- For all found pairs: $M_{\ell\ell} > 15$ GeV ($e\mu$) or $|M_{\ell\ell} - M_Z| > 10$ GeV (same flavor)

Figure 3.12 illustrates the preselection efficiency. The preselection described above can be seen in the figure as the dashed black line. For all samples, the preselection is quite close to the AOD cuts Ia–Ie. However the jet-based cuts, which could not be implemented in this preselection, cause many excess events in some samples. This can be seen from the difference between the height of the purple bar (showing the final AOD selection count) and the height of the dashed line (showing the TAG selection).

For the signal sample, there is only a small fraction of excess events. The $Z \to \mu^-\mu^+$ background has been eliminated for the ~10000 event sample. However the preselection is not so effective for $Zb\bar{b}$ and especially $t\bar{t}$ and $Z \to \mu^-\mu^+$, which is particularly unfortunate given the large cross-section of the latter two backgrounds.

In the AOD selection, a substantial fraction of $t\bar{t}$ is reduced by cut 2e, which is based on the η of jets. Considering the η spread displayed in 3.10, it is not possible to achieve this with TAGs.

Still, all three discussed background samples have been reduced by at least one order of magnitude.

3.6.4 A noninclusive preselection

It has been shown that a large contribution of the excess events in the TAG preselection are caused by ignoring the cuts on jet properties and on the muon isolation.

Figures 3.7 also showed that in this context the calorimeter isolation is of greater importance than the track isolation. As a matter of fact, the TAG format does contain the calorimeter isolation for muons. The only problem was the different cone size, which made an inclusive preselection impossible.

We shall discuss in how far the preselection can be improved by giving up the premise of keeping it inclusive. The goal will be to reach a similar background rejection as the original cuts.

Calorimeter isolation

Figure 3.13 shows the distribution of the calorimeter isolation divided by p_T both for cones with $\Delta R = 0.3$ and $\Delta R = 0.4$. The histogram includes events that have two muons with $p_\text{T} > 15$ GeV.

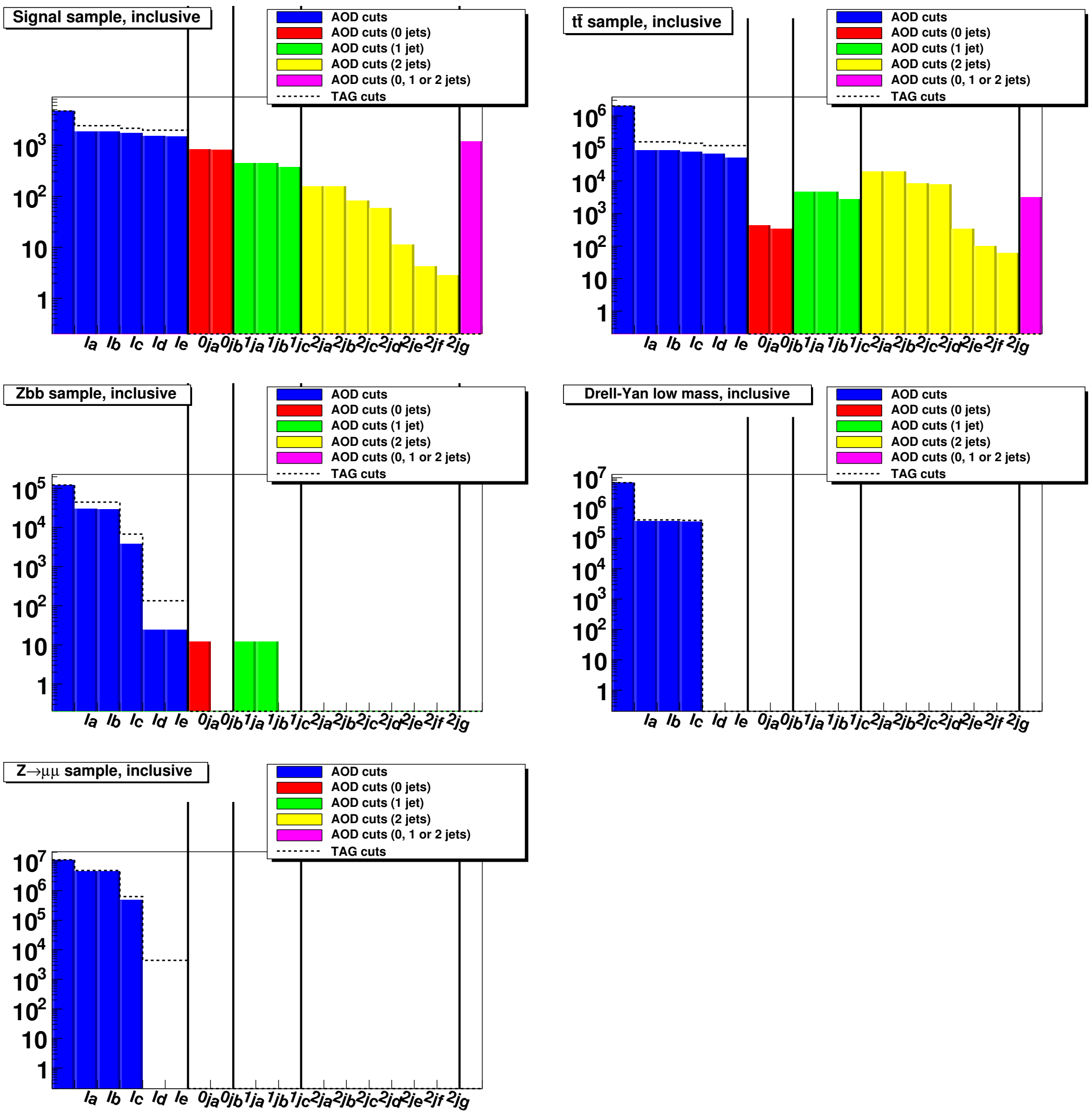


Figure 3.12: Inclusive cut flow. The labels of the x axis refer to the cuts explained in appendix B.

The blue bars represent the cut flow for the cuts that are common for all jet bins.

The blue, green and yellow bars represent cuts that are only done for specific jet bins. The rightmost bar represents the sum of the final cuts (0ja, 1jc, 2jg).

The dashed line shows the cut flow of the inclusive TAG preselection, which has only be implemented for cuts Ia-Ie.

Histograms are scaled to $\int\mathscr{L}\,\mathrm{d}t = 10\ \mathrm{fb}^{-1}$.

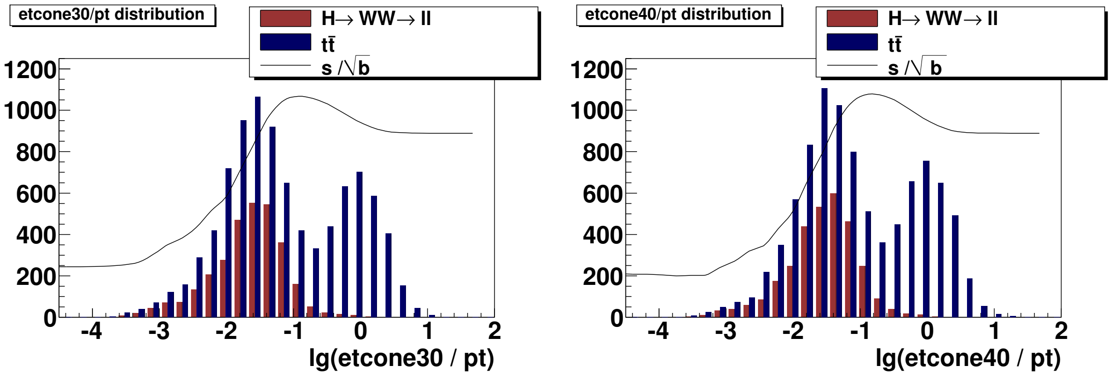


Figure 3.13: Distribution of $E_{T,\mathrm{cone}}/p_\mathrm{T}$ for signal and $t\bar{t}$ background and the cut significance. The bars are not normalized to the same integrated luminosity.

The lepton selection required that

$$\frac{p_\mathrm{T}}{E_{T,\mathrm{cone},30}} < 0.2. \tag{3.7}$$

The figure shows that this cut removes a large portion of the $t\bar{t}$ background while keeping nearly all of the $H \to W^+W^- \to \ell\nu\bar{\ell}\bar{\nu}$ events. The black curve shows $\mathrm{s}/\sqrt{\mathrm{b}}$ as a measure for the cut optimization at the respective point.

The curve has a maximum at 0.2, indicating that this cut is suited to separate the background. In fact, the same is true for the $E_{T,\mathrm{cone},40}$. Therefore this variable will be used for a noninclusive preselection. Similar results have been obtained for the $E_{T,\mathrm{cone}} < 30$ GeV cut.

Jet definition

Figure 3.10 demonstrated that the two jet definitions are largely comparable. Therefore the jet-based cuts can be implemented with TAGs too.

Resulting cuts

The cuts that have been used for the inclusive preselection will be modified to include the calorimeter isolation for muons and the cuts that use variables for jets. Now it is also possible to implement the cuts that depend on the number of jets. Figures 3.14 and 3.15 show the resulting cut flow as well as the fraction of the events from the original selection that will not be found with the TAG selection.

It can be seen that the cut on the number of leptons (Ia) shows hardly any difference between AOD and TAG, which was not the case for the inclusive preselection. This demonstrates the benefit of using the muon isolation.

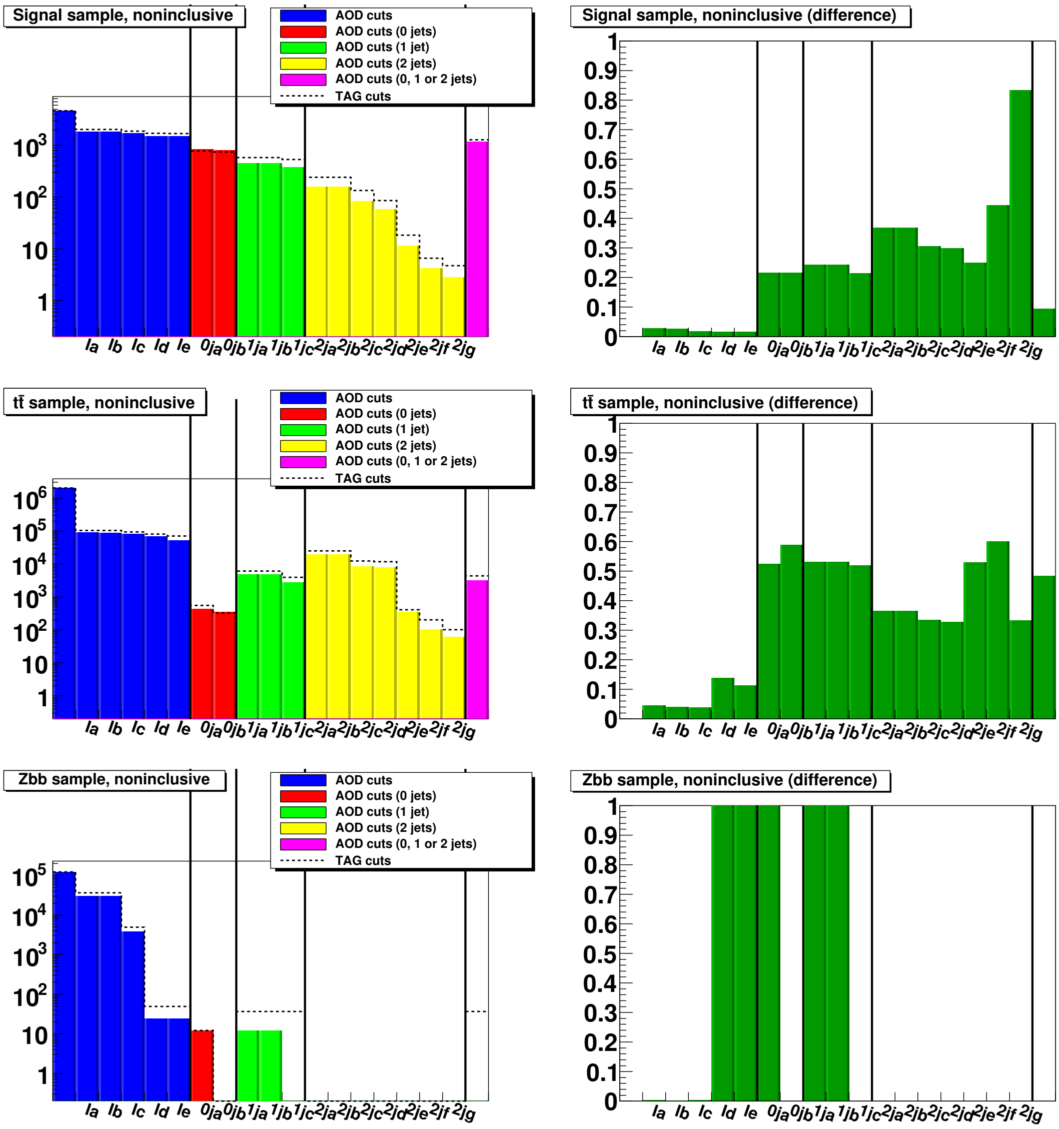


Figure 3.14: Noninclusive cut flow (signal, $t\bar{t}$ and $Zb\bar{b}$).

In the left column, the rightmost bin (purple bar) shows the final selection, i. e. events that passed any of 0jb, 1jc or 2jg (for TAGs: 2je).

The histograms in the right column show the events that are in the AOD selection but not in the TAG selection, normalized to the events in the AOD selection.

Histograms are scaled to $\int \mathscr{L}\,\mathrm{d}t = 10\ \mathrm{fb}^{-1}$.

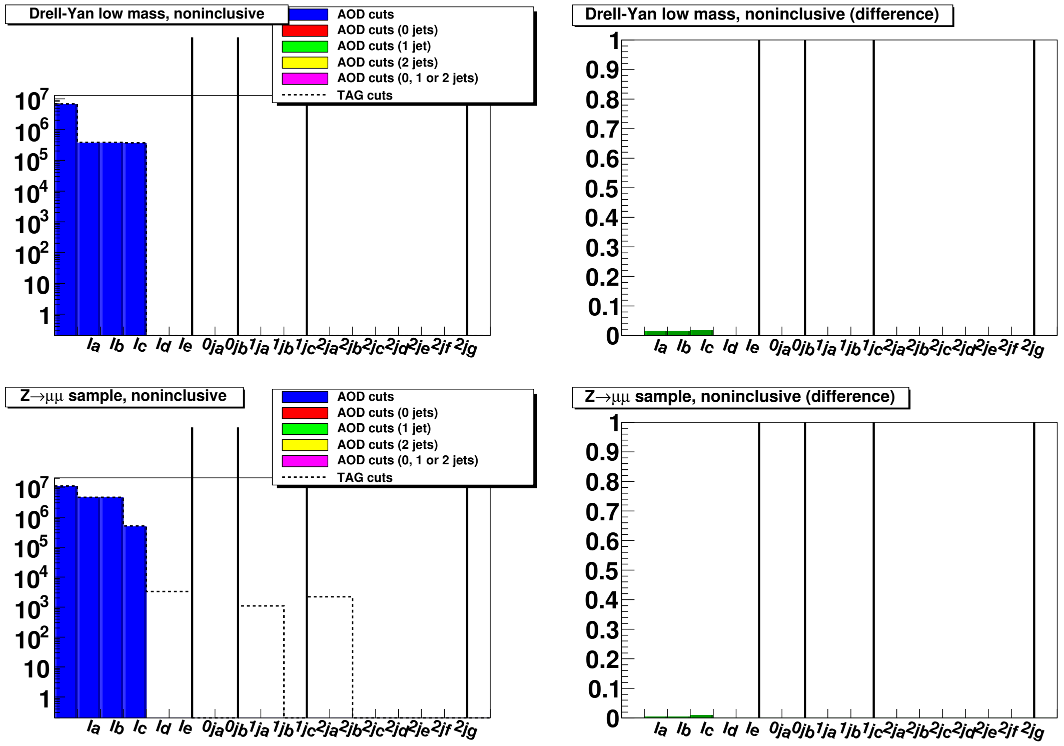


Figure 3.15: Noninclusive cut flow (Drell-Yan).

The histograms in the right column show the events that are in the AOD selection but not in the TAG selection, normalized to the events in the AOD selection. Histograms are scaled to $\int \mathscr{L}\,\mathrm{d}t = 10\ \mathrm{fb}^{-1}$.

Moreover, the jet cuts substantially improved the TAG selection. The final numbers of both selections are very close, which can be seen by comparing the purple bar against the dashed line right above it.

The $Z \to \mu^-\mu^+$ background sample has no events after the final cuts for either selection, and $Zb\bar{b}$ still has a very small amount of events after the TAG selection.

The right column (dark green bars) shows the fraction of the events from the AOD selection that will not be found by the TAG selection. In terms of table 3.1, this is the ratio $a_{10}/\left(a_{10}+a_{11}\right)$, which was required to be zero for the inclusive preselection.

The rightmost bar in the dark green histograms represents the portion of event selection (i. e. events having passed all cuts in any of the three bins) that can not be reached with TAG cuts. In the $H \to W^+W^- \to \ell\nu\bar{\ell}\bar{\nu}$ histogram, this fraction is smaller than all of the three final cuts (0jb, 1jc, 2jg). This is because the TAG cuts classify some events into a different jet bin than the AOD cuts, i. e. some events are selected by both TAG and AOD even though the two algorithms count a different number of jets.

Only about 10% of the signal events are missing in this preselection. The missing events can further be reduced by omitting the cuts 2jf and 2jg for the TAG case. These cuts use the invariant mass of jets and the total transverse momentum, which had to be calculated from several TAG variables that differ from their corresponding in AOD variable, resulting in high discrepancies. This induced a loss of several events in the 2-jet bin, which is very weakly populated (about 10 events in this sample). Omitting cuts 2jf and 2jg does not have a great influence on the $t\bar{t}$ background compared to other cuts that are more effective.

After all cuts, the Drell-Yan samples have no remaining events with either selection. The $Zb\bar{b}$ background is also zero after the AOD cuts, however the TAG selection leaves a small amount of events in this sample. The only background that has a considerable amount of events left after all cuts is $t\bar{t}$, with only few more events in the TAG selection. The next section will show that the resulting TAG cuts are in fact very usable.

3.7 The Higgs mass peak

A transverse Higgs mass can be defined [31, 5] in terms of reconstructed variables as

$$m_{\mathrm{T},H} := \sqrt{\left(E_{\mathrm{T},\ell\ell} + \not{E}_{\mathrm{T}}\right)^2 - \left(\vec{p}_{\mathrm{T},\ell\ell} + \vec{\not{p}}_{\mathrm{T}}\right)^2} \tag{3.8}$$

where

$$E_{\mathrm{T},\ell\ell} := \sqrt{p_{\mathrm{T},\ell\ell}^2 + M_{\ell\ell}^2}. \tag{3.9}$$

	All channels	$\mu\mu$ channel
AOD selection	4 341	1 276
Inclusive TAG selection	124 693	41 384
Noninclusive TAG selection	5 980	1 665

Table 3.3: Number of events after the final selections (sum of all samples, scaled to $\int\mathscr{L}\,\mathrm{d}t = 10$ fb)

Figure 3.16 shows the distribution of $m_{\mathrm{T},H}$ for the samples at hand after various cuts. Because of $m_{\mathrm{T},H} \leq m_{\mathrm{H}}$, the right edge of the histogram is an indicator to determine the Higgs mass once H has been detected.

In the early stages of the selection, the large background contribution completely hides the $H \to W^+W^- \to \ell\nu\bar{\ell}\bar{\nu}$ events and therefore the Higgs mass peak. After all cuts, although $t\bar{t}$ still has a considerable presence in the histogram, a clear peak is recognizable in the AOD selection, diminishing at the simulated Higgs mass (170 GeV).

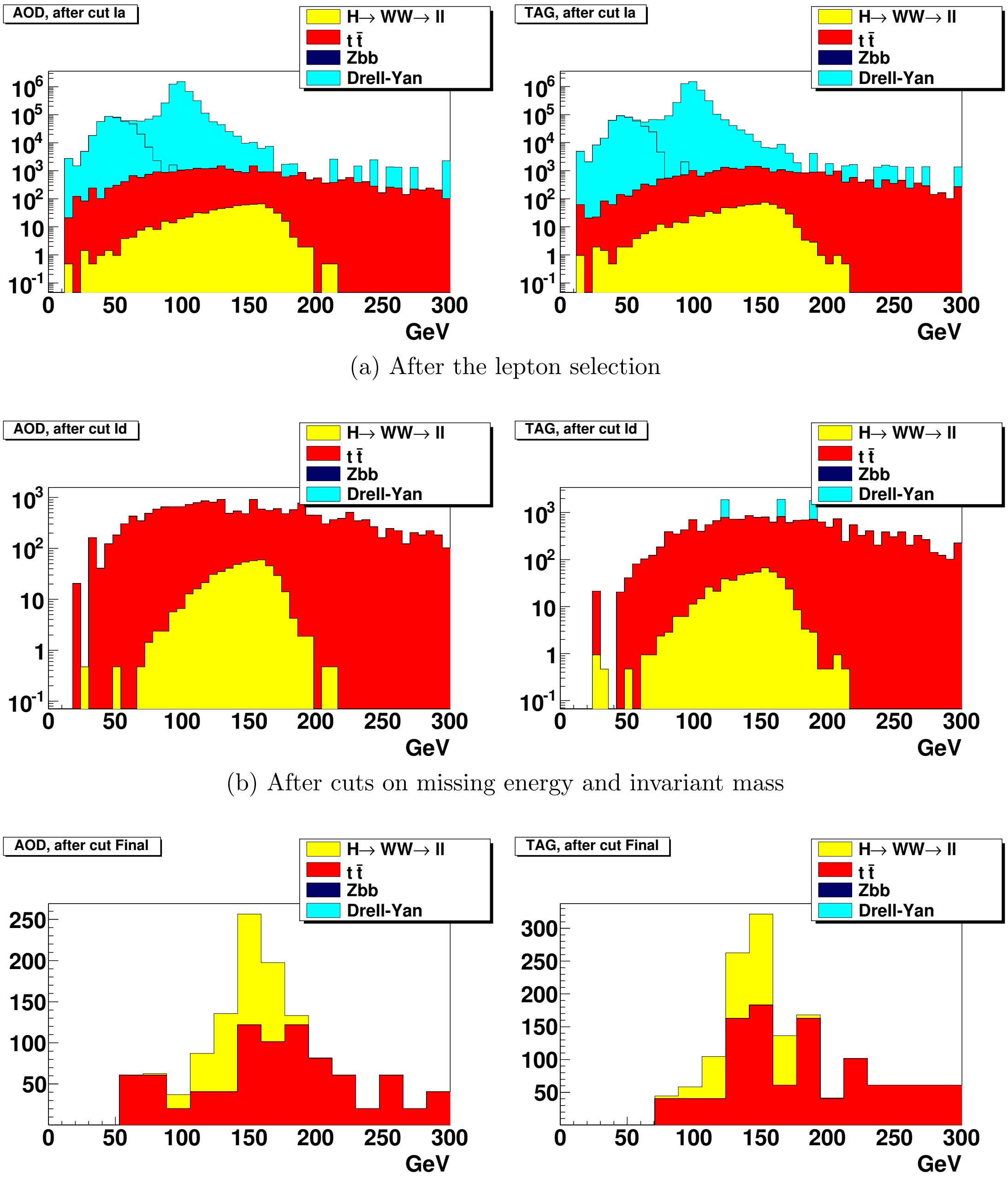
Both the event selection and the calculation of $m_{\mathrm{T},H}$ can be accomplished based on TAG files. The right column of figure 3.16 shows the result, which was obtained using the non-inclusive event selection described in the previous section. After cut Ia or Id, the histogram looks very similar to the original one on the left-hand side, reassuring the correctness of the TAG selection. After the final cut, the H peak can be seen too. It is not as clear as in the AOD case, however it does allow a good approximation of the Higgs mass.

3.8 Summary

$t\bar{t}$, $Zb\bar{b}$ and $Z \to \mu^-\mu^+$ are large backgrounds for the $H \to W^+W^- \to \ell\nu\bar{\ell}\bar{\nu}$ analysis. However the latter two can be easily reduced by a few well-chosen cuts (assuming the Higgs mass of 170 GeV). $t\bar{t}$ poses a greater problem, as it seems not to be possible to completely separate it from the signal process. The required cuts also reduce a large amount of the Higgs events. Still it is possible to visualize the Higgs mass spectrum in the $m_{\mathrm{T},H}$ histogram, which can also be achieved by using only a restricted set of variables like those found in TAGs.

The results presented in this chapter have been normalized to $\int\mathscr{L}\,\mathrm{d}t = 10$ fb, which can be recorded during one year at the design luminosity of LHC. In section 2.8 it has been estimated that this corresponds to a muon stream with 2×10^9 events and 200 TB of AODs.

As shown in table 3.3, for the $\mu\mu$ channel the TAG selection leaves 41 384 and



(a) After the lepton selection

(b) After cuts on missing energy and invariant mass

(c) After cuts 0jb/1jc/2jg (for TAG: 2je)

Figure 3.16: The Higgs transverse mass. For these histograms, events with two muons have been used. ($\int \mathscr{L}\,\mathrm{d}t = 10\ \mathrm{fb}^{-1}$)

1 665 in the inclusive and noninclusive case, respectively. Using these events as a preselection for an AOD analysis can combine the quickness of TAGs with the accuracy of AODs.

For 200 TB, the number of AOD files will be of the order of 100 000, which is higher than the number of events in the preselection. Therefore, in a distributed analysis based on the TAG preselection, the greater part of the AOD files need not be accessed at all. The remaining AOD files will be accessed by the back-navigation mechanism. In each of these files, only one or a few events will be accessed. It can be assumed that most of the analysis time will be taken up by the queuing of the analysis jobs (up to one hour), and most of the actual job run time will be spent in the startup phase of the Athena framework (about one minute on each computing element). Note that these numbers do not increase much for larger datasets, because they represent overhead that is only needed once per analysis.

This is a high speedup compared to the full AOD analysis, which was estimated in section 2.8 to take more than one day, a time that can only be achieved if the Grid is not occupied with other jobs.

Chapter 4

Conclusion and Outlook

TAGs provide a means to make event selections for many purposes. If only a few runs are to be processed, TAG information can be accessed with few resources. For this task the TAG database can be used, or TAG files may be downloaded from the ATLAS storage infrastructure in order to process them locally, which is a larger undertaking even for a single AOD file.

Although TAG selections can get very close to the ones that are designed with maximum significance in mind, there are still some discrepancies. Despite this, an inclusive preselection can be achieved, which enables one to perform an analysis that uses the full accuracy of AOD events and still take advantage of fast event selections as provided by TAGs. However, TAG variables alone can also accomplish tasks like searching for $H \to W^+W^- \to \ell\nu\bar{\ell}\bar{\nu}$ events and reduce important backgrounds. Taking advantage of this, the small footprint of TAG analyses permits monitoring of LHC data. Creating a $m_{\mathrm{T},H}$ histogram with TAGs alone can under ideal circumstances provide a fast answer whether a data sample contains Higgs candidates.

It is possible to further reduce the above-mentioned discrepancies. Optimizations in the TAG definition, but also the consideration of TAGs when designing a selection could lead to substantial improvements:

- Changing the TAG specification by replacing some variables with more sophisticated ones could do some great benefit. As a matter of fact, some changes have already been done. The latest version of TagTool includes `MET_RefFinal` rather than `MET_Final` for the missing energy [24]. This may improve the suitability of TAGs for reducing backgrounds such as $Zb\bar{b}$.

- Keeping TAG information in mind while designing physics analyses could also give large benefits. For instance, the $E_{T,\mathrm{cone},40}$ and $E_{T,\mathrm{cone},30}$ have been shown to be similarly helpful for a separation of $t\bar{t}$ from $H \to W^+W^- \to \ell\nu\bar{\ell}\bar{\nu}$, so one might consider preferring the variables that TAGs provide. One might even explicitly include a TAG-compatible preselection as part of the full selection.

Appendix A

Contents of TAG files

Variable	Name in TAG files
General	
Run number	`RunNumber`
Event number	`EventNumber`
Event type	`isCalibration, isTestBeam, isSimulation, isRealData`
Number of tracks	`NTrk`
Number of vertices	`Nvx`
Primary vertex	`VtxX, VtxY, VtxZ`
Streaming criteria/results	`Stream`
Random number	`RandomNumber`
Time stamp	`TimeStamp`
Bunch by bunch luminosity	`Luminosity`
Luminosity block number	`LumiBlockN`
Missing energy	`MissingET`
φ of missing E_{T}	`MissingETPhi`
Summed cell E_{T}	`SumET`
Data quality	
Detector Status	`Status***` (49 variables)
Trigger information	
CTP decisions	`CTPWord0–23`
Level 1 trigger type	`Level1TriggerType`
Level 2 trigger masks	`L2PassedTrigMask0–31`
Event filter masks	`EFPassedTrigMask0–31`
Back references	
AOD reference	`StreamAOD_ref`
ESD reference	`StreamESD_ref`
RAW reference	`Stream1_ref`
Physics attributes	
Electron/photon identification	`EgammaWord`
Muon identification	`CombinedMuonWord`
Jet missing E_{T} identification	`JetMissingEtWord`
Tau identification	`TauIdWord`
Jet tagging	`JetTagWord`
B-physics analysis	`BPhysWord`
Exotic physics analysis	`ExoticWord`
Higgs physics analysis	`HiggsWord`
SUSY physics analysis	`SUSYWord`
SM physics analysis	`SMWord`
Top physics analysis	`TopWord`
Heavy ion analysis	`HeavyIonWord`

Table A.1: Contents of TAG files: Event-specific information [22]

Variable	Name in TAG files
Electrons (4)	
Total number of loose electrons	NLooseElectron
Loose electron p_{T} (signed)	LooseElectronPt1-4
Loose electron η	LooseElectronEta1-4
Loose electron φ	LooseElectronPhi1-4
Loose electron tightness	LooseElectronTightness1-4
Muons (4)	
Total number of loose muons	NLooseMuon
Loose muon p_{T} (signed)	LooseMuonPt1-4
Loose muon η	LooseMuonEta1-4
Loose muon φ	LooseMuonPhi1-4
Loose muon tightness	LooseMuonTightness1-4
Loose muon isolation E_{T}	LooseMuonIsolEt1-4
Loose muon track isolation	LooseMuonIsolN1-4
Photons (2)	
Total number of loose photons	NLoosePhoton
Total number of loose converted photons	NLooseConvertedPhoton
Loose photon p_{T}	LoosePhotonPt1-2
Loose photon η	LoosePhotonEta1-2
Loose photon φ	LoosePhotonPhi1-2
Loose photon tightness	LoosePhotonTightness1-2
Jets (6)	
Total number of jets	NJet
Total number of b-tagged jets	NBJet
Jet p_{T}	JetPt1-6
Jet η	JetEta1-6
Jet φ	JetPhi1-6
B-tag likelihood	BJetLikelihood1-6
Summed E_{T} over jets	JetSumET
Tau jets (2)	
Total number of tau jets	NTau
Tau jet p_{T}	TauJetPt1-2
Tau jet η	TauJetEta1-2
Tau jet φ	TauJetPhi1-2
Tau jet number of tracks	TauJetNTrk1-2
Tau jet likelihood	TauJetLikelihood1-2

Table A.2: Contents of TAG files: Particles and jets [22]

Version 14 variable	Version 15 variable
(not present)	`BunchId`
`CTPWord0–23`	`L1PassedTrigMask0–23`
(not present)	`DPDWord` (DPD information)

Table A.3: Contents of TAG files: Differences between versions 14 and 15 [23]

Appendix B

Cuts used for $H \to W^+W^- \to \ell\nu\bar{\ell}\bar{\nu}$ event selection

B.1 Lepton and jet selection

First, some cuts are made on the leptons and jets in each event to decide which objects are to be considered for further analysis.

Muon selection

This analysis uses muons from the STACO reconstruction algorithm.

m1 $p_{\mathrm{T}} > 15$ GeV

m2 $|\eta| < 2.5$

m3 Muons must have the "combined" flag.

m4 $\chi^2 < 100$ (from the combined muon matching in the STACO algorithm)

m5 Impact parameter significance with respect to primary vertex must be smaller than 10.

m6 If after the above cuts there are two muons are within a cone of $\Delta R < 0.3$, then only the one with higher p_{T} is taken.

m7 Calorimeter isolation: $\frac{E_{T,\mathrm{cone,30}}}{p_{\mathrm{T}}} < 0.2, \quad E_{T,\mathrm{cone,30}} < 10$ GeV

m8 Track isolation (.30 cones): $\frac{\mathtt{trkIso}}{p_{\mathrm{T}}} < 0.1, \quad \mathtt{trkIso} < 10$ GeV.

Electron selection

e1 $p_{\mathrm{T}} > 15$ GeV

e2 $|\eta| < 2.47$, excluding the crack-region: $1.37 < |\eta| < 1.52$

e3 Electron has been reconstructed with the cluster-based algorithm

e4 *Tight* electrons; *medium* for the 2-jet channel.

e5 Impact parameter significance with respect to primary vertex must be smaller than 10.

e6 If after the above cuts there are two electrons within a cone of $\Delta R < 0.3$, then only the one with the higher p_{T} is taken.

e7 If after the above cuts there is a muon within a cone of $\Delta R < 0.3$ around an electron, then the electron is discarded.

e8 Calorimeter isolation: $\frac{E_{T,\mathrm{cone},30}}{p_{\mathrm{T}}} < 0.2, \quad E_{T,\mathrm{cone},30} < 10$ GeV

e9 Track isolation (.30 cones): $\frac{\texttt{trkIso}}{p_{\mathrm{T}}} < 0.1, \quad \texttt{trkIso} < 10$ GeV.

Jet tagging

In j3 and j4, the electrons and muons are taken from the lepton selection described above.

j1 $p_{\mathrm{T}} > 20$ GeV

j2 $|\eta| < 3.0$

j3 If there are a *medium* electron and a jet within a 0.4-cone, then the jet is discarded.

j4 If there are a muon and a jet within a 0.4-cone, then the jet is discarded.

B.2 Higgs candidate preselection

After the leptons and jets have been selected, several cuts determine whether the event is to be included in the analysis:

Ia Exactly two opposite sign leptons from the selection above are required.

Ib Veto event if another electron is found that is a medium electron from the cluster-based algorithm with the following criteria: $p_\mathrm{T} > 15$ GeV, $|\eta| < 2.47$. The veto electron is ignored if it is within a 0.05-cone near any of the two selected leptons.

Ic Invariant mass of the two leptons:

$$\begin{cases} M_{\ell\ell} > 15 \text{ GeV} & \text{for } e\mu \\ |M_{\ell\ell} - M_Z| > 10 \text{ GeV} & \text{for same flavor} \end{cases}$$

(with $M_Z = 91.2$ GeV).

Id Missing energy cut:

$$\not{E}_\mathrm{T} > \begin{cases} 30 \text{ GeV} & \text{for } e\mu \\ 40 \text{ GeV} & \text{for same flavor} \end{cases}$$

Ie $m_\mathrm{T} > 30$ GeV

For this cut, a preliminary Higgs transverse mass is defined as

$$m_\mathrm{T} := \sqrt{2\left(\vec{p}_{\ell 1} + \vec{p}_{\ell 2}\right)_\mathrm{T} \not{E}_\mathrm{T} \left(1 - \cos \mathrm{d}\varphi_{\ell\ell}\right)} \tag{B.1}$$

where $\mathrm{d}\varphi_{\ell\ell}$ is the angle between the di-lepton vector and the missing transverse momentum vector in the transverse plane.

B.3 Final Higgs search selection

The following are specific to the number of tag jets from the jet selection above:

0-jet bin

0ja No tag jets in the event

0jb $(\vec{p}_{\ell 1} + \vec{p}_{\ell 2})_\mathrm{T} > 30$ GeV

1-jet bin

1ja Exactly one tag jet

1jb b-veto: For the one jet, if $|\eta| < 2.5$ and the SV0 b-tagging weight is > 5.7, then the event is removed.

1jc $p_{\mathrm{T,tot}} < 30$ GeV where $p_{\mathrm{T,tot}}$ is the magnitude of the two-dimensional vector sum $\vec{p}_{\mathrm{T},\ell 1} + \vec{p}_{\mathrm{T},\ell 2} + \vec{p}_{\mathrm{T},j} + \vec{\not{p}}_\mathrm{T}$.

2-jet bin

2ja Exactly two tag jets

2jb b-veto as in 1jb

2jc Jets should be in opposite hemispheres: $\eta_{j1}\eta_{j2} < 0$.

2jd $p_{\mathrm{T}} > 40$ GeV for the leading jet

2je $|\eta_{j1} - \eta_{j2}| > 3.8$

2jf Invariant mass of the two jets: $M_{jj} > 500$ GeV

2jg $p_{\mathrm{T,tot}} < 30$ GeV where $p_{\mathrm{T,tot}}$ is the magnitude of the two-dimensional vector sum $\vec{p}_{\mathrm{T},\ell 1} + \vec{p}_{\mathrm{T},\ell 2} + \vec{p}_{\mathrm{T},j1} + \vec{p}_{\mathrm{T},j2} + \vec{\not{p}}_{\mathrm{T}}$.

Bibliography

[1] ATLAS homepage: *What is ATLAS?* `http://atlas.ch/what_is_atlas.html`, retrieved on 2010-05-10.

[2] Joachim Tuckmantel: *Synchrotron Radiation Damping in LHC and Longitudinal Bunch Shape.* CERN-LHC-Project-Report-819, Geneva, 2005.

[3] ATLAS Photos – Detector Site – Surface, `http://www.atlas.ch/photos/detector-site-surface.html`

[4] Christoph Berger: *Elementarteilchenphysik – Von den Grundlagen zu den modernen Experimenten.* Aachen, 2006.

[5] Johannes Ebke: private communication.

[6] The ATLAS Collaboration: *Expected Performance of the ATLAS Experiment: Detector, Trigger and Physics.* Technical report, CERN-OPEN-2008-020, Geneva, 2008.

[7] `https://twiki.cern.ch/twiki/bin/view/Atlas/LuminosityBlock`

[8] ATLAS Collaboration: *ATLAS Detector and Physics Performance – Technical Design Report.* Volume 1, revision 0, 1999.

[9] Benjamin Ruckert: *Muon Reconstruction and the Search for Leptoquarks at LHC.* Diploma thesis, LMU München, 2006.

[10] ATLAS Computing Group: *ATLAS Computing – Technical Design Report,* revision 3, 2005.

[11] Torbjorn Sjostrand, Stephen Mrenna, and Peter Skands: *PYTHIA 6.4 Physics and Manual.* JHEP, 0605:026, 2006.

[12] S. Frixione and B. R. Webber: *Matching NLO QCD computations and parton shower simulations.* JHEP, 0206:029, 2002.

[13] M. L. Mangano, M. Moretti, F. Piccinini, R. Pittau, and A. D. Polosa: *ALPGEN, A Generator for Hard Multiparton Processes in Hadronic Collisions.* JHEP, 0307:001, 2003.

[14] G. Corcella, I. G. Knowles, G. Marchesini, S. Moretti, K. Odagiri, P. Richardson, M. H. Seymour, and B. R. Webber: *HERWIG 6.5: an event generator for Hadron Emission Reactions With Interfering Gluons (including supersymmetric processes).* JHEP, 0101:010, 2001.

[15] The GEANT4 web site:
`http://geant4.web.cern.ch/geant4/`

[16] Matthias Schott: *Study of the Z Boson Production at the ATLAS Experiment with First Data.* PhD thesis, LMU München, 2007.

[17] L. Asquith et al: *Performance of Jet Algorithms in the ATLAS Detector.* ATL-COM-PHYS-2009-630, 2009.

[18] B. Abbott22 et al: *Study on reconstructed object definition and selection for top physics.* ATL-COM-PHYS-2009-633, 2009.

[19] Gabriele Reiter: *Studien zur Higgs-Produktion im Kanal $H \rightarrow W^+W^- \rightarrow \mu^+\mu^-\nu_\mu\bar{\nu}_\mu$ mit dem ATLAS Detektor am LHC.* Diploma thesis, LMU München, 2007.

[20] *Full Dress Rehearsal,*
`https://twiki.cern.ch/twiki/bin/view/Atlas/FullDressRehearsal`

[21] *FDR-2 Data for Users,*
`https://twiki.cern.ch/twiki/bin/view/Atlas/FDR2ForUsers`

[22] *TagForEventSelection – Release 14.2.10,*
`https://twiki.cern.ch/twiki/bin/view/AtlasProtected/TagForEventSelection14`

[23] *TagForEventSelection – Release 15.X.0,*
`https://twiki.cern.ch/twiki/bin/view/AtlasProtected/TagForEventSelection15`

[24] Change in the ATLAS source code from 2009-10-16,
`https://svnweb.cern.ch/trac/atlasoff/changeset/219848/PhysicsAnalysis/JetMissingEtID/JetMissingEtTagTools/trunk/src/JetMissingEtTagTool.cxx`

[25] Source code of `HiggsPhysTagTool`, the version that was current as of 2010-05,
`https://svnweb.cern.ch/trac/atlasoff/browser/PhysicsAnalysis/HiggsPhys/HiggsPhysTagTools/trunk/src/HiggsPhysTagTool.cxx?rev=219846`

[26] Source code of `ElectronTagTool`,
`https://svnweb.cern.ch/trac/atlasoff/browser/PhysicsAnalysis/ElectronPhotonID/ElectronPhotonTagTools/trunk/src/ElectronTagTool.cxx`

[27] Source code of `MuonTagTool`,
`https://svnweb.cern.ch/trac/atlasoff/browser/PhysicsAnalysis/MuonID/MuonTagTools/trunk/src/MuonTagTool.cxx`

[28] `ChainTagMap.xml`,
`http://alxr.usatlas.bnl.gov/lxr/source/atlas/Trigger/TriggerCommon/TriggerMenuXML/data/ChainTagMap.xml?v=head`

[29] The HSG3: *Analysis Strategies of the HSG3 Subgroup with the First Data,* March 28, 2009.

[30] *MET correction for Higgs WG,*
`https://twiki.cern.ch/twiki/bin/view/AtlasProtected/MetCorrectionForHiggsWG` (accessible for ATLAS members)

[31] *Higgs WG Subgroup: $H \to WW$ (gg, VBF, WH, ttH, VBF invisible H),*
`https://twiki.cern.ch/twiki/bin/view/AtlasProtected/HiggsWW` (accessible for ATLAS members)

[32] *Datasets for HSG3: $H \to WW$ (10* TeV*),*
`https://twiki.cern.ch/twiki/bin/viewauth/AtlasProtected/HiggsWGHSG3Dataset` (accessible for ATLAS members)

[33] *Combined CDF and DZero Upper Limits on Standard Model Higgs-Boson Production with up to 4.2* fb^{-1} *of Data.* FERMILAB-PUB-09-060-E, 2009.

Acknowledgements

Ich möchte Prof. Dorothee Schaile für das Anfertigen des Erstgutachtens danken und dafür, dass sie mir durch die Aufnahme an ihrem Lehrstuhl und die Hilfe bei der Themenauswahl diese Diplomarbeit ermöglicht hat.

Für die Erstellung des Zweitgutachtens möchte ich Dr. Hans von der Schmitt meinen Dank ausdrücken.

Günter Duckeck danke ich für die Betreuung und das Korrekturlesen meiner Diplomarbeit. Ferner danke ich Johannes Ebke, Benjamin Ruckert und Johannes Elmsheuser für weitere Hilfe in physikalischen und technischen Angelegenheiten.

Den Lehrstuhlmitgliedern danke ich für die gute Arbeitsatmosphäre und weitere Unternehmungen wie die Bergtour und die Cocktailabende. Insbesondere vielen Dank an Johannes Ebke, Benjamin Ruckert, Thomas Langer, Jonas Will, Andreas Schacker und Stefan Petrovics für die nette Stimmung im Büro.

Meinem Chef Martin Kerscher danke ich für sein Verständnis, dass ich in diesem Jahr meinem Nebenjob nicht viel Aufmerksamkeit schenken konnte. Ich möchte Martin Kerscher, Jens Schmalzing und Walter Spann auch dafür danken, dass sie mir Gelegenheit gegeben haben, während des Studiums viele Erfahrungen im Rechenzentrum zu sammeln.

Ich danke vielen Kommilitonen und anderen Freunden, die mich in meiner Studentenzeit begleitet haben. Insbesondere bin ich meiner Kommilitonin Halina für ihre vielseitige Unterstützung während des Diploms dankbar.

Meinen Eltern und meinem inzwischen leider verstorbenen Großvater danke ich für die Unterstützung meines Studiums. Meinen Eltern und meinem Bruder auch für die nette Familie, die ich in den Semesterferien gerne besucht habe.

Selbstständigkeitserklärung

Ich versichere hiermit, die vorliegende Arbeit selbstständig verfasst zu haben und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet zu haben.

Christoph Bußenius