IEKP-KA/2006-20

Suche nach $H \to b\bar{b}$ in Assoziation mit einem $t\bar{t}$ Paar in Proton-Proton Kollisionen bei $\sqrt{s} = 14$ TeV

Alexander Schmidt

Zur Erlangung des akademischen Grades eines
DOKTORS DER NATURWISSENSCHAFTEN
von der Fakultät für Physik der Universität (TH)
Karlsruhe

genehmigte

DISSERTATION

von

Dipl.-Phys. Alexander Schmidt
aus Schwetzingen

Tag der mündlichen Prüfung: 1.12.2006

Referent: Prof. Dr. G. Quast, Institut für Experimentelle Kernphysik

Korreferent: Prof. Dr. Th. Müller, Institut für Experimentelle Kernphysik

Deutsche Zusammenfassung

Die Elementarteilchenphysik beschäftigt sich mit den elementaren Bausteinen von Materie und Strahlung sowie deren Wechselwirkungen. Bereits im sechsten Jahrhundert vor Christus wurde das philosophische Konzept des Atomismus von griechischen Philosophen, wie zum Beispiel Demokrit, diskutiert. Aber erst im Jahre 1897 konnte das erste Elementarteilchen, das Elektron, von J. J. Thomson in Kathodenstrahlen experimentell nachgewiesen werden. Im zwanzigsten Jahrhundert folgte die Etablierung der Kernphysik und Quantenmechanik, die in den 50er und 60er Jahren in die Entdeckung einer Vielfalt von Teilchen mittels Streuexperimenten mündete. Dieser zunächst unerklärliche "Teilchenzoo" konnte dann in den 70er Jahren im Standardmodell der Teilchenphysik auf Kombinationen von wenigen elementaren Teilchen zurückgeführt werden. Insbesondere die Erkenntnis, dass die Strukturen der elementaren Bausteine und deren Wechselwirkungen grundlegenden mathematischen Konzepten folgen, kann als eine der herausragenden Entdeckungen des zwanzigsten Jahrhunderts betrachtet werden. Eine wichtige Rolle bei der mathematischen Beschreibung spielen lokale Eichtheorien, die die Existenz von Austauschbosonen auf die Forderung nach Symmetrieprinzipien zurückführen. Dem gegenwärtigen Verständnis entsprechend lässt sich das Standardmodell durch eine $\mathrm{SU}(3) \times \mathrm{SU}(2) \times \mathrm{U}(1)$ Symmetrie beschreiben. Aus dieser folgt die Existenz von acht Gluonen der starken Wechselwirkung und vier Bosonen der elektroschwachen Wechselwirkung, darunter das Photon. Die Massen der elektroschwachen Austauschbosonen spielen hierbei eine besondere Rolle, da sie nicht durch einen Massenterm in der Lagrangedichte erzeugt werden können. Ein vielversprechender Erklärungsansatz ist eine spontane Symmetriebrechung, der sogenannte Higgsmechanismus, der mit der Existenz eines skalaren Bosons, dem Higgs Boson, verbunden ist. Die Suche nach diesem Higgs Boson ist eine der Herausforderungen der heutigen Hochenergiephysik, der sich auch die vorliegende Studie widmet. Die Entdeckung oder die Bestimmung von Ausschlussgrenzen stellen einen wichtigen Schritt zu einer Vervollständigung unseres Wissens über die elementaren Vorgänge der Natur dar.

Am Europäischen Zentrum für Nuklearforschung (CERN) in Genf wird zur Zeit ein neuer Teilchenbeschleuniger aufgebaut, der "Large Hadron Collider" (LHC). Mittels dieses Beschleunigers können Schwerpunktsenergien von 14 TeV erreicht werden. Damit wird der Weg in ein neues Kapitel der Teilchenphysik geebnet, denn die Frage nach der Existenz des Higgs Bosons des Standardmodells kann damit mit hoher Wahrscheinlichkeit beantwortet werden. Doch nicht nur die Physik des Higgs Bosons wird am LHC untersucht werden, sondern auch andere Theorien, wie Supersymmetrie und Extradimensionen. Darüber hinaus spielen Präzisionsmessungen im Standardmodell eine wichtige Rolle, um komplementäre Messungen zu vorhergehenden Experimenten zu erhalten und eventuelle Abweichungen von Vorhersagen des Standardmodells, die auf neue Physik hinweisen, zu finden. Auch Physik mit Schwerionen wird am LHC studiert werden. Vier Teilchendetektoren, darunter das "Compact Muon Solenoid" (CMS) Experiment, werden an vier Punkten des 27 km langen Beschleunigerrings

aufgebaut. In diesen Detektoren werden die beschleunigten Hadronen zur Kollision gebracht und die entstehenden Spuren und Energiedepositionen der Sekundärteilchen mit höchster Präzision aufgezeichnet, um daraus Schlussfolgerungen auf den Primärprozess zu ziehen.

In der vorliegenden Arbeit wurde die Möglichkeit untersucht, das Higgs Boson des Standardmodells der Elementarteilchenphysik im Zerfallskanal $H \rightarrow b\bar{b}$ im CMS Experiment zu entdecken. Dieser Zerfallskanal hat im Massenbereich knapp oberhalb der derzeitigen experimentellen Massenuntergrenze, die durch die LEP Experimente bei 114,4 GeV/c^2 festgelegt wurde, das größte Verzweigungsverhältnis. Aufgrund der Vielfalt anderer Quellen von b-Quarks am LHC muss diese Suche in Assoziation mit top-Quark Produktion realisiert werden, denn diese liefert eine klarere Signatur, die eine Entdeckung ermöglichen könnte.

Diese Studie zum Entdeckungspotential von $H \rightarrow b\bar{b}$ wurde im Hinblick auf eine möglichst realistische Abschätzung konzipiert. Es wurde eine vollständige Monte Carlo Simulation des CMS Detektors durchgeführt. Die Trigger- und Rekonstruktionsalgorithmen entsprechen den Algorithmen, die letztendlich auf die realen Daten angewandt werden. Die Analyse, die in dieser Arbeit präsentiert wird, ist die erste, die eine derartige vollständige Simulation und Rekonstruktion verwendet. Demzufolge wurden Beschränkungen gefunden, die bisher nicht bekannt waren, was im Folgenden näher erläutert wird.

Die Anforderungen an die Leistungsfähigkeit des Detektors und an die Rekonstruktionswerkzeuge sind enorm. Aus diesem Grund wurde der $t\bar{t}H$, $H \rightarrow b\bar{b}$ Kanal als sogenannter "benchmark"-Kanal für den "Physics Technical Design Report" (PTDR) [1] ausgewählt, mit dem Ziel, die zur Verfügung stehenden Analyse- und Rekonstruktionsmethoden zu validieren. Die Leistungsfähigkeit der vorhandenen Werkzeuge reichte nicht aus, um zufriedenstellende Ergebnisse zu erhalten. Daher wurden große Anstrengungen unternommen, um diese Werkzeuge zu beurteilen, weiterzuentwickeln und zu optimieren. Insbesondere die Algorithmen zur Identifikation von b-Quark Jets, welche aufgrund der Anwesenheit von vier b-Quarks im Endzustand die mit Abstand wichtigsten Komponenten der $t\bar{t}H$ Analyse darstellen, wurden untersucht und verbessert.

Neben dem Programm zur vollständigen Simulation des CMS Detektors wird im Rahmen der CMS Software eine weitere, schnelle Detektorsimulation entwickelt. Für diese wurde im Verlauf der Arbeit ein signifikanter Beitrag geleistet, indem eine Schnittstelle zu den Identifikationsalgorithmen für b-Quark Jets entwickelt und optimiert wurde. Die Ergebnisse der schnellen Simulation wurden mit denjenigen der vollständigen Simulation und Rekonstruktion verglichen. Verschiedene Ansätze wurden verfolgt um eine möglichst genaue Übereinstimmung zwischen den beiden Simulationsprogrammen zu erhalten. Das wichtigste Resultat besteht in der Erkenntnis, dass die Observablen am Zerfallsvertex des b-Hadrons in sehr guter Übereinstimmung stehen. Die verbleibenden Unterschiede stehen im Zusammenhang mit der Zahl der Spuren, die vom primären Ereignisvertex ausgehen. Diese Studien und Ergebnisse können als wichtiger Schritt hin zu einer zufriedenstellenden Leistungsfähigkeit der schnellen Detektorsimulation und -rekonstruktion verstanden werden.

Des Weiteren wurde im Rahmen der Arbeit ein Beitrag zur Optimierung der b-Quark Identifikationsalgorithmen geleistet. Indem ein Algorithmus, der auf der Rekonstruktion von Sekundärvertizes basiert, mit einem "Soft lepton" Algorithmus kombiniert wurde, welcher Informationen über leptonische b-Hadron Zerfälle berücksichtigt, konnte eine Verbesserung der Leistungsfähigkeit erreicht werden. Insbesondere die Rate von fehlidentifizierten u-, d- und s-Quark Jets sowie von Gluon-Jets konnte um mehr als 15% reduziert werden. Eine weitere Verbesserung bestand in der Verwendung eines Vertexrekonstruktionsalgorithmus, der eine weiterentwickelte Einbindung von Spuren von Tertiärvertizes anwendet. Dadurch konnte die

Steigerung der Leistungsfähigkeit auf insgesamt 25% erhöht werden.

Eine wichtige Neuerung im Zusammenhang mit der $t\bar{t}H$ Analyse bestand in der Abschätzung des Einflusses von systematischen Fehlern. Sowohl die Daten des Detektors als auch die darauf angewandten Algorithmen sind mit systematischen Unsicherheiten verbunden. Beispiele sind die Fehler in Bezug auf die Energieskala von Jets sowie die Identifikations- und Fehlidentifikationsraten von b-Quark Jets. Außerdem sind die Wirkungsquerschnitte der meisten Untergrundprozesse nur in führender Ordnung bekannt. Die Auswirkungen der systematischen Fehler auf die resultierenden Ereignisraten sowie auf das Entdeckungspotential des $t\bar{t}H$ Prozesses wurden von verschiedenen Standpunkten aus betrachtet. Zum einen ist die zu erwartende Leistungsfähigkeit des CMS Detektors und der zugehörigen Analysewerkzeuge entsprechend dem gegenwärtigen Wissensstand und der momentanen Unsicherheiten auf die Vorhersagen zu betrachten. Zum anderen muss erörtert werden, auf welchem Niveau sich die Fehler bewegen, sobald die ersten echten Kontroll- und Kalibrationsdatensätze verfügbar sein werden. Eine dritte Fragestellung behandelt die notwendige Präzision der Instrumente, um eine Entdeckung von $H \rightarrow b\bar{b}$ zu ermöglichen. Die verschiedenen Ansätze gehen mit verschiedenen Fehlermodellen einher. Diese unterschiedlichen Hypothesen wurden untersucht, und es hat sich herausgestellt, dass die Ergebnisse stark von den zugrundeliegenden Annahmen über Art und Modell der Fehler abhängen.

Ein weiteres Themengebiet, das im Verlauf der Arbeit eine Rolle spielte, ist die technische Verwirklichung der Analyse mittels "Grid"-Technologien. Diese trugen dazu bei, die enormen Anforderungen an Datenspeicher und Rechenkapazität, die mit dieser Studie verbunden sind, zu bewältigen. Um die riesigen Datenraten, die an den LHC Experimenten anfallen, zu verarbeiten, müssen verteilte Computing-Konzepte zur Anwendung kommen. Es wird nicht mehr möglich sein, alle Daten an einem Ort bereitzuhalten und zu verarbeiten. In diesem Zusammenhang wurde eine erste Verwirklichung der CMS Datenanalyse via Grid-Werkzeuge am deutschen "Tier 1" Zentrum, GridKA, erreicht. Ein Prototyp von automatisierten Ereigniskatalogen und Datenbanken wurde am GridKA installiert, konfiguriert und im weiteren betreut, um die Grid-Datenanalyse für alle Benutzer der CMS Daten zu ermöglichen.

Aus den Resultaten dieser Arbeit können verschiedene wichtige Schlussfolgerungen gezogen werden. Die in [2] und [1] publizierten Ergebnisse wurden bestätigt, und es wurde eine Gegenprobe mittels einer Neuimplementierung und Optimierung des Quellcodes erreicht. Darüber hinaus wurden einige signifikante Verbesserungen der Analyse erzielt. Diese Verbesserungen bestanden in der Optimierung kinematischer Schnitte und Ereignisselektionen, in den bereits angesprochenen Weiterentwicklungen der b-Quark Jet Identifikationsalgorithmen sowie in Ergänzungen der zur Verfügung stehenden Menge an simulierten Ereignissen, um die statistische Zuverlässigkeit der Ergebnisse zu erhöhen. Im Vergleich zu den Resultaten in [2] konnte die Signifikanz im semileptonischen Kanal um etwa 10% erhöht werden, während die Reinheit um mehr als 90% gesteigert werden konnte. Da die Reinheit, sobald systematische Fehler berücksichtigt werden, der entscheidende Faktor ist, konnte auch das Resultat nach Einbeziehung dieser Fehler verbessert werden. Jedoch ändert sich die Aussage in Bezug auf die Beobachtbarkeit nicht wesentlich, sofern Gauß-verteilte Fehler angenommen werden. In einem neu optimistischeren Fehlermodell konnte eine Verbesserung der Aussage bezüglich der Beobachtbarkeit erzielt werden. Die Signifikanz kann Werte von $\sigma > 3$ erreichen, für eine hypothetische Higgs Boson Masse von $m_H = 115\ \mathrm{GeV}/c^2$ und für eine Dauer der Datennahme von drei Jahren bei einer Luminosität von $2 \cdot 10^{33} \mathrm{cm}^{-2}\mathrm{s}^{-1}$, entprechend einer integrierten Luminosität von 60 fb^{-1}. Dazu sind jedoch genaue Kenntnisse über die erwarteten Untergrundraten erforderlich. Das zentrale Resultat ist in Abbildung 1 dargestellt, welche das rekonstruierte

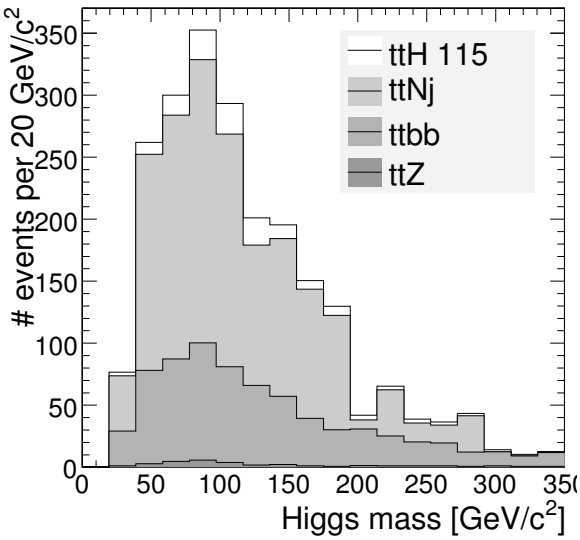

Abbildung 1: Invariantes Massenspektrum des Higgs Bosons nach Optimierung der Analyse im semileptonischen Zerfallskanal. Gezeigt sind alle relevanten Untergrundprozesse, $t\bar{t}Z$, $t\bar{t}b\bar{b}$ und $t\bar{t}N\mathrm{j}$ (in grau) sowie die Signalereignisse (in weiß), welche im Histogramm stufenweise aufsummiert dargestellt sind. Die hypothetische Masse des Higgs Bosons beträgt $m_H = 115$ GeV/c^2. Die Ergebnisse sind entsprechend einer integrierten Luminosität von 60 fb^{-1} berechnet.

Spektrum der invarianten Masse des Higgs Bosons für Signal- und Untergrundprozesse zeigt.

Eine wichtige Erkenntnis, die der vollständigen, realistischen Simulation und Rekonstruktion der Ereignisse zu verdanken ist, besteht in der Neubeurteilung des Einflusses von $t\bar{t}$ plus Jets Untergründen. Diese Prozesse wurden in bisherigen Studien unterschätzt, da die Unterdrückung dieser Untergründe nur durch b-Jet Identifikation möglich ist, die in früheren Studien in der vollständigen Detektorsimulation nicht zur Verfügung stand. Die Mehrheit der fehlidentifizierten $t\bar{t}$ + Jets Ereignisse besteht aus Gluonen, die in b- oder c-Quarks aufspalten und daher nicht unterdrückt werden können, sowie aus c-Quark Jets, die aus Zerfällen von W Bosonen stammen. Eine Berechnung der Effizienz der Untergrundunterdrückung muss diese Effekte berücksichtigen.

Offensichtlich muss die Entdeckung des $H \to b\bar{b}$ Zerfalls als große Herausforderung betrachtet werden, die nur auf längerer Zeitskala nach einigen Jahren des Detektorbetriebes zu bewältigen ist. Die theoretischen Unsicherheiten der Wirkungsquerschnitte der wichtigsten Untergrundprozesse haben einen sehr großen Einfluss auf die Endresultate. Eine vielversprechende Weiterführung der Analyse besteht in der Entwicklung von Methoden zur Messung der Untergrundraten aus echten Daten.

Abschließend kann gefolgert werden, dass der $t\bar{t}H$, $H \to b\bar{b}$ Prozess messbar sein wird, jedoch mit hoher Wahrscheinlichkeit erst nach einer vorangehenden Entdeckung des Higgs Bosons und einer Bestimmung seiner Masse. Auch wenn dieser Kanal nicht als Entdeckungskanal geeignet ist, handelt es sich hierbei trotzdem um einen wichtigen Beitrag in Form einer komplementären Messung zur Bestätigung des Standardmodells. Insbesondere die Messung der kombinierten top-Higgs, Higgs-bottom Yukawa-Kopplung wird durch die Bestimmung des Wirkungsquerschnittes des Signalprozesses ermöglicht.

Search for $H \to b\bar{b}$ in Association with a $t\bar{t}$ Pair in Proton-Proton Collisions at $\sqrt{s} = 14$ TeV

PhD Thesis
Faculty for Physics
University of Karlsruhe (TH)

by

Alexander Schmidt

Supervisors: Prof. Dr. G. Quast and Prof. Dr. Th. Müller
Institut für Experimentelle Kernphysik
University of Karlsruhe (TH)

Contents

Chapter 1

Introduction

Particle physics is the branch of physics that deals with the elementary constituents of matter and radiation and the interactions between them. Therefore, it can be considered one of the most fundamental topics of science. The philosophical concept of atomism was discussed as early as in the 6th century BC by Greek philosophers like Democritus. It was not until the year 1897, however, that the first elementary particle, the electron, was discovered experimentally in cathode rays by J. J. Thomson. Another milestone was the discovery of the atomic nucleus in an experiment conducted by Ernest Rutherford in the year 1909. In this experiment, the deflection of alpha particle rays directed on a thin gold foil was measured. The scattering angles were sometimes larger than 90 degrees leading to the conclusion that the atom contains a small positively charged nucleus.

The establishment of nuclear physics and quantum mechanics followed in the twentieth century leading to the discovery of a large variety of particles using scattering experiments in the 50s and 60s. This "particle zoo" was confusing at first, but all these particles could be reduced to combinations of a small number of elementary constituents in the Standard Model developed in the 70s. Especially the realization that the structure of the elementary components and their interactions follow basic mathematical principles can be understood as one of the most outstanding discoveries of the twentieth century. An important role in the mathematical description is attributed to local gauge theories which deduce the existence of exchange bosons from symmetry principles. According to the current understanding, the Standard Model is based upon a $\mathrm{SU}(3) \times \mathrm{SU}(2) \times \mathrm{U}(1)$ symmetry which predicts the existence of eight gluons of the strong nuclear interaction and four bosons of the electroweak interaction, the photon and the massive $\mathrm{W}^{\pm}$ and Z bosons.

The masses of the electroweak vector bosons play a very special role in this context, because they cannot be constructed from mass terms in the Lagrange density. A promising approach towards an explanation is the concept of spontaneous breaking of a local gauge symmetry, the so-called Higgs mechanism, introduced by Peter Higgs in the year 1964. This mechanism is connected to the existence of a scalar boson, the Higgs boson. The search for this Higgs boson is one of the major challenges of today's high energy physics. This thesis is also dedicated to this search. Its discovery or non-discovery represents an important step towards a completion of our knowledge about elementary processes in nature. A more detailed coverage of gauge theories and the Higgs mechanism is given in Chapter 3 of this thesis.

The next generation particle accelerator LHC ("Large Hadron Collider") and its associated particle detectors, which are described in Chapter 2, have a very high potential to answer

the question about the existence of the Higgs boson. These experiments, which are carried out by large international collaborations, are expected to provide first results in 2008. But not only the physics of the Higgs boson will be studied at these experiments. Other theories beyond the Standard Model, like Supersymmetry or extra dimensions, will be investigated as well. Furthermore, precision measurements in the Standard Model will be very important in order to obtain complementary measurements to preceding experiments and to find deviations from predictions of the Standard Model indicating new physics. Also heavy ion physics will be studied at the LHC.

The main topic of this thesis is the determination of the discovery potential of the Higgs boson in the decay channel $H \rightarrow b\bar{b}$ with the CMS detector. The exclusion limit of the Higgs mass has been determined to 114.4 GeV/c^2 by the experiments at the Large Electron-Positron Collider (LEP). The mass range just above this exclusion limit is very interesting because constraints from experiments at Tevatron and LEP indicate a low Higgs mass [3]. According to predictions within the Standard Model, the masses of the W boson and the top quark are connected with the Higgs boson mass through radiative loop corrections to the gauge boson propagators [4]. Figure 1.1 shows the dependency of the allowed Higgs masses on m_W and m_t as well as the current measurements. Therefore, a direct measurement of the Higgs boson mass represents an important consistency check of the Standard Model. More details about

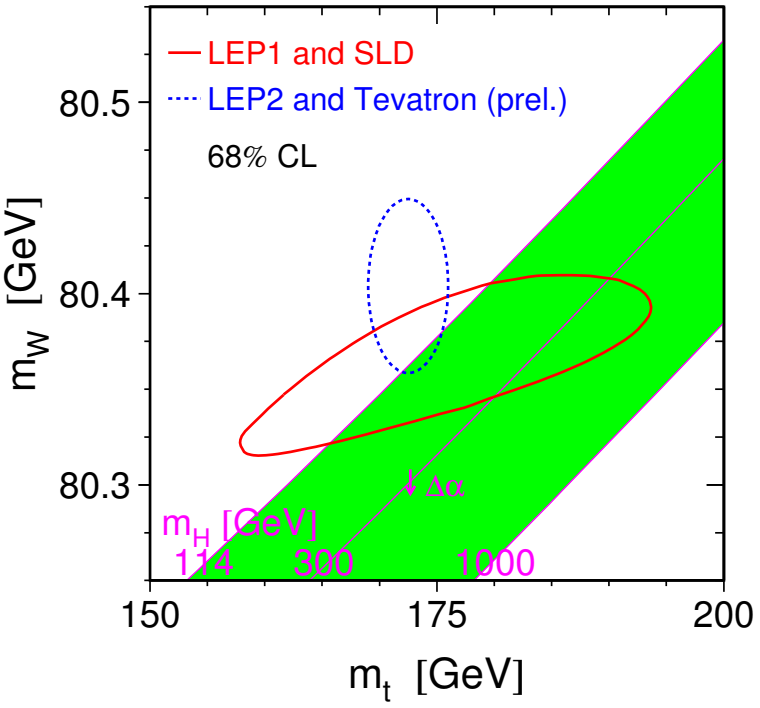

Figure 1.1: Comparison of indirect measurements (solid contour) of m_W and m_t and direct measurements (dashed contour). The 68% confidence levels are plotted in both cases. Also shown is the relationship of the masses as a function of the Higgs mass. [3]

the current status of the experimental search for the Higgs boson is given in Chapter 3.

At the LHC there are many production and decay processes of Higgs bosons. The decay channel $H \rightarrow b\bar{b}$ has the highest branching ratio in the very low mass region up to about 130 GeV/c^2. For slightly higher masses, the $H \rightarrow WW$ and $H \rightarrow ZZ$ decay modes start to contribute significantly until they are the dominant channels at $m_H \approx 160$ GeV/c^2. Due to the large abundance of other sources of b-quarks at the LHC, especially $g \rightarrow b\bar{b}$, the search for $H \rightarrow b\bar{b}$ has to be carried out in association with top quark production. This particular production mode has a reasonable cross section compared to background cross sections, which is discussed in Chapter 5 of this thesis.

The full analysis of $t\bar{t}H$ with $H \rightarrow b\bar{b}$ has been implemented in terms of a full simulation of the CMS detector and realistic reconstruction algorithms have been applied. This specific channel has been selected as "benchmark" channel in Volume 2 of the Physics Technical Design Report (PTDR) [1] because of the high demands on track reconstruction and b-flavour tagging performances. Therefore, a large amount of effort has been invested in the development and improvement of the analysis and reconstruction tools presented in Chapter 4. Especially the algorithms for b-quark identification have been studied and optimized since these algorithms are the most powerful component of the $t\bar{t}H$ analysis considering the presence of four b-quark jets.

An important advancement of the $t\bar{t}H$ analysis compared to preceding studies is the estimation of the impact of systematic uncertainties due to various sources like the energy scale of jets or b-jet tagging efficiencies. The analysis methods, the results and the impact of these systematic errors are presented in Chapter 5.

Finally, a short summary of the topics of this thesis, particularly the main results and the conclusions, is given in Chapter 6.

Chapter 2

The LHC and the CMS Experiment

At CERN, the "European Organization for Particle Physics Research" in Geneva, Switzerland, a new hadron collider experiment (LHC[1]) is under construction. It is a proton-proton collider that reaches a center of mass energy of $\sqrt{s} = 14$ TeV. Particle detectors are placed at four interaction points. Two of these detectors (ALICE[2] and LHCb[3]) are designed for special purposes (heavy ion and b-physics), while ATLAS[4] and CMS[5] are general purpose detectors. The LHC machine and one of the detectors, the CMS detector, are desribed in the following.

2.1 The Large Hadron Collider

The LHC is being installed in the 27 km long LEP[6] tunnel. Figure 2.1 shows the geographical situation and the location of the four experiments at the LHC. The proton beams circle in opposite directions in two separate beamlines that are filled with 2835 bunches of 10^{11} particles. These bunches are formed in the 26 GeV Proton Synchrotron (PS), which was CERN's first major particle accelerator built in 1959 and which is being reused for the purpose of forming the correct spacing of 25 ns between the bunches. In a next step, the beam is accelerated to 450 GeV in the Super Proton Synchrotron (SPS) which was built in 1976 and which was used in the beginning of the 1980s as proton-antiproton collider for the UA1 and UA2 experiments leading to the discovery of the W and Z bosons, earning the Nobel Price for Carlo Rubbia and Simon van der Meer. Subsequently, the beam is transferred to the LHC and accelerated to 7 TeV.

The commissioning of the LHC machine is planned to start at the end of the year 2007. At the beginning, the machine will start running with a few bunches in single beam operation. In 2008 this will be followed by a low luminosity pilot physics run in which a small number of 43 bunches with only 10^{10} protons will continue the evolution of the machine. The proton density will continuously increase, while the spacing will be decreased until the nominal numbers will

[1] Large Hadron Collider
[2] A Large Ion Collider Experiment
[3] Large Hadron Collider beauty experiment
[4] A Toroidal LHC ApparatuS
[5] Compact Muon Solenoid
[6] Large Electron Positron collider

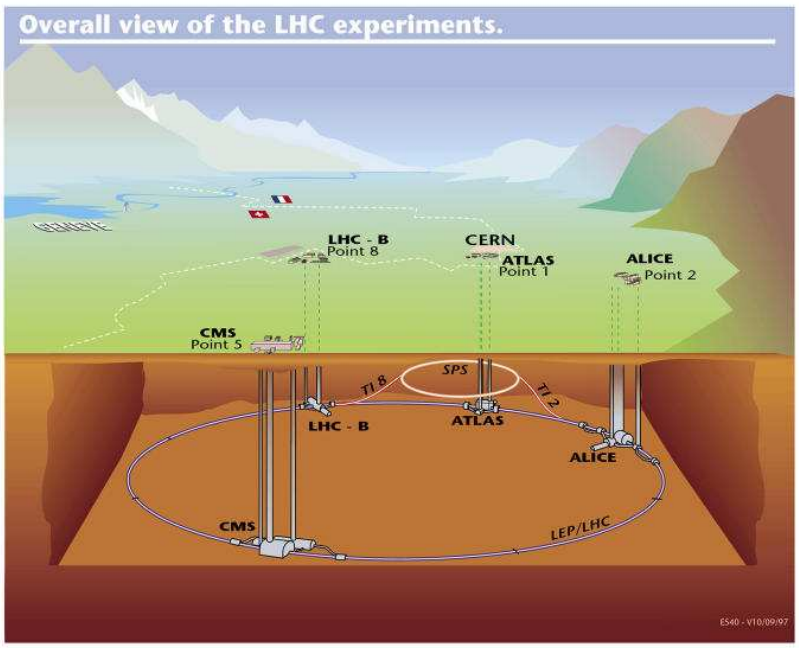

Figure 2.1: Geographical situation at the LHC.

be reached.

The design value of the luminosity at the interaction points is $L = 10^{34}\text{cm}^{-2}\text{s}^{-1}$ during the so called "high luminosity" runs and $L = 10^{33}\text{cm}^{-2}\text{s}^{-1}$ during the "low luminosity" phases. To focus the beams and to force them into the right trajectories, about 1232 superconducting niobium-titanium magnets are installed. These magnets produce fields up to 8.36 Tesla.

In one year of operation, the LHC will collect an integrated luminosity of 10 fb^{-1}, but this value is likely to be much less than 5 fb^{-1} in the first year, since the machine development will probably encounter unforeseen inefficiencies.

2.2 The Compact Muon Solenoid

The CMS detector project is one of the largest scientific collaborations in history. More than 2000 people from all over the world are working for CMS. The construction and commissioning of the detector is therefore not only a technical but also an administrative challenge. Despite all difficulties, the installation of the detector is progressing well, as the very successful completion of the Magnet Test and Cosmic Challenge (MTCC) in the year 2006 has shown. During the MTCC, the muon system and parts of the tracking system have been commissioned and have been used in order to reconstruct muons from cosmic rays. A central role during this MTCC was the cooling and subsequent startup of the magnet, a superconducting solenoid which sits in the heart of the detector. The solenoid is 13 m long, has an inner diameter of 5.9 m and reaches a magnetic field of 4 T. It is the largest superconducting solenoid ever built. During the MTCC the magnet has proven to be operable in reliable conditions and a mapping of its magnetic field has been performed. A profile view of the CMS detector, showing the position of the magnet and muon systems is displayed in Figure 2.2.

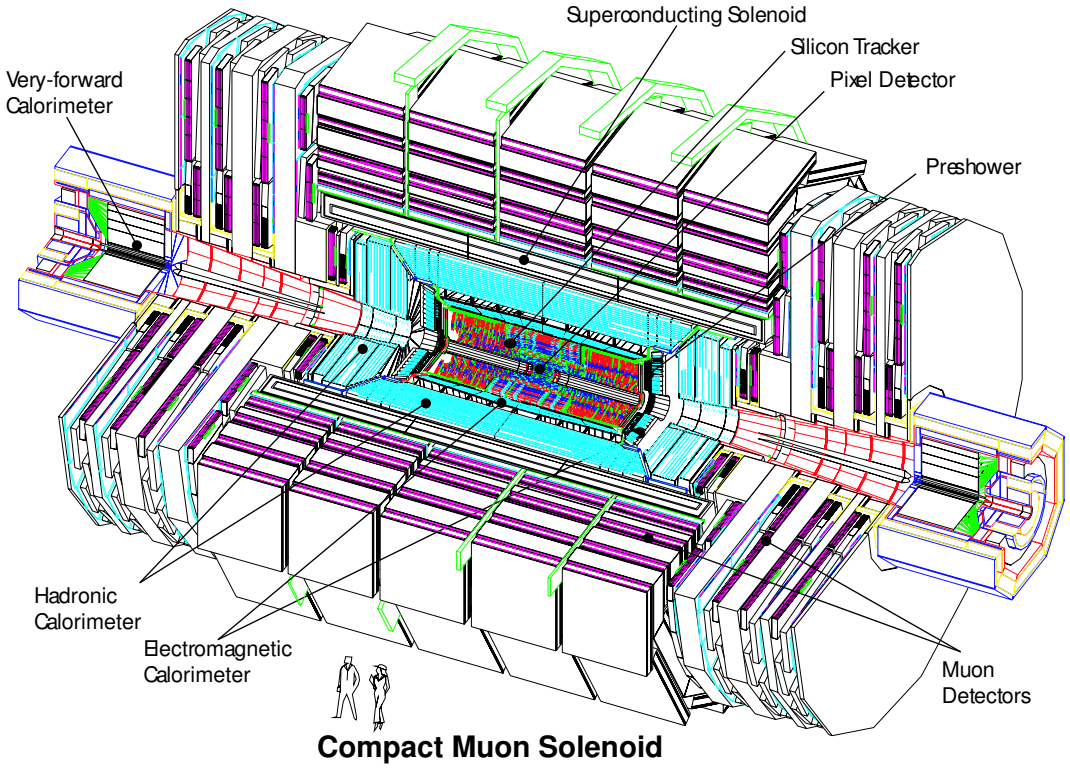

Figure 2.2: Profile view of the CMS detector.

The design of the solenoid has been chosen in order to create a high magnetic field which is necessary to achieve a sufficient bending of the charged particle tracks in order to get a good momentum resolution in the compact muon spectrometer. The return yoke is interleaved with four muon "stations" consisting of aluminium Drift Tube chambers (DT) in the barrel and Cathode Strip Chambers (CSC) in the endcaps. In both cases, Resistive Plate Chambers (RPC) are placed sandwich-like in between the DTs and CSCs, respectively.

The DT chambers consist of 12 layers of tubes. Eight of these layers measure the r, ϕ coordinate in the bending plane using wires parallel to the beam line. The remaining four layers measure the z coordinate. The CSCs are trapezoidal in shape and consist of 6 gas gaps, each gap having a plane of radial cathode strips and a plane of anode wires running almost perpendicularly to the strips. An RPC detector consists of a double-gap bakelite chamber, operating in avalanche mode. The gaps have a 2 mm width.

The DT and CSC detectors provide a good position and therefore momentum measurement, while the RPC detector has a good time resolution. Therefore, both detectors types are combined in order to achieve an improved overall measurement. This emphasis on the layout of the muon system in CMS has its origin in the importance of muons in all kinds of physics analyses, especially in the search for Higgs bosons and physics beyond the Standard Model.

An exhaustive description of the magnet and muon detectors can be found in [5] and [6], respectively. Besides the muon system and the solenoid which are responsible for the naming of CMS, the tracking system is the next important key ingredient.

2.2.1 The Tracking System

The innermost part of the CMS detector is a silicon pixel system which provides precise three dimensional position measurements of charged particles passing through the sensitive volumes. The pixel detectors in the barrel region has three layers at the distances 4.3 cm, 7.2 cm and 10 to 11 cm to the beam axis, covering a pseudorapitity up to $|\eta| < 2.2$. Only two layers are integrated in the endcap at z positions of 32.5 cm and 46.5 cm which increase the $|\eta|$ covering to $|\eta| < 2.5$.

There are in total 50 million pixels at a pitch of 100 μm $\times$ 150 μm yielding a spacial resolution of 15 μm exploiting the shape of the charge distribution on the sensor surface [7]. The pixels are covering a total area of about 1 m^2.

The following part of the tracking system consists of silicon strip detector modules which are arranged in ten concentric layers in the barrel. Four of these layers belong to the Inner Barrel (TIB) while six layers constitute the Outer Barrel (TOB). The barrel part covers a radial distance of 20 to 110 cm and a $|z|$ distance of 280 cm. The strips are oriented parallel to the beam axis to allow a precise azimuthal measurement. The two inner layers of TIB and TOB are double-layered with a stereo angle of 100 mrad and therefore allows a three diemensional measurement. There are also nine End Cap disks (TEC) located between $z = 120$ cm and 280 cm with a radial orientation of the strips and two double-layers at the first and last disk.

The methods for track reconstruction and the resulting resolutions are summarized in Section 4.2.5. More information about the tracking system is available in [8].

2.2.2 The Electromagnetic Calorimeter

The Electromagnetic Calorimeter (ECAL) consists of ~80000 lead tungstate ($PbWO_4$) crystals to perform the accurate measurement of electron and photon energies and their directions of flight. $PbWO_4$ crystals are chosen mainly because of their short radiation length of $X_0 = 0.89$ cm, due to the high density 8.2 g/cm^3, and small Molière radius of $R_M = 2.2$ cm. This allows a compact ECAL design with narrow showers. The crystals are about 23 cm long, corresponding to almost 26 X_0, thereby containing more than 99% of the shower energy. A second advantage of using $PbWO_4$ is that the scintillating process is fast: 80% of the light is emitted within 20 ns, matching the LHC bunch crossing time of 25 ns.

The lateral granularity of the ECAL is $\Delta\eta \times \Delta\phi = 0.0175 \times 0.0175$, corresponding to a crystal front face of about 22 x 22 mm^2. The fine lateral size is required because of the need for a good π^0 rejection, to avoid that two photons from energetic π^0s, which are emitted close to each other are reconstructed as a single photon. All the crystals are mounted in a projective geometry with a 3 degree tilt in η and ϕ with respect to the mean position of the primary interaction vertex in order to limit the effects of the inter-crystal gaps. The barrel section (EB) has an inner radius of 129 cm. It is structured as 36 identical "supermodules," each covering half the barrel length and corresponding to a pseudorapidity interval of $0 < |\eta| < 1.479$. The endcaps (EE) are located at a distance of 314 cm from the vertex and are covering a pseudorapidity range of $1.479 < |\eta| < 3.0$. Further details can be found in the ECAL TDR [9].

2.2.3 The Hadron Calorimeter

The Hadron Calorimeter (HCAL) is realized as a copper alloy calorimeter and allows the measurement of the energies of hadrons, that are not stopped in or before the ECAL. The HCAL

surrounds the ECAL completely and is used in conjunction with the latter for energy measurements. It provides a hermetic coverage to allow missing transverse energy measurements, therefore it is separated in a barrel part ($|\eta| < 1.3$) and two endcaps ($1.3 < |\eta| < 3$), including a Forward calorimeter, situated 6 m down the beam pipe which increases the hermeticity to $|\eta| < 5$.

The active part of the HCAL are plastic scintillators with wavelength shifting fibre readout. Layers of these tiles alternate with layers of 5 cm thick brass absorber to form the sampling calorimeter structure. The tiles are arranged in projective towers with fine granularity to provide good di-jet separation and mass resolution. In the barrel the calorimeter has a thickness of 79 cm corresponding to five nuclear interaction lengths. This is not enough for a full shower containment leading to low energy tails in the hadron distributions and mismeasurements of missing transverse energy. Therefore, an Outer Hadron calorimeter (HO) is placed outside the solenoid which consists of one scintillator layer. The HCAL is described in detail in [10].

2.2.4 Level-1 Trigger

In order to reduce the event rate of 40 MHz to about 100 Hz, two trigger levels are realized in CMS. The first one is a hardware based Level-1 Trigger which sits directly at the detector and which is shortly described in the following. The second trigger level is the High-Level Trigger (HLT), which is software based and is described in Section 4.2.1.

At the first trigger level, all information about the event is preserved and a decision about the acceptance is made in negligible deadtime using a subset of the available information. The information used at Level-1 involves calorimeter and muon system. The muon trigger is organized into subsystems for each muon detector type: DT, CSC and RPC. The information from these three triggers is combined in the global muon trigger. Afterwards, the information from the global muon trigger is sent to the Level-1 global trigger, where the muon information is combined with calorimter information. Based on objects like photons, electrons, jets and muons and after employing sums of E_t and p_t thresholds, the trigger decides in less than 1 μs if the event is accepted or not. The maximum design trigger rate of 100 kHz corresponds to a minimal rejection rate of 10^4. The Level-1 Trigger project is described in [11].

Chapter 3

The Higgs Boson in the Standard Model

In the year 1930, the first evidence for a "weak" interaction was found in the nuclear beta decay. The observed energy spectrum of the electron was continuous in contrast to nuclear γ-emission. If a two-body decay is assumed, then this observation contradicts energy and momentum conservation. To resolve this problem, Wolfgang Pauli proposed an additional neutral particle, the neutrino, that is emitted with the electron. The first approach to describe this process was a four-fermion-point interaction. Today this is properly explained by virtual W boson exchange as in Figure 3.1.

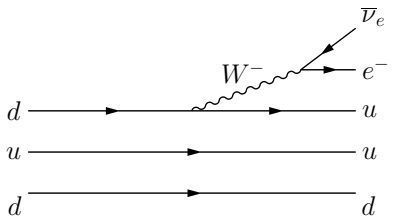

Figure 3.1: Feynman diagram of the neutron decay. Two of the quarks are "spectators" and do not participate directly.

In 1973, experiments at CERN revealed the existence of uncharged weak interactions. The first reaction observed was $\overline{\nu}_\mu + e \rightarrow \overline{\nu}_\mu + e$. At this time, the theory was already established and the reaction was supposed to be mediated by an uncharged partner of the W, the Z boson. Even though the experimental confirmation of the presence of the electroweak gauge bosons became quickly indisputable, the question of how these bosons acquire their mass is still not answered finally. One of the most promising theories answering this question is the Higgs mechanism.

Based on the theory of electroweak interactions, which is being developed in the following sections, the necessity for a Higgs mechanism will be discussed followed by a short overview of the experimental searches for the Higgs boson. Much of the presentation of the material in the subsequent sections has been inspired by [12, 13].

3.1 Parity Non-Conservation and V–A Theory

One of the properties of the weak interaction contradicting intuition is parity non-conservation. This was first observed by Wu *et al.* in nuclear beta decay of polarized ^{60}Co nuclei in a magnetic field. The relative electron intensities along and against the field direction show a forward-backward asymmetry, which implies that the reaction violates parity conservation.

It was found that neutrinos occur in left-handed helicity states only and anti-neutrinos in right-handed states[1]. The charged leptons produced in weak interactions are left-handed with a degree of polarization of $\beta = -v/c$. This behavior can be explained with V–A theory (vector minus axial-vector).

In this theory any Dirac-spinor u can be divided into two chirality components:

$$u = \frac{1}{2}(1+\gamma^5)u + \frac{1}{2}(1-\gamma^5)u = u_R + u_L$$

$P_R = \frac{1}{2}(1+\gamma^5)$ and $P_L = \frac{1}{2}(1-\gamma^5)$ are the chirality-projection-operators and the spinor u_R is called right-handed and u_L left-handed. For $E \gg m$, P_L and P_R become the projection-operators for negative and positive helicity.

In the according Feynman rules for the calculation of the invariant amplitude, only the appropriate chirality components are to be used. For example, at the $\mu\nu$-vertex, the associated factor becomes

$$\overline{u}(\mu)\gamma_\mu u_L(\nu) = \overline{u}(\mu)\gamma_\mu \frac{1-\gamma^5}{2} u(\nu) \ .$$

Now there are two terms at the neutrino-muon-vertex:

- The vector-current
$$V^\mu = \overline{\psi}(\mu)\gamma^\mu\psi(\nu) \ ,$$
that transforms like a four-vector.

- The axial-vector-current
$$A^\mu = \overline{\psi}(\mu)\gamma^\mu\gamma^5\psi(\nu) \ ,$$
that transforms like an axial-vector. This vector behaves like a four-vector under Lorentz-transformations, but it keeps its sign under parity transformation.

The V–A construct therefore violates parity conservation. Because of $P_L^2 = P_L$ the matrix element contains only the left-handed components of the spinors and the right-handed components of the anti-spinors:

$$\overline{u}_e\gamma_\mu \frac{1-\gamma^5}{2} u_\nu = \overline{u}_e\gamma_\mu \left(\frac{1-\gamma^5}{2}\right)^2 u_\nu = \overline{(u_e)}_L\gamma_\mu(u_\nu)_L \ .$$

This means that only left-handed components participate in this kind of weak interactions.

However, the coupling of the Z-boson is not so simple. Instead of the purely V-A vertex factor of the W boson, it is necessary to use a mixed form:

$$\gamma^\mu(c_V^f - c_A^f\gamma^5) \ ,$$

where the coefficients depend on the particular quark or lepton (f) involved. These numbers and also the coupling constants and masses of the vector-bosons are determined by one fundamental parameter, the "Weinberg angle" or "weak mixing angle". This angle can be calculated from electroweak unification, discussed in Section 3.3.

[1] The discovery of neutrino-oscillations shows that this is an approximation.

3.2 U(1) Local Gauge Invariance and QED

In classical electrodynamics a global gauge transformation of the vector field $A^\mu = (\phi, \vec{A})$ with

$$A^\mu \to A'^\mu = A^\mu - \partial^\mu \chi$$

leaves the fields $\vec{E}$ and $\vec{B}$ invariant since

$$\vec{B} = \nabla \times \vec{A} \ , \quad \vec{E} = -\nabla\phi - \frac{\partial \vec{A}}{\partial t} \ .$$

If the principle of gauge invariance is applied to quantum mechanics, the combined transformation turns out to be

$$\begin{aligned} \vec{A} &\to \vec{A}' = \vec{A} + \nabla\chi \ , \\ \phi &\to \phi' = \phi - \frac{\partial \chi}{\partial t} \ , \\ \psi &\to \psi' = e^{iq\chi}\psi \ , \end{aligned}$$

to fulfill the Schrödinger equation[2]. If the principle is extended to *local* invariance, one gets the result that this leads to the interaction of particles with fields. To see this, we start with the Lagrangian of a free Dirac particle

$$\mathcal{L} = i\overline{\psi}\gamma^\mu \partial_\mu \psi - m\overline{\psi}\psi \tag{3.1}$$

which is not invariant under local phase transformations of the form

$$\psi(x) \to e^{i\alpha(x)}\psi(x) \ .$$

If $\overline{\psi}$ transforms like

$$\overline{\psi} \to e^{-i\alpha(x)}\overline{\psi}$$

then the last term in $\mathcal{L}$ is invariant, but not the derivative:

$$\partial_\mu \psi \to e^{i\alpha(x)}\partial_\mu \psi + ie^{i\alpha(x)}\psi\partial_\mu \alpha \tag{3.2}$$

The $\partial_\mu \alpha$ term is the cause for the break of the invariance. If one postulates local gauge invariance, then a modified derivative D_μ, that transforms like ψ itself, is necessary.

$$D_\mu \psi \to e^{i\alpha(x)} D_\mu \psi \ .$$

If this covariant derivative is used instead of ∂_μ in (3.1), the Lagrangian becomes invariant. Now, a derivative that cancels the additional $\partial_\mu \alpha$ term in (3.2) has to be found. To do this, it is necessary to introduce a vector field A_μ with appropriate transformation properties:

$$D_\mu \equiv \partial_\mu - ieA_\mu \ ,$$

[2]The Schrödinger equation of a particle with the charge q in an electromagnetic field is

$$\left(\frac{1}{2m}(-i\nabla - q\vec{A})^2 + q\phi\right)\psi(t,\vec{x}) = i\frac{\partial \psi}{\partial t}$$

with

$$A_\mu \to A_\mu + \frac{1}{e}\partial_\mu \alpha \ .$$

If the new field is regarded as the photon field and if an invariant term corresponding to its kinetic energy is added, we get the Lagrangian of QED:

$$\mathcal{L} = \overline{\psi}(i\gamma^\mu \partial_\mu - m)\psi + e\overline{\psi}\gamma^\mu A_\mu \psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}$$

This means that the interacting field theory QED is deduced by postulating local gauge invariance on the free fermion Lagrangian.

3.3 $\mathrm{SU}(2)_L \times \mathrm{U}(1)$-Symmetry

The attempt to extend the U(1) local gauge invariance to SU(2) leads to electroweak unification, but with lots of additional requirements, e.g. the masses of the bosons have to be included. This is outlined in the following discussion.

The particles that experience electroweak transitions by emission of field bosons can be arranged in multiplets of a "weak isospin" in analogy to the spin-formalism. The left-handed fermions constitute doublets with $I = 1/2$:

$$\begin{array}{lccc} & & & I_3 \\ \text{Leptons:} & \begin{pmatrix} \nu_e \\ e^- \end{pmatrix}_L & \begin{pmatrix} \nu_\mu \\ \mu^- \end{pmatrix}_L & \begin{pmatrix} \nu_\tau \\ \tau^- \end{pmatrix}_L \quad \begin{array}{c} 1/2 \\ -1/2 \end{array} \end{array}$$

$$\begin{array}{lccc} & & & I_3 \\ \text{Quarks:} & \begin{pmatrix} u \\ d' \end{pmatrix}_L & \begin{pmatrix} c \\ s' \end{pmatrix}_L & \begin{pmatrix} t \\ b' \end{pmatrix}_L \quad \begin{array}{c} 1/2 \\ -1/2 \end{array} \end{array}$$

The right-handed leptons and quarks do not couple to the charged weak currents, therefore they are arranged in singlets:

$$I = 0: \quad e_R^-,\ \mu_R^-,\ \tau_R^-,\ u_R,\ d_R,\ s_R,\ c_R,\ b_R,\ t_R$$

The Dirac-wave-function of the left-handed leptons can then be expressed as a product of a left-handed Dirac-spinor $\psi_L(t, \vec{x})$ and a weak isospinor χ:

$$\nu_L = \psi_L(t,\vec{x})\begin{pmatrix}1\\0\end{pmatrix},\ e_L = \psi_L(t,\vec{x})\begin{pmatrix}0\\1\end{pmatrix} \ .$$

The transition $e^- \to \nu_e$ proceeds by emission of a W^--boson and is mediated by the "step-up" operator τ_+, the transition $\nu_e \to e^-$ by the "step-down" operator τ_-. The matrices $\tau_\pm$ are linear combinations of the first two Pauli spin matrices:

$$\tau_\pm = \frac{1}{2}(\tau_1 \pm i\tau_2)$$

There should be another operator τ_3, that leaves I_3 unchanged, which is later identified with the neutral weak current.

In analogy to the U(1) case, where q is the coupling strength and α the transformation angle, a phase transformation in the weak isospin space is defined. The $\mathrm{SU}(2)_L$ group describes transformations of the left-handed weak isospin multiplets, e.g.

$$\begin{pmatrix} \nu_e \\ e^- \end{pmatrix}'_L = \exp\left(i\frac{g}{2}\vec{\tau}\cdot\vec{\beta}(x)\right)\begin{pmatrix} \nu_e \\ e^- \end{pmatrix}_L ,$$

where g is the coupling strength and $\vec{\tau}$ represents the pauli matrices (τ_1, τ_2, τ_3). Now there are three angles β_j.

To get invariance under a local transformation it is necessary to introduce a triplet of vector-fields $W_1^\mu, W_2^\mu, W_3^\mu$ for the $\mathrm{SU}(2)_L$ group. The covariant derivative for (ν_e, e_L^-), (ν_μ, μ_L^-), (ν_τ, τ_L^-) is

$$D^\mu = \partial^\mu + i\frac{g}{2}\vec{\tau}\cdot\vec{W}^\mu \ .$$

The right-handed fermions have to be included as well since the neutral electroweak interaction couples right-handed states. This is accomplished by introducing the "weak hypercharge" Y:

$$Q = I_3 + \frac{1}{2}Y \ .$$

Q is the charge and I_3 the third component of weak isospin. The associated weak hypercharge current then involves left-handed and right-handed chirality states. The weak hypercharge can be associated with phase transformations as well:

$$\begin{aligned} \begin{pmatrix} \nu_e \\ e^- \end{pmatrix}'_L &= \exp\left(i\left(\frac{g'}{2}Y_L\right)\chi(x)\right)\begin{pmatrix} \nu_e \\ e^- \end{pmatrix}_L , \\ e'_R &= \exp\left(i\left(\frac{g'}{2}Y_R\right)\chi(x)\right)e_R \ . \end{aligned}$$

$\frac{g'}{2}Y$ is the coupling constant instead of the charge q in the electromagnetic case. These transformations form a U(1) group. A single vector field is necessary for local gauge invariance in the U(1) group as derived earlier. If this is combined with the three vector fields of the $\mathrm{SU}(2)_L$ group one gets the covariant derivative of $\mathrm{SU}(2)_L \times \mathrm{U}(1)$:

$$D^\mu = \partial^\mu + ig\vec{T}\cdot\vec{W}^\mu + i\frac{g'}{2}YB^\mu$$

For left-handed leptons it is

$$\vec{T} = \vec{\tau}/2 \ , \ Y = -1$$

and for right-handed leptons

$$\vec{T} = 0 \ , \ Y = -2 \ .$$

This means that for $(\nu_e, e_L^-), (\nu_\mu, \mu_L^-), (\nu_\tau, \tau_L^-)$

$$D^\mu = \partial^\mu + i\frac{g}{2}\vec{\tau}\cdot\vec{W}^\mu - i\frac{g'}{2}B^\mu$$

and for e_R^-, μ_R^-, τ_R^-

$$D^\mu = \partial^\mu - ig'B^\mu \ .$$

For the left-handed leptons this can also be expressed as

$$D^\mu = \partial^\mu + i\frac{g}{\sqrt{2}}\left(\tau_+ W^{(-)\mu} + \tau_- W^{(+)\mu}\right) + i\frac{g}{2}\tau_3 W_3^\mu - i\frac{g'}{2}B^\mu \ .$$

To get the transition matrix-elements, the covariant derivative has to be substituted in the Dirac-equation. For the process $e^- \to \nu_e$ we get

$$\mathcal{M} \propto -i\frac{g}{\sqrt{s}}\overline{\nu}_L\gamma_\mu\tau_+ e_L W^{(-)\mu} \ .$$

The transition $\nu_e \to \nu_e$ is mediated by

$$i\frac{g}{2}\tau_3 W_3^\mu \text{ and } -i\frac{g'}{2}B^\mu$$

and contributes to the matrix-element

$$-i\frac{g}{2}\overline{\nu}_L\gamma_\mu\tau_3\nu_L W_3^\mu \text{ and } +i\frac{g'}{2}\overline{\nu}_L\gamma_\nu\nu_L B^\mu \ .$$

The electromagnetic field A^μ can not be identified with W_3^μ or B^μ since the coupling to the neutrino does not disappear. It is possible to construct a linear combination of these two fields to get a vanishing coupling to the neutrino:

$$A^\mu = aW_3^\mu + bB^\mu \text{ with coupling } \sim a\left(-\frac{g}{2}\right) + b\frac{g'}{2} = 0$$

so it is

$$a = \frac{g'}{\sqrt{g^2+g'^2}} \ , \ b = \frac{g}{\sqrt{g^2+g'^2}} \ .$$

This defines the "weak mixing angle" or "Weinberg angle" θ_w:

$$\cos\theta_w = \frac{g}{\sqrt{g^2+g'^2}} \ , \ \sin\theta_w = \frac{g'}{\sqrt{g^2+g'^2}}$$

It follows that

$$A^\mu = B^\mu\cos\theta_w + W_3^\mu\sin\theta_w$$

If the field Z^μ of the neutral weak current is supposed to be orthogonal to A^μ it is

$$Z^\mu = -B^\mu\sin\theta_w + W_3^\mu\cos\theta_w \ .$$

The fundamental relation between the charge e and the coupling constants g and g' is

$$e = g'\cos\theta_w = g\sin\theta_w \ .$$

This follows if right-handed electrons that couple to B^μ only are considered:

$$ig'\overline{u}_R\gamma_\mu u_R B^\mu = ig'\cos\theta_w\overline{u}_R\gamma_\mu u_R A^\mu - ig'\sin\theta_w\overline{u}_R\gamma_\mu u_R Z^\mu$$

In QED the coupling of electrons is equal for right- and left-handed states:

$$ie\overline{u}\gamma_\mu u A^\mu$$

So one can see that $e = g'\cos\theta_w$. A similar argumentation leads to the coupling of the electron to the Z:

$$-\frac{ig}{2\cos\theta_w}\overline{u}\gamma_\mu(v_e - a_e\gamma^5)uZ^\mu \ ,$$

where the vector- and axial-vector couplings are

$$v_e = 2\sin^2\theta_w - 1/2 \ , \ a_e = -1/2$$

3.4 The Origin of Mass

In gauge theories, the interacting bosons are required to be massless. This is no problem for the photon and gluons, but if this is applied to weak interactions with massive bosons, we run into trouble. If mass terms of the form $M^2 W_\mu W^\mu$ are introduced into the Lagrangian, it is no longer gauge-invariant. One possibility to explain the masses of the particles is the introduction of a background field, the Higgs field, in analogy to the theory of superconduction.

3.4.1 Spontaneous Symmetry Breaking

The Lagrangian of a scalar field is for example

$$\mathcal{L} = \frac{1}{2}(\partial_\mu \phi)^2 - (\frac{1}{2}\mu^2\phi^2 + \frac{1}{4}\lambda\phi^4) \ ,$$

where only the first two terms in the expansion of the potential $V(\phi)$ are kept. If $\lambda > 0$ and $\mu^2 > 0$ the ground state is $\phi = 0$ and the Lagrangian is symmetric. But if $\mu^2 < 0$ the Lagrangian has a mass term with the wrong sign and two minima at

$$\phi = \pm\sqrt{-\mu^2/\lambda}$$

If the field is translated to $\sqrt{-\mu^2/\lambda}$ by

$$\phi(x) = \sqrt{-\mu^2/\lambda} + \eta(x)$$

it is

$$\mathcal{L}' = \frac{1}{2}(\partial_\mu \eta)^2 - \lambda v^2\eta^2 - \lambda v\eta^3 - \frac{1}{4}\lambda\eta^4 + const.$$

with $v = \sqrt{-\mu^2/\lambda}$. Now there is a mass term of the correct sign. The mass is

$$m_\mu = \sqrt{2\lambda v^2} = \sqrt{-2\mu^2}$$

and the higher order terms in η represent the interaction of the field with itself.

It is necessary to note that the Lagrangians $\mathcal{L}$ and $\mathcal{L}$' are equivalent. Surprisingly they yield different masses. The fact that perturbation theory is always expanded around the minimum of a potential resolves this ambiguity. The Feynman calculus is a perturbation theory and it would not converge if it would be expanded around $\phi = 0$ because this is not a stable minimum. This means that the mass was "generated" by a "spontaneous symmetry breaking" because the original reflection symmetry of the Lagrangian has been broken by the choice of of the ground state $\phi = v$.

3.4.2 Spontaneous Breaking of a Local SU(2) Gauge Symmetry

It is necessary to extend the results of the previous chapter to SU(2) gauge symmetry. The Lagrangian

$$\mathcal{L} = (\partial_\mu \phi)^\dagger(\partial^\mu \phi) - \mu^2\phi^\dagger\phi - \lambda(\phi^\dagger\phi)^2 \ ,$$

where ϕ is a SU(2) doublet of complex scalar fields

$$\phi = \begin{pmatrix}\phi_\alpha \\ \phi_\beta\end{pmatrix} = \sqrt{1/2}\begin{pmatrix}\phi_1 + i\phi_2 \\ \phi_3 + i\phi_4\end{pmatrix}$$

is invariant under global transformations of the form

$$\phi \to \phi' = e^{i\vec{\alpha}\vec{\tau}/2}\phi \ .$$

To extend the invariance to local transformations, ∂_μ has to be replaced by the covariant derivative

$$D_\mu = \partial_\mu + i\frac{g}{2}\vec{\tau}\cdot\vec{W}_\mu \ .$$

The invariant Lagrangian is then

$$\mathcal{L} = \left(\partial_\mu\phi + i\frac{g}{2}\vec{\tau}\cdot\vec{W}_\mu\phi\right)^\dagger\left(\partial^\mu\phi + i\frac{g}{2}\vec{\tau}\cdot\vec{W}^\mu\phi\right) - V(\phi) - \frac{1}{4}\vec{W}_{\mu\nu}\cdot\vec{W}^{\mu\nu} \ ,$$

with

$$V(\phi) = \mu^2\phi^\dagger\phi + \lambda(\phi^\dagger\phi)^2 \ .$$

Again, a kinetic energy term is added to the Lagrangian with

$$\vec{W}_{\mu\nu} = \partial_\mu\vec{W}_\nu - \partial_\nu\vec{W}_\mu - g\vec{W}_\mu\times\vec{W}_\nu \ .$$

If $\mu^2 < 0$ and $\lambda > 0$ one gets a potential $V(\phi)$ with a minimum at a finite value

$$\phi^\dagger\phi = |\phi|^2 = -\frac{\mu^2}{2\lambda} \ .$$

An arbitrary point in this minimum can be chosen and $\phi(x)$ can be expanded around this point, e.g.

$$\phi_1 = \phi_2 = \phi_4 = 0 \ , \ \phi_3^2 = -\frac{\mu^2}{\lambda} = v^2 \ .$$

The expansion

$$\phi(x) = \sqrt{\frac{1}{2}}\begin{pmatrix}0\\ v+h(x)\end{pmatrix}$$

is then substituted into the Lagrangian. This means that of the four scalar fields only one remains, the Higgs field $h(x)$. It is sufficient to substitute

$$\phi_0 = \sqrt{\frac{1}{2}}\begin{pmatrix}0\\ v\end{pmatrix}$$

into the Lagrangian. The effect of this procedure is the spontaneous breaking of the SU(2) symmetry because the expansion around the groud state hides the symmetry of the original Lagrangian. The mass can then be read off the new Lagrangian. The relevant term is

$$\left(i\frac{g}{2}\vec{\tau}\cdot\vec{W}_\mu\phi\right)^\dagger\left(i\frac{g}{2}\vec{\tau}\cdot\vec{W}^\mu\phi\right) = \frac{g^2v^2}{8}\left((W_\mu^1)^2 + (W_\mu^2)^2 + (W_\mu^3)^2\right)$$

and the mass is $M = \frac{1}{2}gv$.

It should be noted that in general, a procedure like this leads to the existence of massles Goldstone bosons. But these bosons can be gauged leading to the longitudinal polarizations of the massive vector bosons.

3.4.3 Masses of the Gauge Bosons

This formalism has to be applied to the weak interaction so that $W^{\pm}$ and Z become massive and the photon remains massless. An SU(2) × U(1) gauge invariant Lagrangian $\mathcal{L}_2$ has to be added to the Lagrangian that describes the electroweak interaction:

$$\mathcal{L}_2 = \left| \left(i\partial_\mu - g\vec{T}\cdot\vec{W}_\mu - g'\frac{Y}{2}B_\mu \right)\phi \right|^2 - V(\phi) \ ,$$

where $|\ |^2 = (\ )^\dagger(\ )$. The form of this Lagrangian has been discussed in the previous section. The "Weinberg-Salam model" now makes a choice for the fields so that the vacuum is invariant under U(1) transformations and the photon remains massless. Four fields are arranged in an isospin doublet with hypercharge $Y = 1$:

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}$$

with

$$\begin{aligned} \phi^+ &= (\phi_1 + i\phi_2)/\sqrt{2} \\ \phi^0 &= (\phi_3 + i\phi_4)/\sqrt{2} \ . \end{aligned}$$

The Higgs potential is chosen as in the previous section with $\mu^2 < 0$ and $\lambda > 0$ and the vacuum expectation value becomes

$$\phi_0 = \sqrt{\frac{1}{2}}\begin{pmatrix} 0 \\ v \end{pmatrix} \ .$$

ϕ_0 is then substituted into the Lagrangian $\mathcal{L}_2$ and the masses can be read off:

$$\left| \left(-i\frac{g}{2}\vec{\tau}\cdot\vec{W}_\mu - i\frac{g'}{2}B_\mu \right) \right|^2 = (\frac{1}{2}vg)^2 W_\mu^+ W^{-\mu} + \frac{1}{8}v^2 (W_\mu^3, B_\mu) \begin{pmatrix} g^2 & -gg' \\ -gg' & g'^2 \end{pmatrix} \begin{pmatrix} W^{3\mu} \\ B^\mu \end{pmatrix} \ .$$

The mass term $M_W^2 W^+ W^-$ yields

$$M_W = \frac{1}{2}vg \ .$$

The remaining term is

$$\frac{1}{8}v^2 \left(g^2 (W_\mu^3)^2 - 2gg' W_\mu^3 B^\mu + g'^2 B_\mu^2 \right) = \frac{1}{8}v^2 \left(gW_\mu^3 - g'B_\mu \right)^2 + 0 \left(g'W_\mu^3 + gB_\mu \right)^2 \ .$$

Since

$$\begin{aligned} A_\mu &= \frac{g'W_\mu^3 + gB_\mu}{\sqrt{g^2 + g'^2}} \\ Z_\mu &= \frac{gW_\mu^3 - g'B_\mu}{\sqrt{g^2 + g'^2}} \ , \end{aligned}$$

the mass terms $\frac{1}{2}M_Z^2 Z_\mu^2 + \frac{1}{2}M_A^2 A_\mu^2$ can be identified and we get

$$\begin{aligned} M_A &= 0 \\ M_Z &= \frac{1}{2}v\sqrt{g^2 + g'^2} \ . \end{aligned}$$

Re-expressed in terms of the Weinberg angle:

$$\frac{M_W}{M_Z} = \cos\theta_w \ .$$

We get a massless photon and massive gauge bosons W and Z. The masses of W and Z are not equal because of the mixing between W_μ^3 and B_μ. This is just one possibility to generate masses for the weak gauge bosons. More complicated choices for the Higgs field lead to different relations between M_W and M_Z.

3.4.4 Masses of the Fermions

The same Higgs doublet which generates the masses of the gauge boson can be used to give mass to the leptons and quarks by introducing a third term in the Lagrangian, which is called Yukawa-Interaction term. In case of the electron this looks like:

$$\mathcal{L}_3 = -G_e \left[(\overline{\nu}_e, \overline{e})_L \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} e_R + \overline{e}_R (\phi^-, \overline{\phi}^0 \begin{pmatrix} \nu_e \\ e \end{pmatrix}_L \right].$$

After spontaneous symmetry breaking and the substitution

$$\phi = \sqrt{\frac{1}{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix},$$

this reduces to

$$\mathcal{L}_3 = -m_e \overline{e} e - \frac{m_e}{v} \overline{e} e h.$$

The electron mass became

$$m_e = \frac{G_e v}{\sqrt{2}}.$$

Besides the mass term, this Lagrangian contains a term coupling the electron to the Higgs scalar which is very small and does not produce any detectable effect. Heavier fermions like top quarks have a much stronger coupling to the Higgs field since the coupling is proportional to the fermion masses. The value G_e is arbitrary, thus the mass m_e cannot be predicted. Also the mass m_H of the neutral Higgs boson itself cannot be predicted by this formalism.

This model has five free parameters for one generation of leptons:

- The two gauge couplings for SU(2) and U(1): g and g'
- The two parameters in the scalar potential $V(\phi)$
- The Yukawa coupling constant G_e

For each added lepton generation, an additional Yukawa coupling parameter occurs. These parameters fully determine the observables. They can be reexpressed in terms of parameters that are measurable, namely e, $\sin\theta_W$, and the masses m_H, m_W, m_e.

The motivation for constructing the theory like this instead of simply adding mass terms to the Lagrangian that break local gauge invariance resides in the renormalizability which is preserved by the Higgs mechanism.

3.4.5 The Higgs Boson at the LHC

In the Standard Model, one weak isospin Higgs doublet is introduced leading to the existence of one elementary Higgs boson after electroweak symmetry breaking. The only unkown parameter in this model is the mass m_H of the Higgs boson itself. All production and decay properties of the Higgs boson are fixed with its mass [14]. The search for the Higgs boson is therefore connected to the search for the characteristical final state signatures that depend on its mass. The decay modes of the Higgs boson can roughly be divided into two mass ranges above and below 135 GeV/c^2 as visible on the left side of Figure 3.2. For $m_H < 135\text{GeV}/c^2$

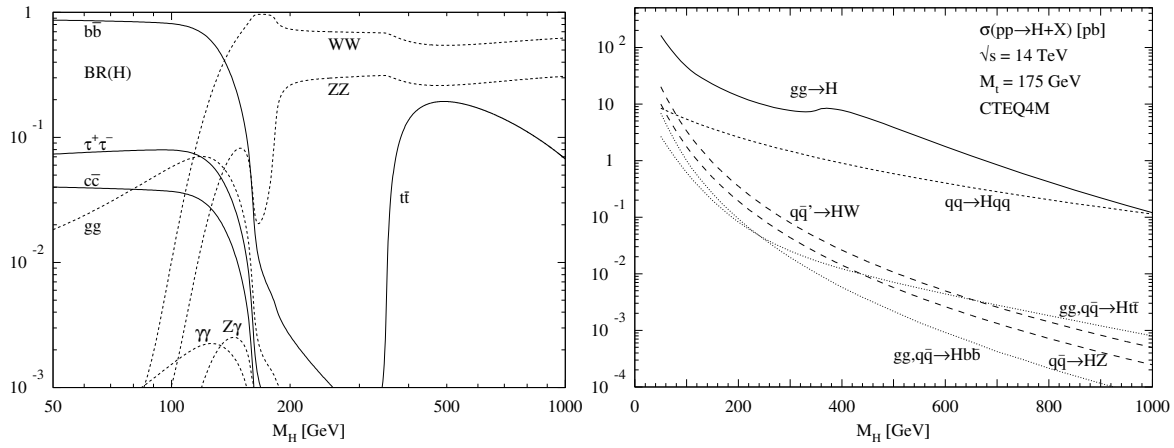

Figure 3.2: On the left: Decay branching ratios of a Standard Model Higgs boson [15]. On the right: Leading order cross sections for different Higgs boson production processes in dependence on the Higgs boson mass m_H in the Standard Model [16].

the main decay channels are $b\bar{b}$ and $\tau^+\tau^-$ with a total branching ratio of more than 90%. The decay modes into $c\bar{c}$ and gluons do not play a significant role at the LHC. For the decays into quark pairs, QCD corrections have to be taken into account which are known up to three-loop order [17, 18, 19, 20, 21, 22, 23]. The electroweak corrections are known up to NLO [24, 25, 26, 27]. Even though the branching ratio into photons $H \to \gamma\gamma$ is extremely small with values around $2 \cdot 10^{-3}$, it is an important discovery channel in this mass range. This decay is mediated by loops of heavy objects like top quarks, b quarks and W bosons.

The mass range above 135 GeV/c^2 is dominated by the decay into WW and ZZ pairs. Starting with the kinematic threshold for $t\bar{t}$, also this channel contributes a small fraction. The decay width of the Higgs boson increases significantly in the high mass range. For very high Higgs mass values, the width has a similar magnitude as the mass itself, which conceals the resonance interpretation of the Higgs boson.

There are four main production mechanisms for Higgs bosons at the LHC:

- Gluon fusion $gg \to H$, which is the largest contribution in the full Higgs mass range [28], shown in Figure 3.3a
- Vector boson fusion $qq \to qq + (WW \text{ or } ZZ) \to qqH$, for large Higgs masses [29, 30, 31], shown in Figure 3.3b
- Higgs-strahlung $q\bar{q} \to (Z^* \text{ or } W^*) \to H + (Z \text{ or } W)$, which is an alternative in the intermediate mass range $m_H < 2m_Z$ [32, 33], shown in Figure 3.3c

- Associated Higgs production $q\bar{q}$ or $gg \rightarrow t\bar{t}H$, which can contribute in the very low mass region, shown in Figure 3.3d. The analysis of this channel is the main topic of this thesis.

The production cross sections for all these channels are shown on the right side of Figure 3.2. In case of the associated production with top pairs, full NLO QCD corrections are calculated which increase the LO cross section by 20% [34, 35, 36].

3.5 Search for the Higgs Boson

The search for the Higgs boson has already been conducted at the LEP experiments giving no direct indication of a Standard Model Higgs boson production. The exclusion limit has been determined to $m_H > 114.4$ GeV/c^2 at the 95% confidence level [37]. The expected main production mechanism at LEP is the Higgsstrahlung process $e^+e^- \rightarrow HZ$ with $H \rightarrow b\bar{b}$ as the main decay channel, because only low mass Higgs hypotheses are relevant at LEP energies. The searches concentrate on final states with four jets ($H \rightarrow b\bar{b}$, $Z \rightarrow q\bar{q}$), with leptons and missing energy ($Z \rightarrow l\bar{l}$ and $Z \rightarrow \nu\overline{\nu}$) as well as final states with τ leptons ($H \rightarrow b\bar{b}$, $Z \rightarrow \tau^+\tau^-$ and $H \rightarrow \tau^+\tau^-$, $Z \rightarrow q\bar{q}$). The combined result for the determination of a lower bound of the Higgs mass in these channels, using data from all experiments at LEP is shown in Figure 3.4.

Also experiments at Tevatron have invested big efforts in finding direct evidence for Higgs boson production in $p\overline{p}$ collisions at $\sqrt{s} = 1.96$ TeV. A total number of sixteen different channels have been combined at both CDF and D0 in order to obtain upper limits on the Standard Model cross sections. The results are shown in Figure 3.5. A value of < 1 in this

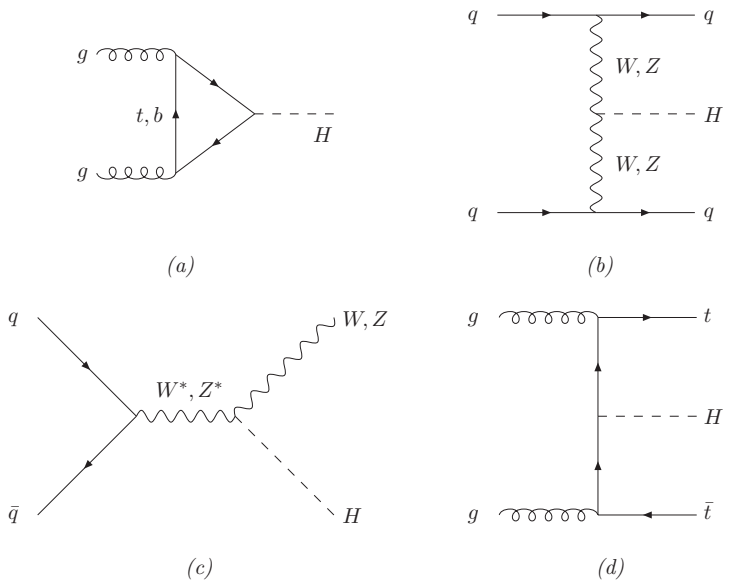

Figure 3.3: The four main production processes of Higgs Bosons at the LHC: a) gluon fusion, b) vector boson fusion, c) Higgs-strahlung, d) associated production with top quarks. [1]

figure indicates an exclusion of the corresponding Higgs mass at the 95% confidence level. It is also visible that the expectations and observations are in good agreement, but do not yet reach the significance to exclude any Standard Model predictions. The best exclusion power is achieved in the mass range between 150 and 180 GeV/c^2. This range is dominated by the H $\rightarrow$ WW channel which has less background processes.

Besides the direct measurements discussed above, indirect measurements have the potential of providing important constraints on m_H. The masses of the W boson, the top quark and the Higgs boson are connected within the Standard Model through radiative corrections. This has already been mentioned and shown in Figure 1.1. A global fit in the Standard Model using all 14 observables measured at the Z pole and including also direct measurements of m_t, m_W and Γ_W predicts a low Higgs mass [4] which is illustrated in Figure 3.6. This diagram shows the χ^2 fits for the Higgs boson mass for different assumptions for $\Delta\alpha^{(5)}_{had}$, where $\Delta\alpha^{(5)}_{had}$ represents the effect from the running of the electromagnetic coupling due to light quark loops in the photon propagator. It is clearly visible that the indirect measurements are in good

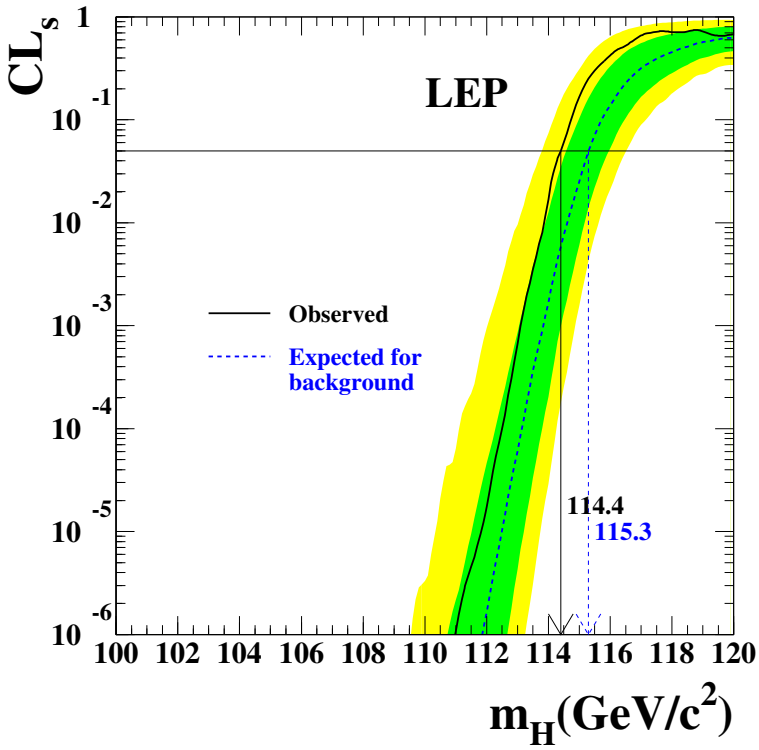

Figure 3.4: $CL_s = CL_{s+b}/CL_b$ represents the ratio of probabilities to obtain the observed event configuration under the assumption of a signal plus background hypothesis (CL_{s+b}) or a background only hypothesis (CL_b). The solid line is the observation, while the dashed line is the median background expectation. The dark and light shaded bands around the background expectation correspond to the 68% and 95% probability. The intersection of the horizontal line at $CL_s = 0.05$ with the observed curve is used to define the lower bound on the Standard Model Higgs boson mass with 95% confidence level. [37]

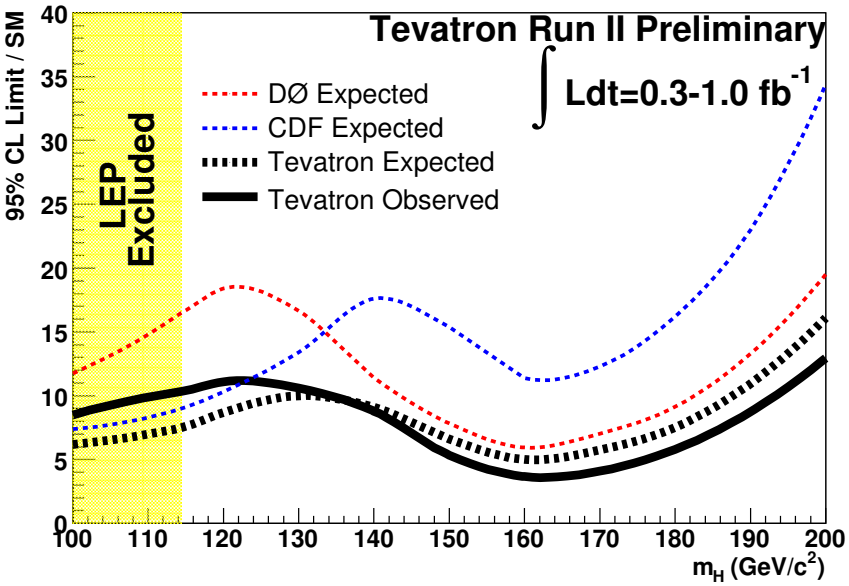

Figure 3.5: Limits on the Higgs boson production cross section normalized to the Standard Model prediction as function of the Higgs mass m_H for the combined CDF and D0 analyses. The dashed lines show the expectations determined from background-only pseudo-experiments. The solid line shows the observation. [38]

agreement with the directly excluded mass of $m_H > 114.4$ GeV/c^2. The conclusion is that the upper limit corresponding to $\Delta\chi^2 = 2.7$ is $m_H < 166$ GeV/c^2 at the 95% confidence level. While these searches do not poof that the Higgs boson actually exists, they can be considered as a guideline for the mass range in which the Higgs boson should be expected.

The search for the Higgs boson will be continued at the LHC where the discovery potential reaches to the theoretical upper limit [40]. Exploiting its properties discussed in Section 3.4.5, the search strategies at LHC cover a large variety. In the mass region $m_H < 150$ GeV/c^2 the small width $\Gamma_H < 1$ GeV/c^2 can be used to find a narrow peak in the $H \to \gamma\gamma$ and $H \to ZZ^* \to l^+l^-l'^+l'^-$ channels. The large cross section of the $gg \to H$ production motivates a discovery in the so called "golden channel" with four leptons in the final state that provides a very clean signature. In contrast to $H \to \gamma\gamma$, which suffers from large jet background, the backgrounds to $H \to ZZ^* \to l^+l^-l'^+l'^-$ are moderate. The $H \to \gamma\gamma$ channel can also be searched for in associated production modes WH and $t\bar{t}H$ with isolated leptons from $W \to l\nu_l$ and in H+jet productions with a high E_t jet. These channels have less backgrounds and less requirements on the resolution of the electromagnetic calorimeter, but have smaller production cross sections. The $H \to b\bar{b}$ decay can only be searched for in the associated $t\bar{t}H$ mode because of large backgrounds as discussed in Chapter 5.

The $H \to Z\gamma$ and $H \to \mu^+\mu^-$ channels have very small branching fractions and can only be discovered with high integrated luminosities exceeding 100 fb^{-1}. The gauge boson fusion processes $qq \to qqH$ provide a low jet activity in the central rapidity region due to the lack of colour exchange in the hard process. Additionally, two "tagging" jets emerge in the forward direction that can be used together with a central jet veto for event selection and background suppression. These production processes are accessible in almost all decay modes.

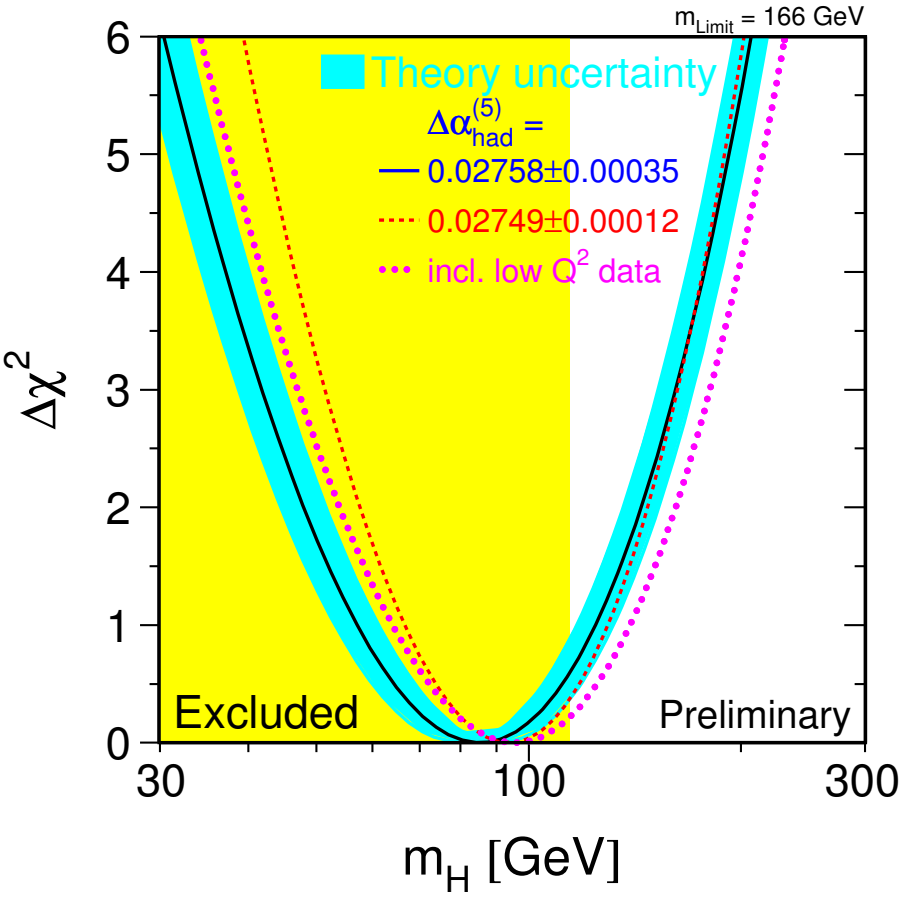

Figure 3.6: $\Delta\chi^2(m_H) = \chi^2_{min}(m_H) - \chi^2_{min}$ as function of m_H. The band around the curve represents the theoretical uncertainty due to missing higher order corrections. The vertical band shows the exclusion limit at $m_H > 114.4$ GeV/c^2 from direct measurements. [4] [39]

The H → WW decays are of special interest because they provide the highest branching fraction in the mass region above 120 GeV/c^2 up to 200 GeV/c^2. All decay modes are accessible, but the fully leptonic modes require good understanding of backgrounds and the Higgs mass can only be reconstructed in the transverse plane because of the two neutrinos. Above 200 GeV/c^2 the sensitivity of H → ZZ → 4l is again the largest one, while above 500 GeV/c^2, where the width is large, the H → (ZZ or WW) decays are used also in semileptonic or fully hadronic final states. The discovery reach for the Standard Model Higgs boson is summarized in Figure 3.7, where the expected significances are shown for various channels in dependence on the Higgs boson mass.

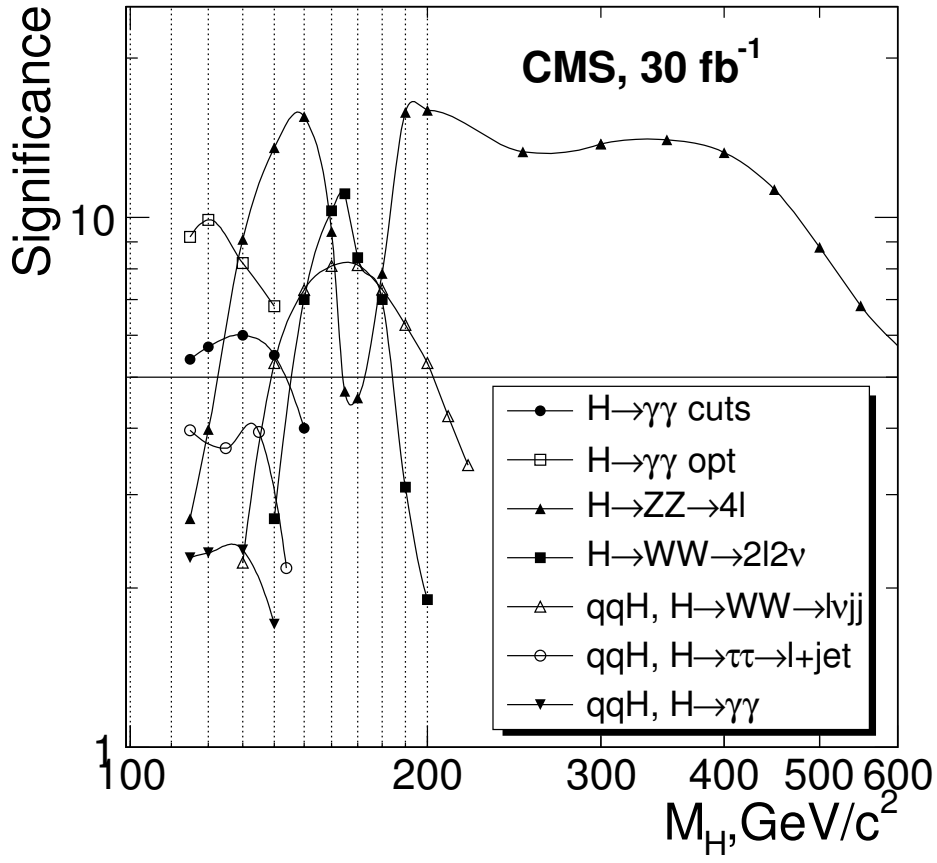

Figure 3.7: Expected signal significances in 30 fb^{-1} for different production and decay channels. [1]

Chapter 4

The CMS Software and Analysis Environment

This chapter describes the software and analysis framework of CMS as published in [41] and tries to give an overview of the general design concept. The topics, which the author of this thesis has contributed to, are described in detail. In particular, improvements in b-tagging and performance studies for jet reconstruction, which are important components of the $t\bar{t}H$ analysis are investigated in Sections 4.2.7 and 4.3. The rest of the material is kept succinct and is given for completeness.

The main goals of the CMS software are to process the detector output in the trigger farms, deliver the data to physicists and provide the necessary instruments and tools to analyze these data in order to produce physics results. The basic application areas can be subdivided into the following categories:

- Event filter and High-Level Trigger
- Simulation including Generation and Digitization of Monte Carlo events
- Reconstruction
- Calibration and Alignment
- Creation of High-Level Objects (muons, electrons, jets, ...)
- Physics Tools and Visualization
- Physics and Data Quality Monitoring

The simulation and digitization is shortly summarized in Section 4.1, Subsection 4.2.1 gives an overview of the event filter and trigger, while the rest of Section 4.2 describes the reconstruction and higher level objects.

The CMS software is based on the framework COBRA ("Coherent Object-oriented Base for Reconstruction, Analysis and simulation") [42] which implements the fundamental architecture. It provides the essential subsystems, like CARF, the "CMS Analysis and Reconstruction Framework", as well as the Detector Description Database "DDD" and the interface to Monte Carlo Generator information. It implements two basic principles, "event driven notification" and "action on demand", which ensure that only the required invocations and

calculations are performed. This is realized by means of the "observer" design pattern and the concept of implicit invocation of reconstruction algorithms.

The collection of reconstruction algorithms is labeled ORCA ("Object Oriented Reconstruction for CMS Analysis"). ORCA provides the physics reconstruction tools, i.e. track, vertex, electron, photon, muon and jet reconstruction which are described in Section 4.2.

During the year 2006, the CMS software has undergone a reorganization. The software framework, including the basic concepts and data formats have been reimplemented in order to account for the rquirements that have been identified, but could not easily be implemented in the old framework. Most of the reconstruction algorithms are supposed to stay untouched and the physics performance should therefore be identical. The descriptions of the simulation and reconstruction software and the performance according to the Physics Technical Design Report (PTDR) [41], which are given in Sections 4.1 and 4.2, should therefore be also valid in the new framework, called CMSSW. At the time of writing of this thesis, CMSSW was not yet in a reliable state of stability and simulated data were not available. Therefore the $t\bar{t}H$ analysis in Chapter 5 is carried out in the old software framework and no further discussion of CMSSW is given here.

4.1 Simulation and Digitization

The full CMS detector simulation package is named OSCAR ("Object oriented Simulation for CMS Analysis and Reconstruction"). It is based on the COBRA framework. OSCAR employs the GEANT4 toolkit [43] for the simulation of all the CMS detector components.

The input for the simulation are events from Monte Carlo generators like PYTHIA [44]. The generated particles are propagated through the detector and the magnetic field, in parallel with the simulation of the interactions with the detector material and the ensuing energy deposition, the so-called creation of detector "Hits", which is entirely performed by GEANT4. In a subsequent step, which is called "digitization", the response of the detector electronics and readout system is simulated. The output of this step needs to be as close as possible to real data that would come from the CMS detector.

During the digitization step, a certain amount of pileup events are merged into the sample: during the low luminosity phase ($L = 2 \times 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$) of the operation of the CMS detector, about 3.5 inelastic proton proton collisions per bunch crossing occur at the same time. During the high luminosity runs ($L = 10^{34}\mathrm{cm}^{-2}\mathrm{s}^{-1}$) there will be 17.5 parallel collisions on average. The collisions to be merged in, are randomly chosen from a pregenerated sample making sure that they are not reused again in the same order.

During the digitization of the response of the inner tracking system, the entrance and exit points together with the deposited energy of particles passing through the sensitive volumes are recorded and a charge distribution is generated which is mapped to the strip (pixel) geometry. The fractional charge contribution for each channel is determined which leads to a collection of hit channels for all tracks. Noise is added to all channels according to a gaussian distribution and a signal-to-noise ratio of 11 (70) in the strip (pixel) detectors. The digitization of each channel is then performed by rounding the collected charges to integer values.

The simulation of the electromagnetic calorimeter ECAL emulates the signal pulse for each hit according to a nominal longitudinal light collection curve which is a function of the distance from the front face of the crystal. In the case of the hadron calorimeter HCAL, the

simulation converts the deposited energy in the scintillators to numbers of photoelectrons and adds poisson fluctuations and noise.

The muon detector digitization is performed by simulating the response of the Time to Digital Convertes (TDC). The behaviour of the muon drift cells is simulated as a function of the muon direction and impact position. The time resolution is smeared according to an intrinsic cell resolution of 220 μm. The output signal is then obtained by adding the time of flight from the primary vertex and signal propagation time along the cell wire.

4.2 Reconstruction and Selection

4.2.1 Event Filter and High Level Trigger

The goal of the CMS Trigger and Data Acquisition System (TriDAS) [11, 45] is to reduce the enormous information flow produced by a bunch crossing rate of 40 MHz to a manageable data stream of about 100 Hz without loosing interesting physics events. This task is achieved by splitting the trigger into several steps. The first step is the Level-1 Trigger which reduces the event rate to less than 100 kHz. A short overview of the Level-1 Trigger is given in Section 2.2.4.

The second step is the so called High-Level Trigger (HLT). It reads out the front-end detector electronics after a Level-1 Trigger accept and collects all data produced by a specific bunch-crossing. This is followed by a fast processing of physics selection algorithms on the particular event. After the event has been accepted, it is forwarded to the monitoring and mass storage system. This way, the event rate is reduced to the final 100 Hz (The exact value of the final event rate is currently under discussion. It will be somewhere in between 100 and 200 Hz.).

Figure 4.1 shows an illustration of the main functional components of the Data Acquisition System (DAQ). The Builder Network is the high-bandwidth connection between the readout

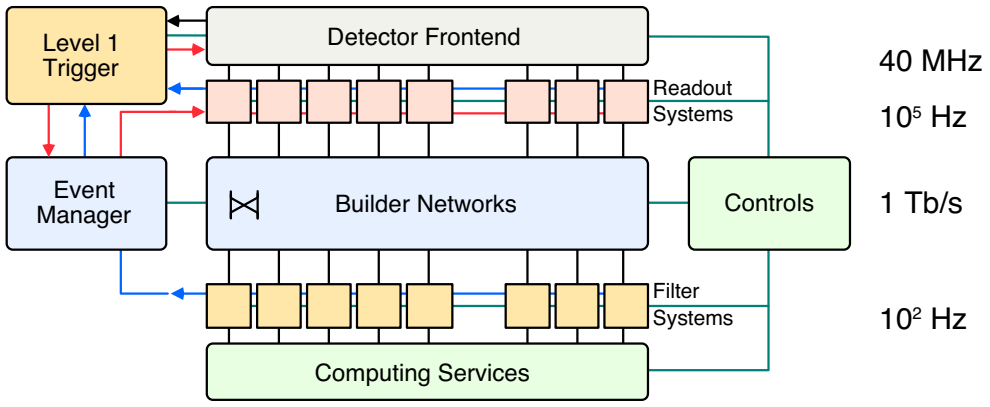

Figure 4.1: Schematic view of the basic Architecture of the Data Acquisition System. [45]

units of the subdetector front-end modules, which provide the data after a Level-1 Trigger signal, and the High-Level Filter System which consists of a processor farm executing the High-Level Trigger algorithms. Approximately 1000 dual-CPU nodes will be installed for this purpose. The Event Manager is responsible for the data flow through the DAQ, while the

Control and Monitor system takes care of configuration, control and monitoring tasks. The Computing Services provide the interface to the storage and offline systems.

The task of the HLT is to filter the event rate to 100 Hz which corresponds to a cross section of 10 nb. Compared to typical physics cross sections, like $W \rightarrow e\nu_e$ which is of the same order of magnitude, it is obvious that already a large part of the physics selection is performed online. In this view, some important requirements on the HLT system have been identified: the dead-time of the whole DAQ system should be less than 2%. All events should have a tag which identifies their specific trigger selection path. Calibration and alignment constants should not have a large impact on the HLT efficiency, in particular, the trigger selection must be computable using only data itself with as little reference to simulation as possible. The uninteresting events should be rejected as soon as possible and the system must not rely on the availability of the full information. Enough information for monitoring purposes should be provided in order to enable quick solutions in case of problems. Also the rejected events must be monitored to a certain extent in order to maintain knowledge about the discarded information. Therefore, the control and monitoring of the HLT algorihms is a crucial aspect.

The optimization of the HLT system, like configuration of algorithms and their corresponding thresholds is a compromise between the physics needs and the total available rate. It is a long-term project that has undergone many changes and is still under development. The current situation of the lepton trigger paths used in the analysis in Chapter 5 is summarized in the following:

- Muons: The HLT muon algorithm is divided into a Level-2 and a Level-3 selection. Level-2 applies calorimeter based isolation criteria and a standalone muon reconstruction is used. After acception, a more CPU intensive Level-3 reconstruction is performed, which is described in more detail in Section 4.2.2. At this level, a tracker isolation using the sum over the transverse momenta p_t of tracks around the muon candidate is applied. For the single muon stream, a p_t threshold of 19 GeV/c is applied, while a threshold of 7 GeV/c is used for the di-muon selection.

- Electrons: The selection proceeds in three steps. At Level-2, only calorimeter information is used. In this step, the energy deposits are clustered to obtain an estimate of the energy and position, which enables the application of cuts on the transverse energy E_t. At Level-2.5 a division into electron- and photon-candidates is obtained by matching the Level-2 information to pixel detector hits. Photon candidates have significantly higher p_t thresholds. Finally, at Level-3 the full track information for electrons is used and an isolation for photons is applied. The final p_t threshold for single electons is 26 GeV/c and 80 GeV/c for single photons, while (12,12) GeV/c are used for di-electons and (30,20) GeV/c for di-photons.

Furthermore, there are HLT trigger paths for jets, τ-leptons, missing transverse energy, b-jets and more complex cross-channel triggers that are not described in detail here. To give an estimate of the expected breakdown of the trigger rates, Figure 4.2 shows a graphical representation of the HLT bandwidth for the different trigger paths. In this figure, the "old" values from the DAQ TDR (2002) are compared to the values from the PTDR (2006). For these values, a luminosity of $L = 2 \times 10^{33}\text{cm}^{-2}\text{s}^{-1}$ is assumed.

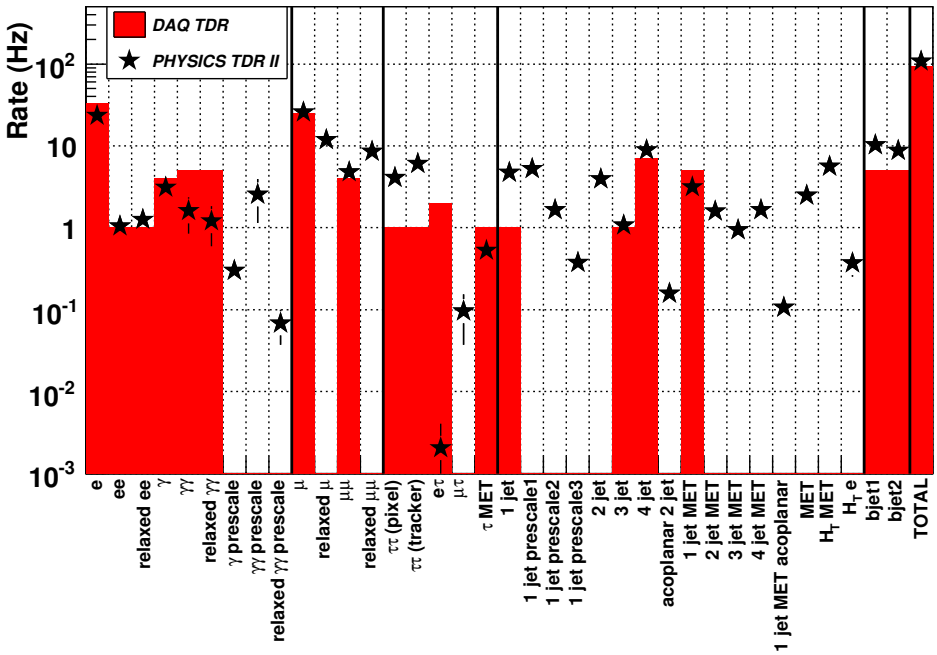

Figure 4.2: Graphical representation of the HLT bandwidth. Compared are the values published in the DAQ TDR [45] and in Volume 2 of the PTDR. [1]

4.2.2 Muon Reconstruction

Muon reconstruction is first performed "standalone" with track segments obtained by the local reconstruction in the muon chambers. The muon trajectories are seeded from the innermost chambers and are worked from inside-out using the Kalman filter [46] method. In this procedure, the predicted state vector at the next measurement surface is compared with real measurements and updated accordingly. The state is propagated through the muon system using the GEANE [47] package which takes care of the estimation of the effects of energy loss in the material, multiple scattering, and non-uniformity of the magnetic field. Afterwards, a backward Kalman filter is performed, from outside-in to determine the track parameters at the innermost muon detector surface. Finally, the track is propagated to the interaction point which is defined by the beam spot size ($\sigma_{xy} = 15$ μm, $\sigma_z = 5.3$ cm). This "standalone" muon reconstruction technique does not include any information from the silicon tracker and is therefore less CPU intensive enabling its use in the Level-2 trigger.

The "global" muon reconstruction method, which is used in the Level-3 trigger extends the muon trajectories in order to include information about hits in the silicon tracker. Again, the GEANE package is used for the extrapolation through the material. According to the trajectory, regions of interest are defined in the tracker in which regional track reconstruction is performed. This track reconstruction consists of three steps. First, the trajectory building (seeded pattern recognition), second, trajectory cleaning (resolution of ambiguities) and third, smoothing (final fit). The resulting p_t resolutions of these two algorithms are shown in Figure 4.3.

The method of "muon isolation" is used to distinguish muons produced in jets from muons

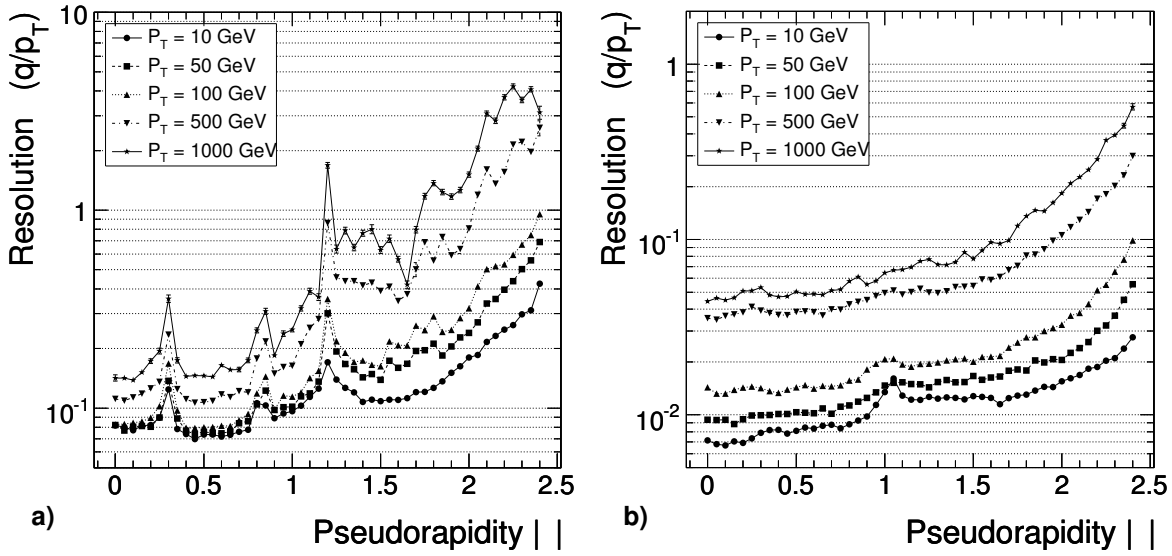

Figure 4.3: q/p_t resolution in dependence on $|\eta|$ according to a gaussian fit to $\frac{q^{rec}/p_t^{rec} - q^{gen}/p_t^{gen}}{q^{gen}/p_t^{gen}}$, where q is the charge and p_t is the transverse momentum of the reconstructed or generated muon, respectively. The left plot shows the standalone muon reconstruction, while the right plot shows global muon reconstruction. [41]

coming from decays of heavy objects, like leptonic W or Z boson decays. For this purpose two basic isolation algorithms are used. First, the calorimeter isolation, which is based on the sum of the calorimeter energy in a cone around the muon. The energy deposit in the cone is defined as a weighted sum of the transverse electromagnetic and hadron calorimeter energy by $E_t = \alpha E_t^{ECAL} + E_t^{HCAL}$, where $\alpha = 1.5$ to account for the better discrimination power of the ECAL.

In a similar procedure, an isolation is defined using the p_t sum of tracks in a cone around the muon direction. This can be performed for pixels only, which is fast, and with full tracker information, which is more precise. In all the isolation algorithms, an optimization is performed by determining the energy and momentum thresholds as a function of the cone size and pseudo-rapidity. A comparison of the performance of the different types of isolation algorithms is given in Figure 4.4. This figure shows the efficiency of selecting non-isolated muons from a $b\bar{b} \rightarrow \mu X$ decay versus the nominal efficiency to select isolated muons from $W \rightarrow \mu\nu$ after cone size and energy threshold optimizations.

4.2.3 Electron Reconstruction

The electron reconstruction starts with the determination of a so called "supercluster" in the electromagnetic calorimeter. A supercluster is a collection of calorimeter clusters, which consist of arrays of ECAL crystals. Typically, a supercluster has an angular extension in ϕ because of the emission of bremsstrahlung along the curved trajectory. The amount of radiated bremsstrahlung depends on the traversed material budget and can be very large. About 50% of the electrons radiate 50% of their energy before reaching the ECAL surface and for 10% of the electrons, more than 95% is radiated. Therefore, advanced superclustering algorithms are employed which search along the ϕ direction for energy deposits, followed by

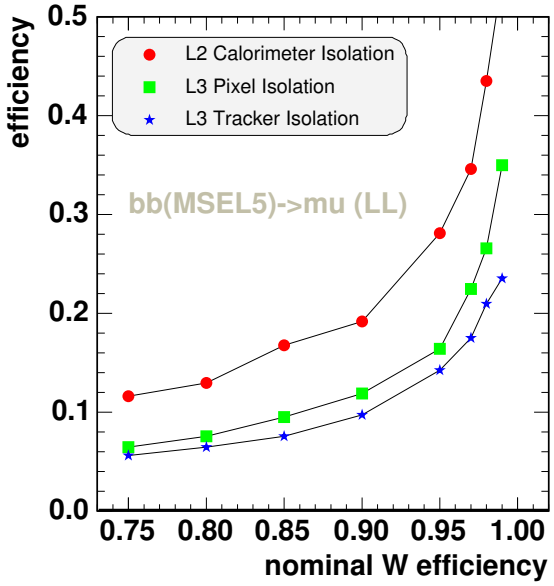

Figure 4.4: Performance of the muon isolation algorithms in comparison for the three algorithm types. The isolation efficiency is shown in dependence on the nominal efficiency for isolated signal muons from W boson decays. [41]

algorithmic energy corrections as explained in [45, 41].

Based on the supercluster, the position of hits in the pixel detector is predicted by backwards propagation through the magnetic field. The efficiency to find two pixel hits with this method is 90% for electrons at $p_t = 10$ GeV/c. These pixel hits serve as seeds for the subsequent track reconstruction with the full tracker. The default track reconstruction method using the Kalman filter technique is only appropriate if the random fluctuations are Gaussian, e.g. in the case of multiple scattering effects. This is not the case for the large amount of energy radiation of electrons. Here, a more complex nonlinear filter approach using a Gaussian Sum Filter (GSF) gives a better description of the propagation of electrons.

The reconstructed energy E^{rec} and momentum p^{rec} of the electron are then matched in order to improve the overall measurement. Depending on p_t and E^{rec}/p^{rec} the relative weight of the calorimetry or tracker information is taken into account. The improvement is due to the opposite behaviour of the energy (momentum) resolution as shown on the right side of Figure 4.5

Since the behaviour of electrons might be very different on an event by event basis, four classes of electrons are defined: the “golden electrons” which have low radiation and tracks well matching the supercluster. The “big brem electrons” have large amount of bremsstrahlung, but still a good measurement and matching between supercluster and track. The “narrow electrons” have less bremsstrahlung than the “big brems” but a relaxed geometrical matching. The “showering electrons” constitute the rest, they are likely to have early hard radiations, and bad energy-momentum matching. The fraction of electrons of the four classes depends on $|\eta|$, on average there are 50% showering electrons and 20% golden electrons. To give an estimation of the precision of the energy measurement in the ECAL, the left side of Figure 4.5 shows the reconstructed energy normalized to the generated energy of electrons in the barrel

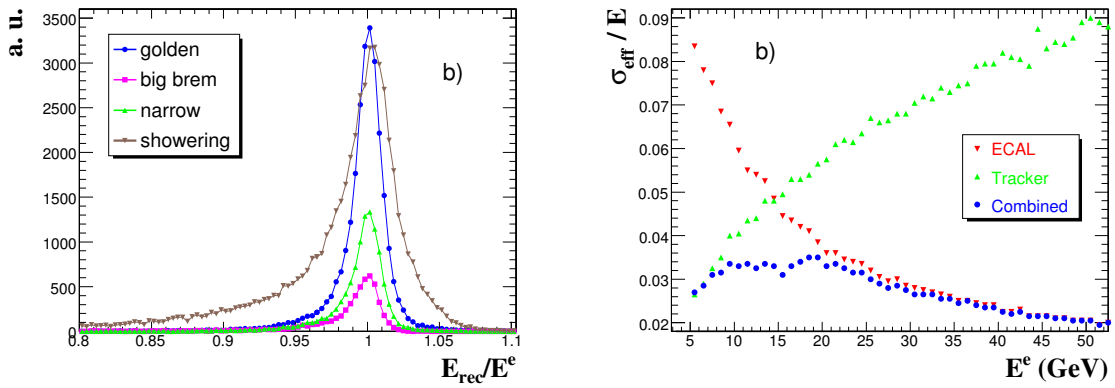

Figure 4.5: On the left: Distribution of the reconstructed and corrected energy of electrons E_{rec} normalized to the generated energy E^e in the barrel only. The electrons are uniformly distributed in energy between 5 and 100 GeV. On the right: Fractional resolution depending on the generated energy E^e, measured with the ECAL, the tracker and the combination of both. [41]

for the different electron classes.

Electron isolation is performed in a similar way as muon isolation using tracks in a cone around the electron direction.

4.2.4 Jet and MET Reconstruction

The primary objects in Jet and also in Missing Transverse Energy (MET or E_T) reconstruction are the electromagnetic and hadron calorimeter (HCAL) cells. In case of the HCAL, the cells are arranged in tower patterns. This tower pattern can be extended to also include the ECAL crystals. This way, a total number of 4176 "ECAL plus HCAL towers" is obtained. These towers serve as input to all jet and MET reconstruction algorithms.

An important part is the preselection of the ECAL plus HCAL towers especially for low p_t jets, because of a significant noise contribution. Therefore, a cut of $E_t > 0.5$ GeV or $E > 0.8$ GeV is applied before a tower is used in the jet reconstruction.

In CMS, three basic jet reconstruction algorithms are used which are briefly described in the following:

- Iterative Cone Algorithm: First, the input objects are ordered by E_t. A cone of size $\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}$ around the highest E_t object is cast and the objects inside the cone are used to form a proto jet. The obtained direction of this jet is used to seed a new proto jet, which is repeated until the energy does not change by more than 1% and the direction does not change by more than $\Delta R < 0.01$. The stable jet is added to the list of jets and the objects inside the cone are removed from the list of objects for the next iteration. The iteration proceeds until no objects above a seed threshold, which is a parameter of the algorihm are available. The cone ΔR is also a parameter.

- Midpoint Cone Algorithm: In a similar way as for the Iterative Cone Algorithm, an iterative procedure to find stable proto jets is applied, but the input objects are not removed from the list for the next iteration. This way, overlapping proto jets are

possible. In a second iteration, a midpoint is calculated for overlapping jets as the direction of the combined momentum. This midpoint serves as seed to find further proto jets. Afterwards, a splitting and merging procedure is performed. In this procedure, two proto jets are merged into one if the shared transverse energy fraction is greater than the parameter f, otherwise the shared objects are assigned to the jet which is closer.

- Inclusive k_T Algorithm: For each input object i and each pair of objects (i, j) two values, d_i and $d_{i,j}$ are calculated: $d_i = E_{t,i}^2 R^2$, where R is a parameter usually set to 1, and $d_{i,j} = \min\left\{E_{t,i}^2, E_{t,j}^2\right\} R_{i,j}^2$ with $R_{i,j}^2 = (\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2$. If the smallest of these values is of type $d_{i,j}$, the two objects i and j are removed from the list, combined and added back to the list as merged object. If an object of type d_i is the smallest in the list, this object is removed from the list and added to the list of final jets. This way, all objects which have a distance $R_{i,j} < R$ are merged and it follows for all final jets i, j that $R_{i,j} > R$.

It should be noted that these algorithms can be applied to all kinds of input objects that behave like fourvectors. This means that not only calorimeter towers can be used but also generated particles, for example.

The energy of the reconstructed jet differs from the true energy. This is due to several effects. For instance, the algorithm itself is not able to collect the exact jet energy because of out-of-cone effects or because of the inclusion of energy from pileup. Furthermore, the energy measurement in the calorimeter is not always precise and suffers from noise and lost energy. Muons and neutrinos in jets are also not included in the energy measurement. Therefore, jet calibration procedures are applied. Two different types of calibration are available: particle and parton-level calibrations. The particle level calibrations, also named MC calibrations, simply apply the identical jet algorithm to generator particles, followed by a matching of the generator jets to calorimeter jets. The energy difference is corrected for, depending on p_t and $|\eta|$. This kind of calibration can only correct a part of the mismeasured energy, because out-of-cone effects, for example, still occur on particle level. Parton level corrections account for the originating parton, before any showering. This kind of calibration depends on the hadronization model and the type of the originating parton. The identification of the primary parton is not always possible in an unambiguous way. Furthermore, a number of data-driven calibration procedures will be used in order to cross check the various calibration methods:

- p_t balance in QCD dijet events.
- p_t balance in γ+jet events.
- W boson mass fit in $\mathrm{t\bar{t}}$ events.

In order to give an example for the performance of the CMS jet finding in terms of precision of the jet energy measurement in various detector regions, Figure 4.6 shows the resolution in dependence on the MC jet energy.

The vector of the missing transverse energy is obtained by summing over all calorimeter towers, assuming the mass of each fourvector to be zero, and the direction to be given by η, ϕ and the collision point. Corrections from muons and jet calibrations can be taken into account, but are mostly specific to the respective analysis. Therefore, the treatment of missing transverse energy is described in more detail in Section 5.3.5.

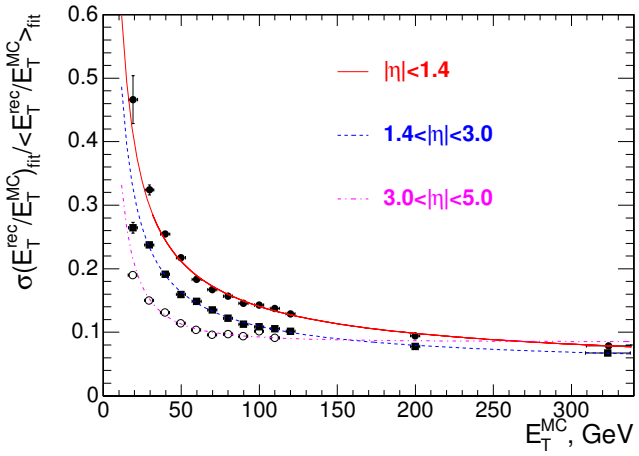

Figure 4.6: Transverse energy resolution of jets as function of the energy of generator particle jets for three different $|\eta|$ ranges. The jets are reconstructed with the Iterative Cone Algorithm with a cone radius of $R = 0.5$. The Monte Carlo jet calibration has been applied. [41]

4.2.5 Track Reconstruction

The reconstruction of charged particle tracks consists of five steps. First, the hits in the pixel and silicon strip detectors are clustered by combining adjacent pixels or strips passing a predefined signal to noise ratio. According to the cluster structure, the position and corresponding error is determined.

The second step is called "seed generation". A seed defines the initial trajectory parameters and at least 3 pixel hits or 2 hits and the beam constraint are necessary to form a seed. An efficient seed finding makes use of a region of interest in which the pixel hits are searched, for example in the case of extending a standalone muon track to the tracker.

After the identification of the seed, the third step consists of the building of the full trajectory. This procedure applies a combinatorial Kalman filter which starts with the track parameters estimated by the seed. The trajectory is propagated iteratively inside-out, including hits on each consecutive layer until the last point. Several hits on each layer can be compatible with the trajectory, therefore all possible combinations are calculated. Quite a large number of parameters define the behaviour of this step. Among these parameters (the default values are given in brackets in the following) is the maximum number of candidates propagated at each step (5), the maximum χ^2 of the hits considered to be compatible with the predicted track state (30), the minimum transverse momentum (0.9) and the minimum number of hits per track (5).

In the fourth step the ambiguities that arose during the trajectory building are resolved. The same track might be reconstructed starting from different seeds, or more than one track candidate might have the same seed origin. In order to avoid double-counting, the track with the least number of hits is discarded, if two tracks have the same number of hits, the one with the highest χ^2 is discarded.

In the final step, the trajectory is refitted. For each valid hit, the position estimate is re-evaluated using the current values of the track parameters. The track parameters and covariance matrix are updated according to the estimates for energy loss and multiple scattering. Afterwards, the track is smoothed by running a second filter backwards from outside-in.

To give an estimate of the performance of the track reconstruction, the resolutions of the transverse momentum p_t and transverse impact parameter d_0 are shown in dependence on $|\eta|$ in Figure 4.7.

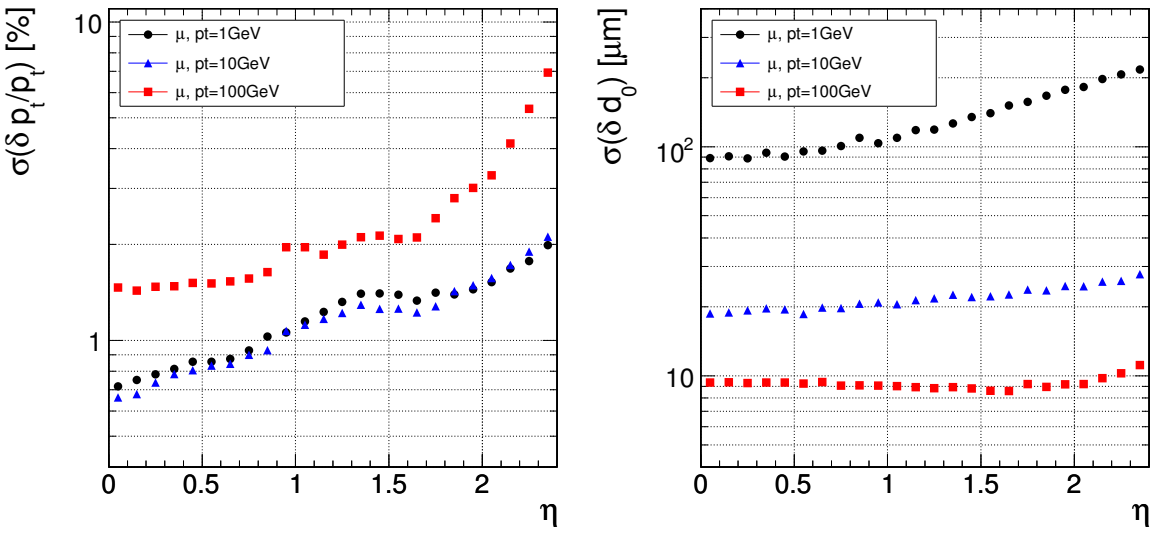

Figure 4.7: On the left: Resolution of the transverse momentum p_t. On the right: Resolution of the transverse impact parameter d_0. These resolutions have been obtained using single muon events at three different p_t values (1, 10 and 100 GeV/c). [41]

4.2.6 b-Tagging

The term b-tagging stands for the identification of jets having a primary parton origin that involves bottom quarks. The methods and tools are published in [48] and [41] and are summarized in the following.

Several properties of the production and decay mechanism of b-hadrons are being exploited for the task of b-tagging. The most important feature is the lifetime of b-hadrons of $\tau \approx 1.5$ ps (with $c\tau \approx 450$ μm) which leads to observable flight distances that can be measured with the high precision of the CMS pixel and silicon strip tracking detectors. The flight distance leads to a significant displacement of the b-hadron's decay vertex which is called secondary vertex. The displacement leads also to charged particle tracks that are not compatible with the primary event vertex. Further useful properties include the large mass of b-hadrons which is around 5 GeV/c^2 and the multiplicity of charged particles in the final state of about 5 on average. Due to the hard b-fragmentation function the b-hadron inside of a b-jet carries a large fraction of the total jet energy. Moreover, the presence of leptons in jets may be used as an indication for a b-jet since about 20% of b-jets contain one or more leptons per lepton flavour, counting only electrons and muons.

The following description of the b-tagging algorithms focuses on the method used in the $t\bar{t}H$ analysis, the "combined" secondary vertex b-tagging, because it yields the best overall performance based on b-hadron lifetime properties. In addition, it can be combined with soft lepton tagging algorithms as described in Section 4.2.7. All the plots shown in this section are obtained with the $t\bar{t}2j$ sample that acts as a background for the $t\bar{t}H$ analysis presented in

Chapter 5. It has been verified that the b-tagging performances do not differ between the $t\bar{t}H$ sample and the $t\bar{t}2j$ sample in the "algorithmic" definition (the term is explained below). The only difference is the absence of original gluon jets in the $t\bar{t}H$ sample in case of the "physics" definition. The following performance plots have been obtained with official analysis tools in the "BReco" subsystem of the CMS reconstruction software ORCA.

The two key ingredients for b-tagging are jets and tracks. For the jets, the default setup using the Iterative Cone algorithm with a cone radius of 0.5 and the "MCJet" calibration, based on correction factors from Monte Carlo simulations, is used (Section 4.2.4). The track finding implies the usage of the Kalman filter method described in Section 4.2.5.

The following track selection has been performed:

- At least 8 reconstructed hits in total (pixel and silicon strip)
- At least 2 reconstructed hits in the pixel detectors
- Transverse momentum $p_t > 1$ GeV/c
- χ^2/ndf of the track fit < 10
- Transverse impact parameter with respect to the primary vertex < 2 mm

These selection cuts are applied in order to obtain a clean set of well reconstructed tracks. The last selection criterion in this list rejects charged particle tracks having a larger displacement from the primary vertex than expected from b-hadron decays, e.g. V^0 decays, photon conversions and nuclear interactions in the beam pipe or first material layers.

For the determination of the performance presented in the following, the true jet flavour has to be known. For this task, two different definitions are utilized. In the "physics definition" the reconstructed jet is matched to the partons from the primary process by analyzing the particle content in a cone around the jet axis. If the distance is within $\Delta R < 0.3$ the matching is considered to be successful. If more than one primary parton fulfills the requirement, the jet is rejected. In this definition, gluon and quark jets with b- or c-content originating from gluon splitting are labelled according to the original gluon or quark. A large fraction of jets can not be identified unambiguously in the "physics" definition, because the direction of the primary parton may deviate too much from the direction of the jet in case of hard gluon radiation. In the "algorithmic" definition on the other hand, almost all jets can be properly identified, because this definition assigns the parton flavour that most likely determines the structure of a jet after the shower evolution. A jet from gluon splitting into $b\bar{b}$ would be labelled b-jet, because the jet looks like an original b-jet from the point of view of a b-tagging algorithm. Therefore, the main difference between these two definitions is the treatment of gluon jets that have a splitting rate of a few percent (roughly 2% for $g \rightarrow b\bar{b}$ and 5% for $g \rightarrow c\bar{c}$).

The "combined" b-tagging algorithm is mainly based on the properties of secondary vertices of weakly decaying b-hadrons. It also uses further topological information about track properties like impact parameter significances which are all combined into one b-tagging discriminator applying a likelihood method. The default secondary vertex finding algorithm is the Trimmed Kalman Vertex Finder [49]. For the purpose of this thesis, an improved vertex finder is introduced and used for the $t\bar{t}H$ analysis in Section 4.2.7. The following cuts are applied to the secondary vertices:

- The transverse distance d between primary vertex and secondary vertex must fulfill $100\ \mu\text{m} < d < 2.5$ cm.
- The transverse distance d divided by its error σ_d must fulfill $d/\sigma_d > 3$.
- The invariant mass of charged particle tracks associated to the vertex must not exceed 6.5 GeV/c^2.
- The vertex must not be compatible with a K^0_S decay. Vertices with two oppositely charged tracks are rejected if their mass is within a window of 50 MeV/c^2 around the nominal K^0_S mass.

Based on this selection, three cases, the so called "vertex categories" can be identified:

1. "RecoVertex": At least one secondary vertex candidate has been found according to the selection criteria.
2. "PseudoVertex": If no vertex is found according to the selection criteria, a so called "PseudoVertex" is created from tracks that are not compatible with the primary vertex, i.e. if they have a signed[1] transverse impact parameter significance of larger than two. This is only possible if at least two such tracks are found.
3. "NoVertex": If no vertex has been found and no PseudoVertex can be constructed.

The distribution of the categories for the different jet flavours is displayed in Figure 4.8. It is

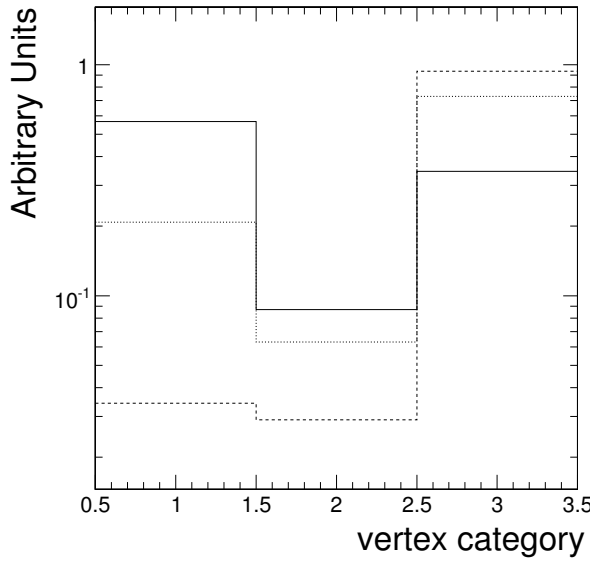

Figure 4.8: Distribution of the vertex categories for the different jet flavours. The solid line refers to b-jets, the dotted line to c-jets and the dashed line represents light flavour (uds-) jets.

[1]The sign of the impact parameter is defined positive if the track is reconstructed "downstream" in the direction of the jet with respect to the primary vertex and negative otherwise.

visible that already the category alone, i.e. the criterion if a secondary vertex has been found or not, has some power to discriminate between b-jets and non-b-jets.

In addition to the signed two-dimensional impact parameter significance of the tracks shown in Figure 4.9, the following observables are used as input for the calculation of the b-tagging discriminator of the "RecoVertex" category.

- Invariant mass of charged particle tracks associated to the secondary vertex. This is motivated by the expectation that b-hadron decays have a larger invariant mass of charged particle tracks than charm or light flavour decays.
- Multiplicity of charged particle tracks associated to the secondary vertex, because b-hadron decays have a characteristic number of charged particles of 5 on average.
- Distance between primary and secondary vertex in the transverse plane, devided by its error, called flight distance significance.
- Energy of charged particle tracks divided by the total energy of charged particles associated to the jet, motivated by the hard b-fragmentation function.
- Rapidities of charged particle tracks associated to the secondary vertex with respect to the jet direction $y = \frac{1}{2} \ln \left(\frac{E+p_\parallel}{E-p_\parallel} \right)$. This enters for each track associated to the secondary vertex.
- The track impact parameter significance of the first track exceeding the charm mass threshold in the transverse plane.

The distributions of these quantities are shown in Figure 4.10. The last observable in this list improves the suppression of charm jets. This is achieved by sorting the tracks in decreasing order according to their impact parameter significances and calculating the invariant mass for tracks 1 to n. The n'th track is the one which causes the invariant mass to exceed the threshold of 1.5 GeV/c^2 which is motivated by the mass of charm hadrons considering only charged particles. The impact parameter significance of the track moving the n-track mass above this threshold is used in the discriminator. For charm jets, this value is expected to be small, because this track does not come from a charm hadron decay and therefore not from a particle with a significant flight distance. For b-jets, however, this value is expected to be larger, because it descends from a b-hadron decay.

For the second category, the "PseudoVertex", most of these variables can still be used except for the distance between primary and secondary vertex, because the spatial position of the pseudo vertex is not fitted. The distributions look similar as in the "RecoVertex" case, but the separation power is reduced.

For the third category without a vertex, none of these variables in the list can be used and only the signed two-dimensional impact parameter significances of the tracks can be incorporated into the discriminator.

All these variables are combined into one single discriminator by the following likelihood function:

$$L^{b,c,q} = f^{b,c,q}(\alpha) \times \prod_i f_\alpha^{b,c,q}(x_i) \tag{4.1}$$

where α denotes the vertex category ($\alpha = 1, 2, 3$), x_i are the individual variables, q stands for light flavour jets including gluons, c or b refers to charm or b-jets, respectively. $f^{b,c,q}(\alpha)$

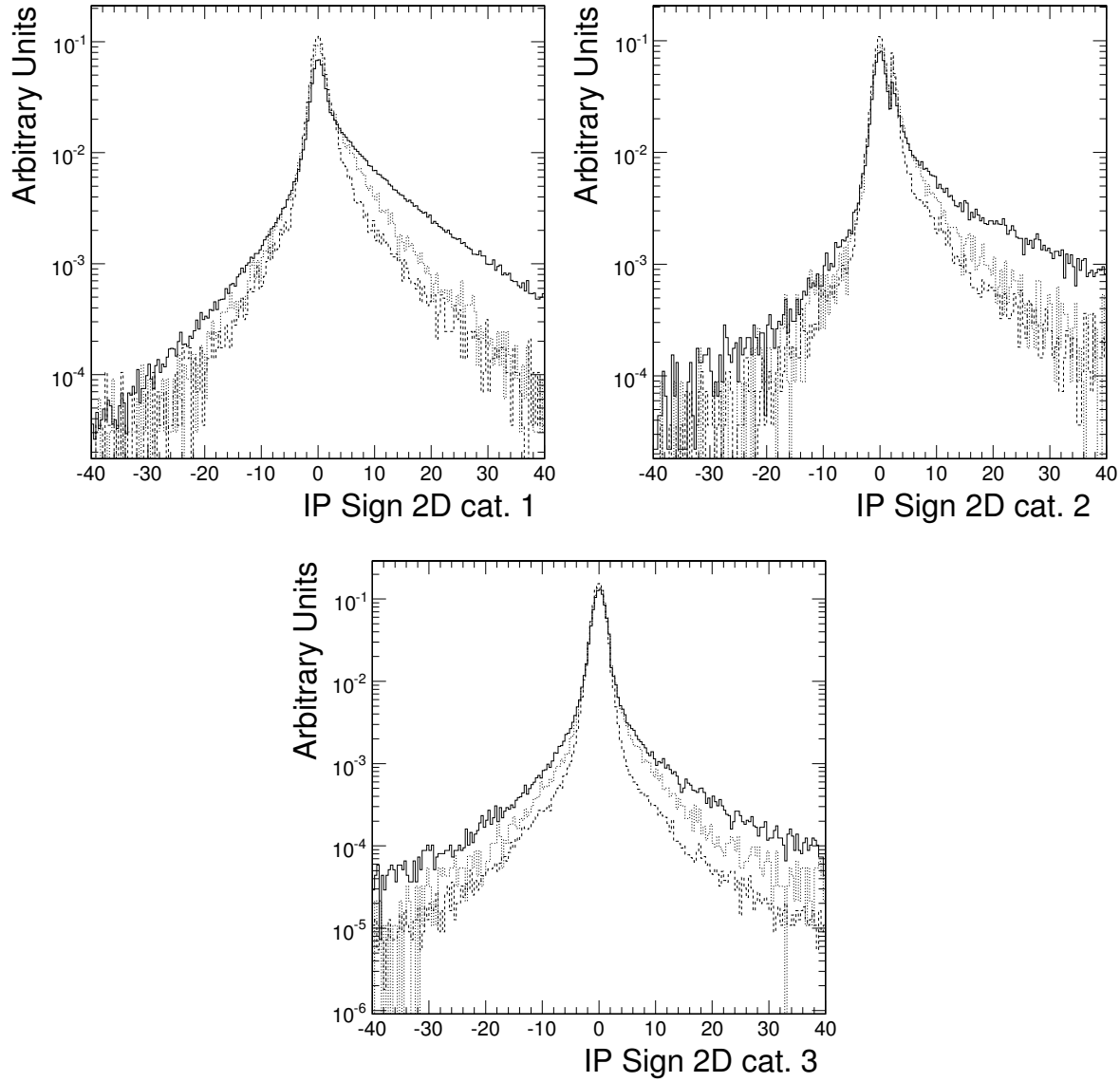

Figure 4.9: Signed transverse impact parameter significance for tracks in b-jets (solid), c-jets (dotted) and light flavour (uds-) jets (dashed) for the three different vertex categories. The second peak in category two is due to the selection criterion for tracks forming the pseudo vertex. These tracks are required to have a signed transverse impact parameter significance of more than two, leading to a cumulation of events having this kind of tracks.

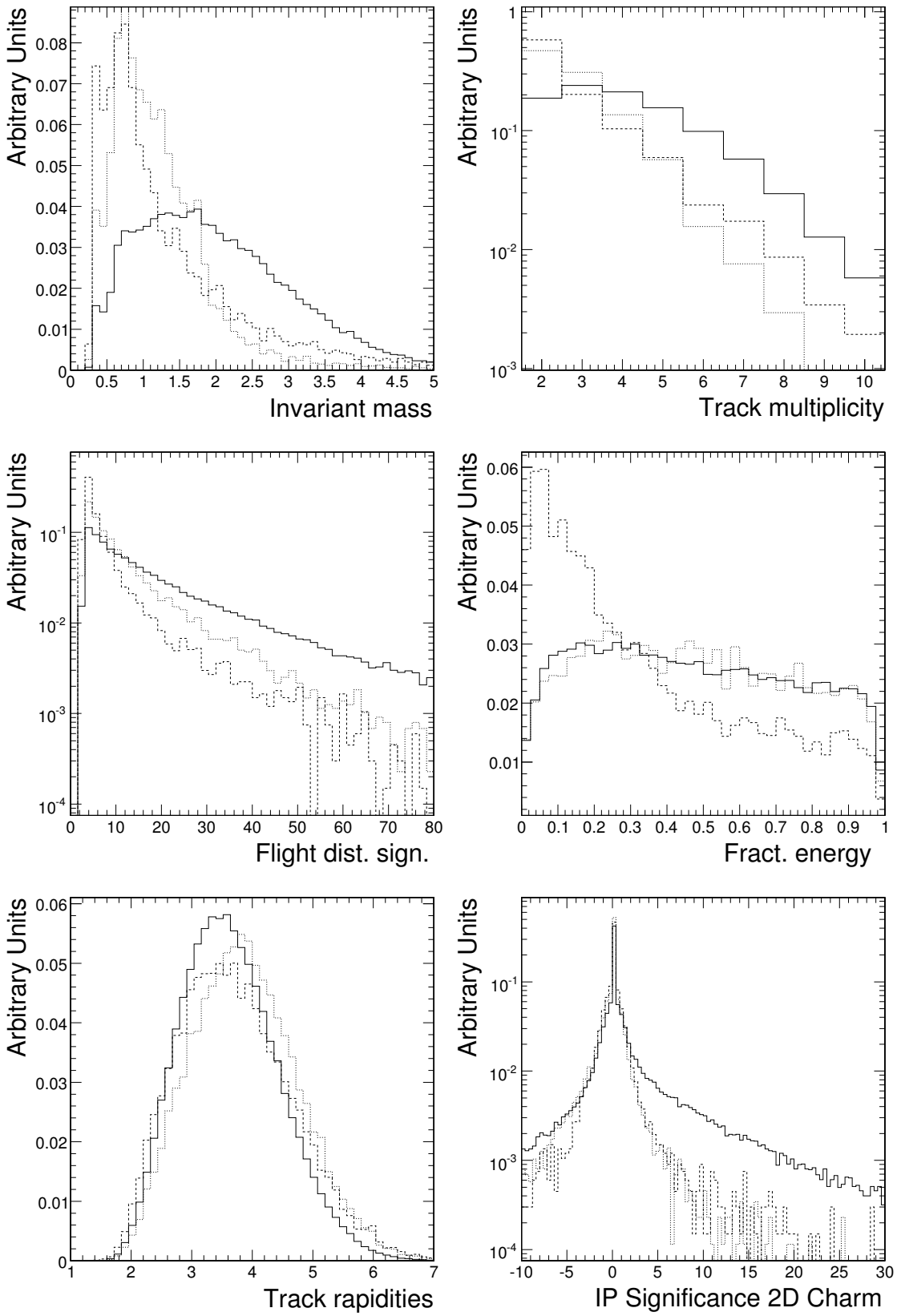

Figure 4.10: Observables used for the calculation of the b-tagging discriminator in the "RecoVertex" category. The solid lines refer to b-jets, the dotted line to c-jets and the dashed line to light flavour (uds-) jets.

is the probability for flavour b, c, q to belong to category α, while $f_\alpha^{b,c,q}(x_i)$ is the probability density function for category α and variable b, c, q, e.g. the impact parameter distribution.

The final discriminator d is then calculated by:

$$d = f_{BG}(c) \times \frac{L^b}{L^b + L^c} + f_{BG}(q) \times \frac{L^b}{L^b + L^q}, \tag{4.2}$$

where $f_{BG}(c)$ and $f_{BG}(q)$ are the a priori probabilities for the c- and q- content in non-b-jets, i.e. $f_{BG}(c) + f_{BG}(q) = 1$. The distribution of this discriminator d is shown in Figure 4.11 for the different jet flavours. It is visible that the discriminator gives a good separation

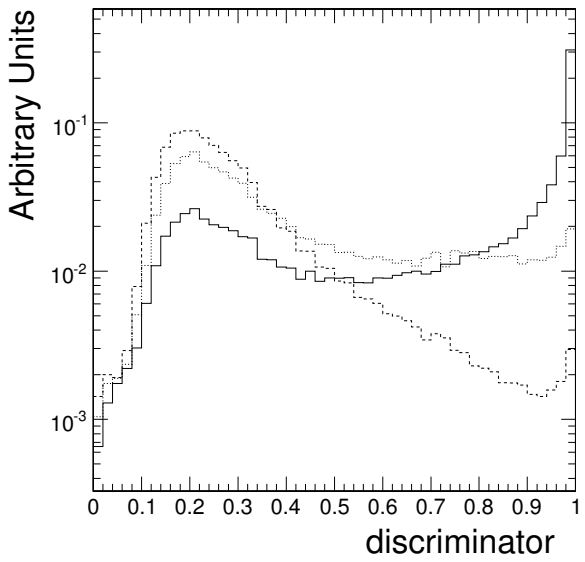

Figure 4.11: Distribution of the b-tagging discriminator for b-jets (solid line), c-jets (dotted line) and light flavour (uds-) jets (dashed line).

between the jets of the different flavours. This likelihood procedure might not be optimal since there are correlations between the observables that are not reflected in the likelihood functions. A possible solution might be a neural network, that automatically takes care of these correlations and could give a few percent improved performance, but this technique has not been implemented yet.

The choice of the cut on this b-tagging discriminator determines the tagging efficiency and misidentification rate. These rates in dependence on the discriminator cut are shown in Figure 4.12. Furthermore, the misidentification rate versus the b-tagging efficiency is shown in Figure 4.13 for the two definitions ("physics" or "algorithmic" definition). It is visible that the gluon misidentification rate is much worse in case of the "physics" definition, because of the occurence of gluon splitting. In the "algorithmic" definition the gluon and u,d,s misidentification rates are almost equal because these jets have similar behaviour in many respects.

Figure 4.13 shows the overall performance for a $t\bar{t}2j$ sample with its characteristical jet distributions which covers a large p_t range. However, the performance strongly depends on the momentum and direction of the jet. This is displayed in Figure 4.14, where the

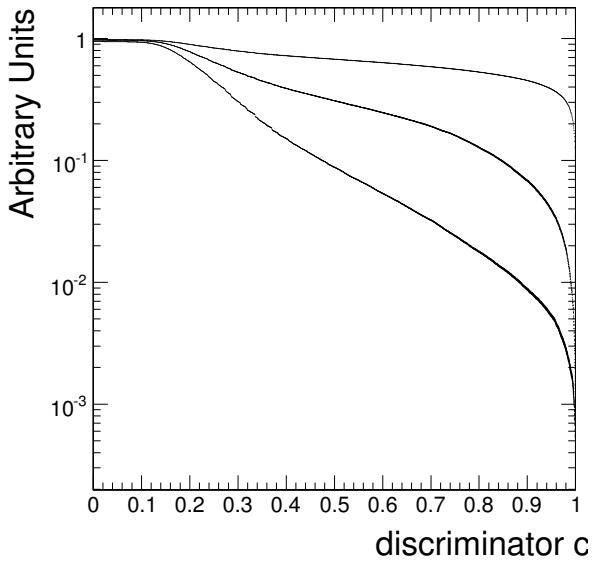

Figure 4.12: b-tagging efficiencies and misidentification rates in dependence on the cut on the b-tagging discriminator. The upper line shows the b-efficiency, the line in the middle shows the charm-efficiency and the lower line, the light flavour efficiency.

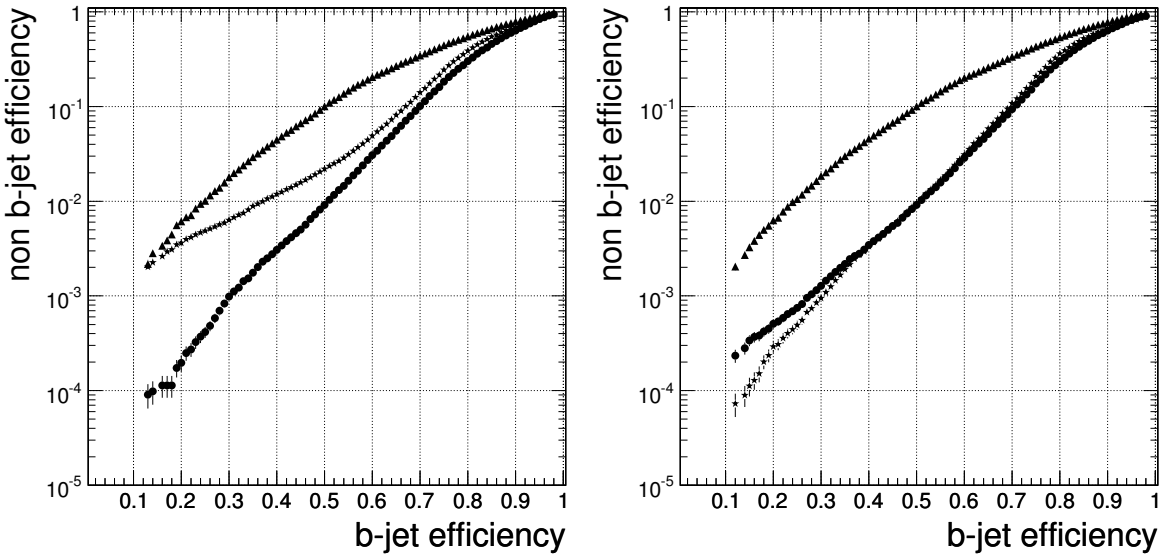

Figure 4.13: Mistagging rates versus b-tagging efficiency. The triangles show the charm misidentification rate, the stars represent gluons, while circles refer to the uds-misidentification rate. On the left side: "physics" definition and on the right side: "algorithmic" definition. The plots are obtained for a $t\bar{t}$2j sample with a minimal jet p_t of 20 GeV/c and $|\eta| < 2.4$.

misidentification rates in dependence on the transverse momentum of the jets for a fixed b-tagging efficiency of 50% are presented. The right plot in this Figure shows the dependence on the pseudo rapidity η. For low p_t values, the performance decreases because of a worse track

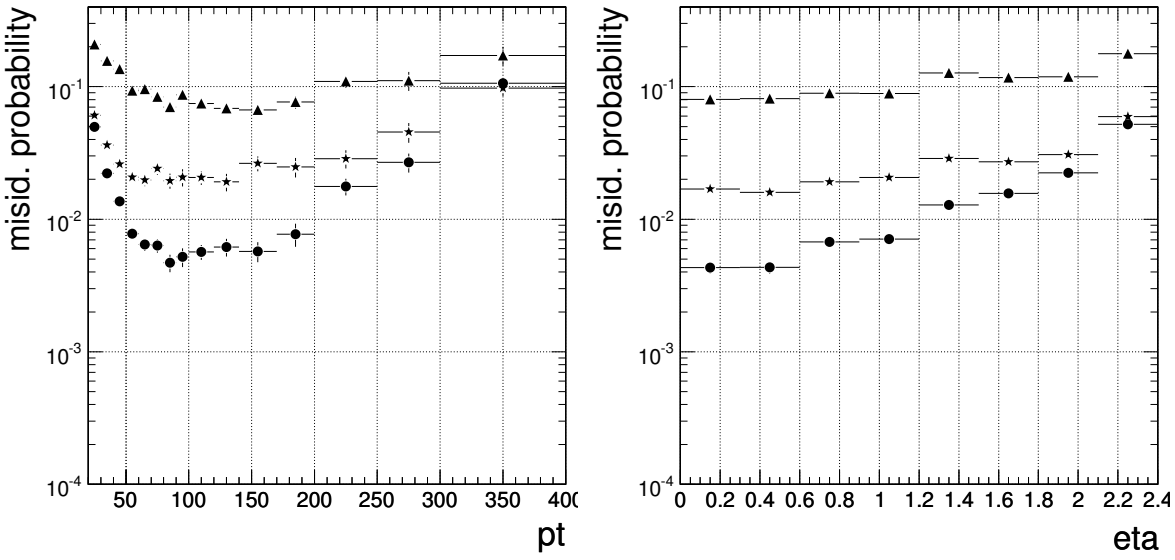

Figure 4.14: Mistagging probability at a fixed b-efficiency of 50% in dependence on p_t (left plot) and $|\eta|$ (right plot) in the "physics" definition. The triangles refer to charm jets, stars to gluons and circles to to uds-jets.

resolution due to an increase in multiple scattering which deteriorates the separation between primary and secondary vertices. For high p_t values, the rate of gluons splitting into heavy quarks increases and the track multiplicity from fragmentation increases which leads to a more difficult pattern recognition in dense jets. Therefore, the optimal b-tagging performance is reached for p_t values between 60 and 90 GeV/c in the central region of the detector. For larger values of the pseudo rapidity η, the performance degrades, because of larger amounts of material that have to be traversed and because of a reduced detector resolution.

4.2.7 Improvements in b-Tagging

The standard algorithm for b-jet identification is the "combined" b-tagging, described in Section 4.2.6, that combines various lifetime based track and vertex properties. This is the algorithm that has been used in the original publication [2]. Since b-tagging is one of the crucial components of the $t\bar{t}H$ analysis, two of the most promising enhancements of the b-tagging performance have been investigated. The first is an improved secondary vertex finding algorithm, namely the "Tertiary Vertex Track Finder" [50]. This algorithm exploits the fact that a b-hadron decay chain does not only contain secondary vertices but also tertiary vertices from charm decays. It takes care of an improved treatment of tracks from tertiary vertices as described in [50, 51]. The improvement of the b-tagging performance due to this algorithm is shown in Figures 4.17.

The second improvement is the combination with soft lepton tagging algorithms [52]. These algorithms make use of the property of b-hadrons to decay into electrons or muons in about 20% of the cases for each lepton family, counting electrons and muons only. The

presence of a lepton in a jet, together with other properties like impact parameter significance, angular distances between jet and lepton, and ratio of lepton momentum to jet energy, are indications for b-decays. These properties are combined into a discriminating variable using a neural network. The distributions of the discriminators of the "combined" b-tagging algorithm and the two lepton tagging algorithms are shown in Figure 4.15.

The two-dimensional performance plots for the "combined" b-tagging have been shown in Figure 4.13. For comparison, the performance of the soft muon tagging algorithm is displayed in Figure 4.16.

Obviously, the "combined" b-tagging gives the best separation between b- and non-b-jets. In the cases where a muon is found in the jet, also the soft muon tagging algorithm has a

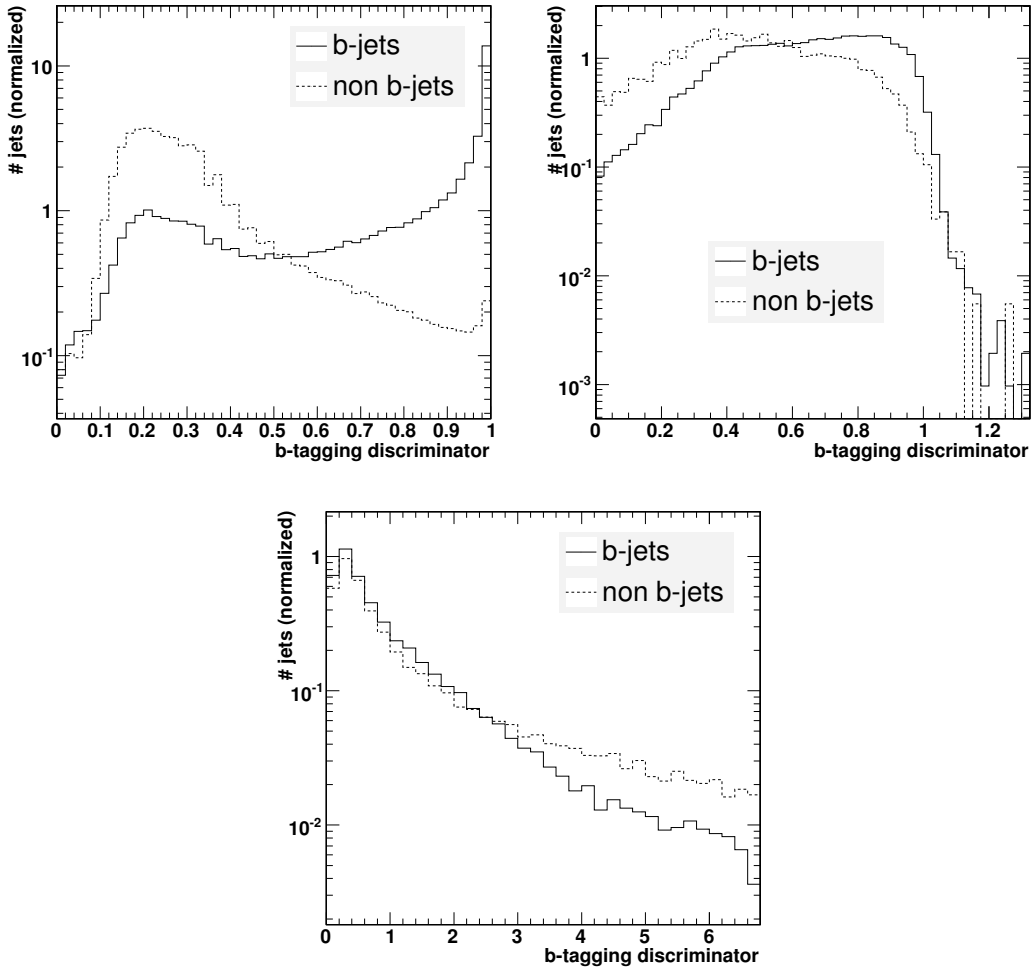

Figure 4.15: Distribution of the b-tagging discriminators for b-jets (solid lines) and non-b-jets (dashed lines). On the top left: "combined" algorithm; Top right: Soft muon algorithm; Bottom: Soft electron algorithm. The semileptonic $t\bar{t}H$ data sample has been used and the jets are required to have $p_t > 20$ GeV/c and $\eta < 2.4$. The distributions for soft lepton algorithms are only shown in the case where a lepton is found in the jet (i.e. 20% of the jets per lepton family).

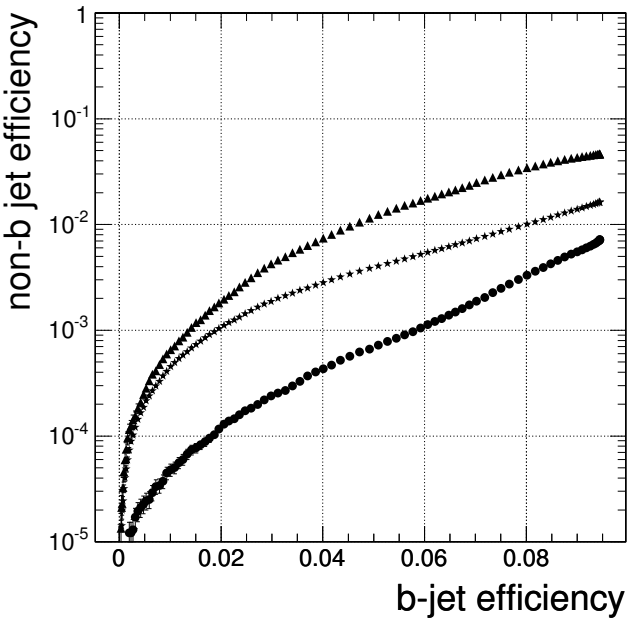

Figure 4.16: Mistagging rates versus b-tagging efficiency for the soft muon b-tagging algorithm. Triangles refer to c-jets, circles to uds-jets and stars to g-jets. The "physics" definition of the true jet flavour has been used. The results have been obtained with fully leptonic and semileptonic $t\bar{t}$ events and QCD events. [41]

significant separation power. The soft electron tagging algorithm does not have a convincing discriminating power, especially not around the peak between 0 and 1. Therefore, the soft electron tagging algorithm is not used in the analysis. Omitting the electron algorithm gives another advantage: since the "combined" algorithm and the lepton tagging algorithms have some correlation, it is beneficial to combine these two algorithms using a multi-dimensional likelihood ratio. Due to the limited amount of Monte Carlo statistics, two dimensions are the maximum. Therefore, only the soft muon tagging algorithm is combined with the "combined" algorithm, using the usual method of calculating a likelihood ratio according to Equation 5.1.

The resulting performance of the "super-combined" algorithm is shown in Figures 4.17. These Figures show the behaviour of the light flavour and charm misidentification rate in dependence on the b-tagging efficiency. Only the most relevant range of the b-tagging efficiency between 50% and 70%, that is usually used in a typical analysis, is displayed. The improvement resulting from the application of the improved tertiary vertex track finder and from the combination with the soft lepton tag are shown separately. The performances are in agreement with [50, 51, 48], if the fact, that the algorithmic definition for the true jet flavour has been used in all diagrams of this section and that light flavour jets and gluon jets are not treated separately in the present plots, is taken into account.

It is visible that the improvements due to the improved tertiary vertex track finder and due to the soft lepton tagging algorithm are of the same order of magnitude, around 15%, in case of the light flavour misidentification rate. The charm misidentification rate shows an improvement between 10% and 3% depending on the b-efficiency working point.

The difference for these two jet flavours is due to the fact that the misidentification is

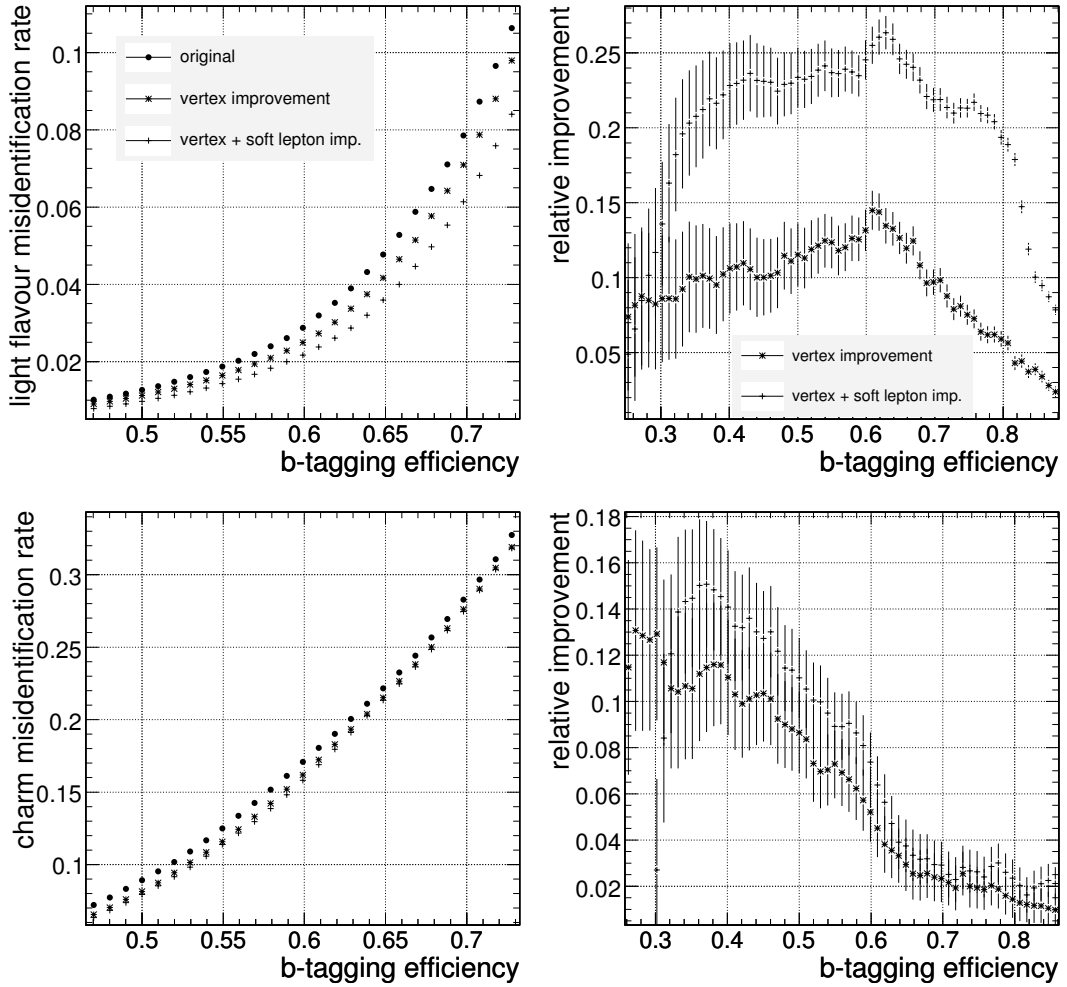

Figure 4.17: Improved Performance of the b-tagging algorithms. On the top left: Light flavour misidentification rate versus b-tagging efficiency for the standard algorithm, the improved tertiary vertex algorithm, and the combination with the soft muon algorithm. The plot on the top right corner shows the relative improvement (in %). The two plots on the bottom show the same, but for the charm misidentification rate. These plots have been obtained with the semileptonic $t\bar{t}H$ datasample and the jets are required to have $p_t > 20$ GeV/c and $\eta < 2.4$. The error bars indicate statistical errors arising from the finite size of Monte Carlo datasets. It should be noted that these plots have large bin-to-bin correlations, because each bin contains also the events of its next bin to the right, since these plots are obtained by applying an increasing cut on the b-tagging discriminator.

caused by different reasons. In case of charm, a real secondary vertex is present, because charm hadrons have a significant life time of the order of $\tau = 10^{-12}$ s and therefore a measurable flight distance with $c\tau \approx 300\mu$m. The charm decay behaves in a certain way like a short b-hadron decay. An improved secondary vertex finding algorithm will in principal not change this situation, but the enhanced inclusion of tracks from tertiary vertices increases also the charm suppression, especially in the high purity region, because charm decays do not provide tertiary vertices. Moreover, leptons are present also in charm decays, thus limiting the possible improvements from the lepton tag. Decays of light flavour hadrons do not have real secondary vertices or leptons and the misidentification is due to pure instrumental mismeasurements that can be improved by better methods of vertex and impact parameter determination.

The largest backgrounds in the $t\bar{t}H$ analysis are $t\bar{t}$ plus N light flavour jets, because of their large cross sections. The improvement in light flavour rejection of around 25% is therefore an important contribution to an improved discovery potential as discussed in the following sections. However, all improvements that rely on more complicated methods are subject to systematic errors.

4.3 Performance of Jet Reconstruction Algorithms

Several jet reconstruction algorithms and parameters in the CMS framework are discussed in Section 4.2.4. These algorithms and their corresponding configuration have been compared in order to determine the setup that gives the best performance. These studies have been carried out using final state generator particles as input for the jet finding. For this purpose, the "Iterative Cone" (IC) algorithm, the "inclusive k_T" (k_T) algorithm and the "Midpoint Cone" (MC) algorithm and their respective configuration parameters have been tested. Comparisons like this have the potential of becoming quite comprehensive, therefore the study has been carried out in a greater extent in conjunction with other channels as published in the proceedings of the 2005 "Les Houches" workshop [53]. In the following, a short summary of these studies will be given.

This study concentrates on the algorithmic task of clustering the input objects for the jet finding, and has to be understood from an analysis perspective. This means that the jet finding is considered to be optimal if the efficiency to reconstruct the complete kinematics of the primary quark event topology is maximized. This reconstruction efficiency is defined in terms of some quality citeria, the so called "quality markers", and has been determined for four different event topologies with two, four, six and eight primary quarks in the final state. In case of the $t\bar{t}H$ channel, there are six jets (primary quarks) in the final state for the semileptonic channel. The di-lepton and the all-hadron channels have four and eight jets, respectively. Therefore, this study covers all cases for $t\bar{t}H$.

The discussed "quality markers" are listed below, together with a short description of their purpose:

- Event Selection Efficiency "ϵ_s": This is the fraction of events that pass the two selection criteria, of a mimimum transverse jet energy of 20 GeV and a maximum pseudo-rapidity $|\eta| < 2.4$.

- Angular Distance between Jet and Parton "$Frac\ \alpha_{jp}^{max}$": For each jet, the ΔR distance α_{jp} to its primary parton is calculated and sorted in increasing order. This way, n α_{jp}

values are obtained, where α_{jp}^{max} is the largest one. To quantify the angular reconstruction performance of an event, the quality marker $Frac\ \alpha_{jp}^{max}$ is defined as the fraction of events with an α_{jp}^{max} value lower than 0.3.

- Energy Difference "$Frac\ \beta_{jp}^{max}$": The reconstructed energy of a jet is usually biased and has a broad resolution as shown in Figure 4.18. This kind of calibration curve can be

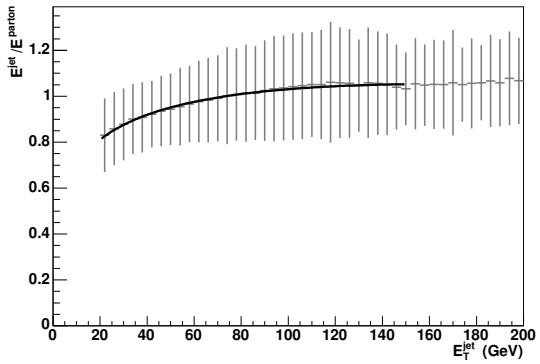

Figure 4.18: E^{jet}/E^{parton} vs. E_T^{jet} for the IC algorithm with a cone radius of 0.4 applied on a final state with four primary quarks. The vertical bars illustrate the resolution. For this plot, only well matched ($\alpha_{jp} < 0.3$) and non-overlapping jets were taken into account in order to determine the optimal energy resolution.

interpreted as an estimator for the expected reconstructed jet energy. The β_{jp} values are defined as the distance from the expected energy fraction (deduced from Figure 4.18) in units of standard deviations. Analogously to the case of the α_{jp} variable, the β_{jp}^{max} value is the largest one of these values and the quality marker $Frac\ \beta_{jp}^{max}$ is defined as the fraction of events with β_{jp}^{max} smaller than 2 (standard deviations).

- Combined Variable "$Frac(\alpha_{jp}^{max} + \beta_{jp}^{max})$": This quality maker is defined as the fraction of events in which both of the two previous criteria (α_{jp} and β_{jp}) are fulfilled. This means that both, energy and direction of the jet are well reconstructed. The left side of Figure 4.19 shows the correlation between the α_{jp}^{max} and β_{jp}^{max} variables. The fraction of events inside the rectangular area in this plot, where both variables are passing the criteria, is defined as the "$Frac(\alpha_{jp}^{max} + \beta_{jp}^{max})$" quality marker. As an illustration of the power of this variable to identify well reconstructed events, the hadronic top quark mass for "good" and "bad" events is shown on the right side of Figure 4.19.

- Overall quality marker "*FracGood*": The fraction of selected and well reconstructed events, i.e. ϵ_s multiplied by $Frac(\alpha_{jp}^{max} + \beta_{jp}^{max})$ is defined as *FracGood*. This gives an estimate of the efficiency to reconstruct the complete primary quark kinematics of an event.

Although the last variable gives a powerful indication of a reasonable jet definition, it is still important to also consider the partial information of the individual quality markers, depending on the priorities of the specific analysis.

Some results for the IC algorithm are shown in Figure 4.20. The left plot in this figure shows the *FracGood* variable for four different jet multiplicities. It is visible that the overall

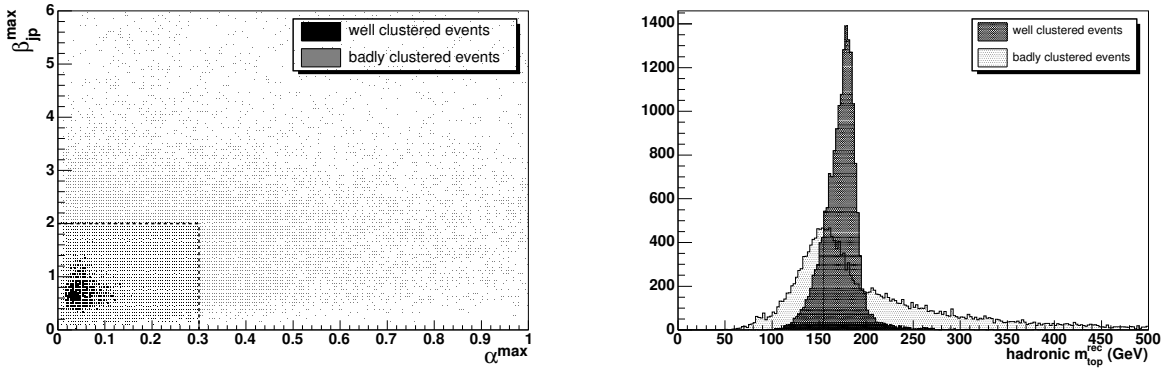

Figure 4.19: On the left: Correlation between the α_{jp}^{max} and β_{jp}^{max} variables. On the right: Hadronic top quark mass for well and badly reconstructed events, according to the combined variable. Both plots have been obtained for the IC algorithm with a cone radius of 0.4.

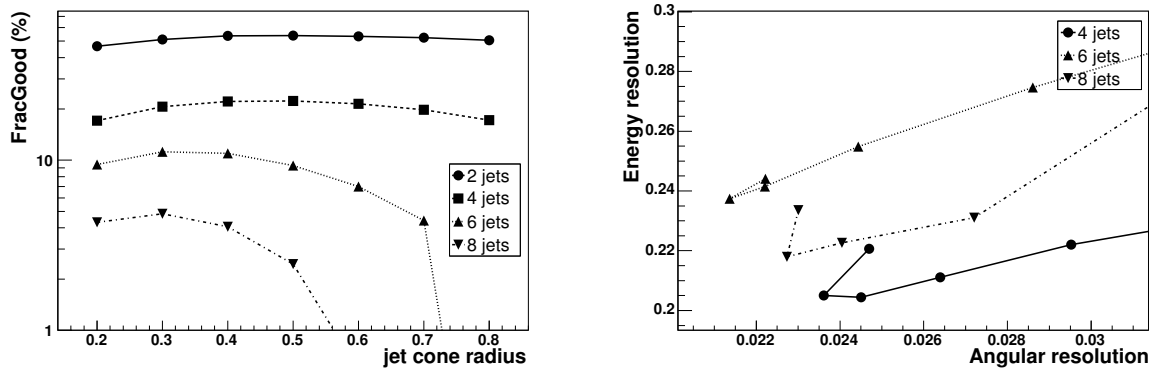

Figure 4.20: On the left: *FracGood* variable for four different jet multiplicities depending on the cone radius (IC algorithm). On the right: Relative energy resolution versus angular resolution. The markers of one type represent one distinct jet multiplicity. The values on the top left end of each line correspond to cone radii of 0.2. The cone radii are increasing with steps of 0.1 along the line to the top right corner. The energy resolution is defined as the RMS divided by the mean value of the E^{jet}/E^{quark} distribution, and the angular resolution is defined by the width of a gaussian fit to the symmetrized ΔR distribution.

efficiency is much smaller for events with high jet multiplicities, but also that a smaller cone radius performs better in these events. The resolutions in energy and direction are shown on the right side of Figure 4.20. As visible in this plot, the resolutions are approximately optimal in the case where also the *FracGood* variable is optimal. The same plots for the k_T algorithm are shown in Figure 4.21. The situation is more complicated in the case of the Midpoint Cone (MC) algorithm. This algorithm is more complicated to configure because it has two additional parameters. The dependence on its cone radius is shown on the left side of Figure 4.22.

Surprisingly, almost no dependence on the shared energy fraction threshold for merging parameter has been found for this algorithm, as shown on the right side of Figure 4.22. Furthermore, this algorithm performs not much better than the IC algorithm, in contrast to the expectations. This might be due to the implemenation in the CMS framework which was not yet mature enough at the time of this study. A new investigation of the performance of this algorithm should be performed as soon as it becomes available in the new CMS

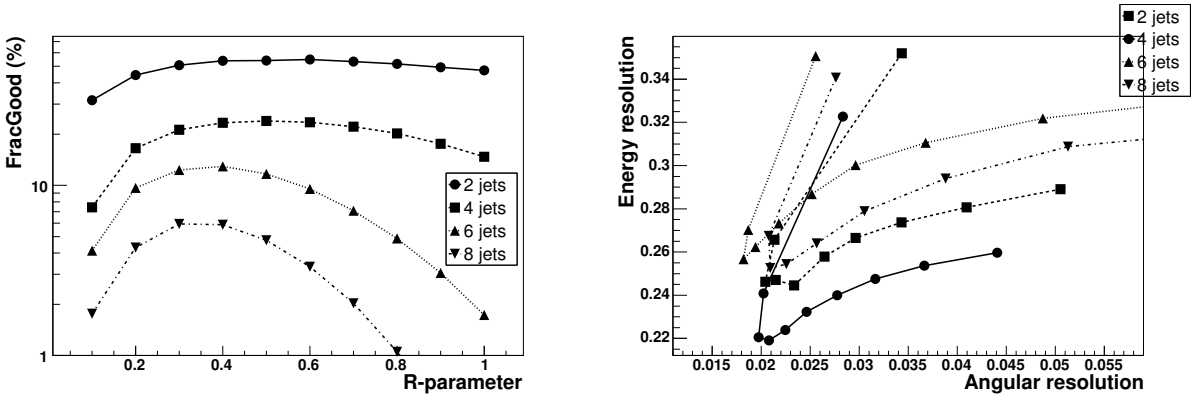

Figure 4.21: The same plots as in Figure 4.20 but for the inclusive k_T algorithm and variation of its R-parameter. The top left values in the right plot correspond to R-parameter values of 0.1 and are increasing with steps of 0.1 along the line.

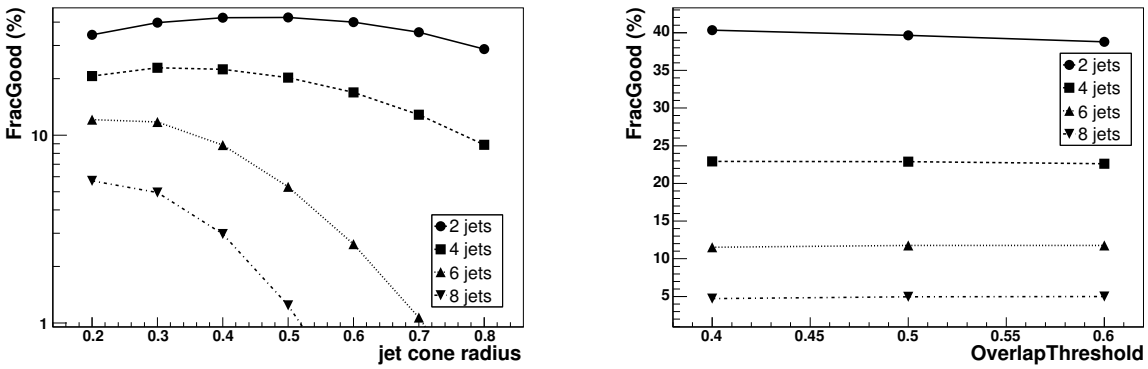

Figure 4.22: On the left: *FracGood* versus the cone radius for a merging threshold of 0.5 and a cone area fraction of 0.25. On the right: *FracGood* versus the threshold for merging for a cone radius of 0.3 and an area fraction of 0.25 (MC algorithm).

software framework, because the experience of other experiments with this algorihm are very promising. In addition, the CPU time consumption of this algorithm has been found to be disproportionally high, therefore the MC algorithm has not been considered for the $t\bar{t}H$ analysis.

The conclusion of these studies is, that, on Monte Carlo level, the Iterative Cone algorithm with a cone radius of 0.4 is a good choice for the semileptonic channel of the $t\bar{t}H$ analysis with six jets in the final state. As soon as reconstructed calorimeter towers are used as input, the cone radius has to be increased due to effects induced by the magnetic field and the resolution in the calorimeter. More details on the particular jet finding setup for the $t\bar{t}H$ analysis are given in Section 5.3.4.

4.4 Fast Detector Simulation and Reconstruction

This Section gives a short overview over the components of the fast simulation program FAMOS, as it has been published in the CMS PTDR [41]. A more detailed description for the part of the fast b-tagging simulation (implemented by the author of this thesis) is given in Section 4.4.1.

The input to FAMOS is a list of generated particles that are propagated through the magnetic field and that are allowed to decay according to their branching ratios. The simulation of the interaction with the detector uses the following processes:

- Electron bremsstrahlung
- Photon conversion
- Charged particle energy loss by ionization
- Charged particle multiple scattering
- Electron, photon, hadron showering

The first four processes are applied in the tracking detector, while the last one in the list is performed in the electromagnetic or hadron calorimeter, respectively. For the muon simulation, a parametrization of the resolutions and efficiencies is applied. The output of FAMOS are higher level objects, like jets, b-tagged jets, muons, electrons, etc. This way, the CPU time to simulate one event can be reduced by up to 3 orders of magnitudes compared to the full simulation and reconstruction.

The tracker geometry in FAMOS uses a simplified model consisting of nested cylinders as shown on the left side of Figure 4.23 in comparison with the geometry in the full detector simulation. The positions of simulated hits in FAMOS are smeared by a gaussian distribution and are turned into reconstructed hits with a certain efficiency. Of special interest for the b-tagging simulation is the impact parameter of which a detailed comparison is shown in the next section. In order to precisely reproduce the tracking performance of the full simulation, the gaussian resolution is parameterized in the silicon tracker using constants for each space dimension. For the pixel tracker a parameterization according to the pixel cluster size and the incident angle with respect to the layer is used. No pattern recognition is applied in FAMOS. Instead, the hits belonging to a track produced by a charged particle, are fit using the same fitting algorithms as in the complete reconstruction. This procedure saves an enormous

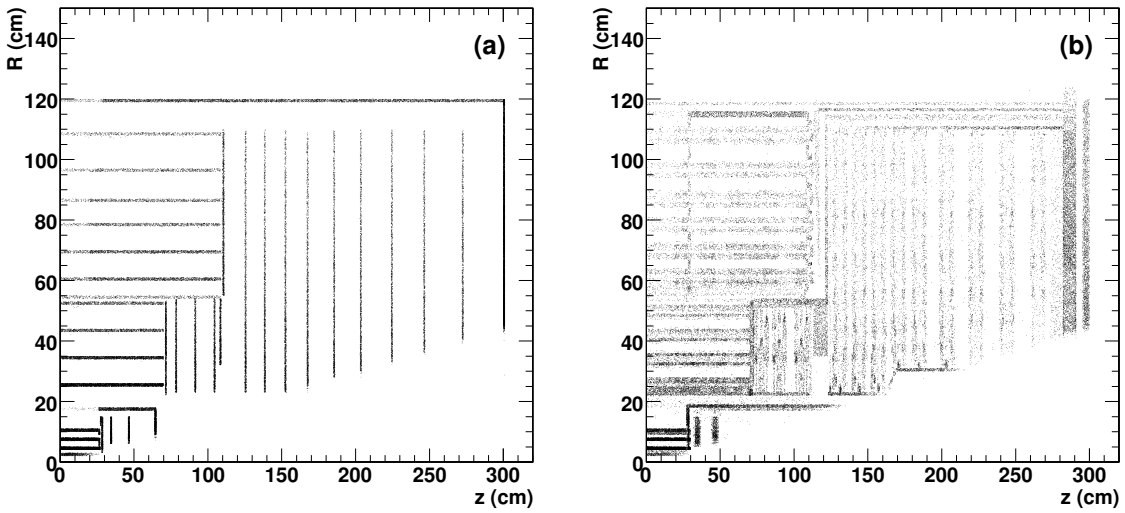

Figure 4.23: Radiography of a quarter of the CMS tracker obtained from vertices of converted photons for the geometry in fast simulation (left) and full simulation (right). [41]

amount of CPU time, but the inclusion of fake hits is not reproduced properly, especially not in a high luminosity environment.

The simulation of electrons uses a shower parametrization under the assumption of a homogeneous material. The energy distribution is then placed into the crystal geometry followed by the simulation of effects like front and rear leakage, energy loss in gaps and effects induced by the magnetic field. Photons are first converted in the ECAL material according to the number of traversed radiation lengths. The resulting e^+e^- pairs are then simulated in the shower evolution as described above.

The calorimeter response to hadrons is parametrized by a gaussian distribution of which the mean value and width depend on η and the energy. These values are taken from the full detector simulation by interpolating between the fully simulated results for discrete p_t values of pions. This smeared energy is then distributed in the calorimeters using parametrized shower profiles.

Muons are propagated in detail through the tracker. The response of the calorimeters is parametrized in a similar way as for pions. The muon chamber response is simply parametrized to reproduce the resolutions and efficiencies of the full simulation.

4.4.1 b-Tagging in FAMOS

The implementation of b-tagging in FAMOS exploits the reusability of reconstruction algorithms in the COBRA framework. The algorithms have originally been developed for the physics reconstruction software ORCA. The adaptation to FAMOS required only small changes that have been connected to requests for elements of the detector geometry that is not fully available in FAMOS, since it applies a simplified detector model. Therefore, the identical b-tagging algorithms as in ORCA are used. The implementation had to be realized in the form of a wrapper that gets the fast tracks and fast jets as input which are passed to the b-tagging algorithms. Therefore, the agreement of the b-tagging performance depends on the

simulation quality of the input objetcts, i.e. jets, tracks and vertices. The secondary vertex reconstruction algorithm is the same as in in the full reconstruction, but it uses fast tracks as input. A dedicated fast vertex reconstruction does not exist. Also the jet reconstruction algorithms are the same as in ORCA using fast calorimeter towers as input.

The most important observables that are used in the "combined" b-tagging algorithm, described in Section 4.2.6, are compared in the following. The same $t\bar{t}2j$ data sample as in Section 4.2.6 has been used. Figure 4.24 shows the distribution of the vertex categories for the different jet flavours in comparison between ORCA and FAMOS. It is visible that the first category has a systematically higher population for b- and charm-jets. Apparently, the reconstruction of the secondary vertex is more efficient in FAMOS than in ORCA. This is due to the cleaner track environment in FAMOS which has less fake hit inclusions which deteriorate the vertex measurement. This fact alone already indicates a better b-tagging performance in FAMOS compared to ORCA.

Besides the secondary vertex, another observable of major importance is the transverse impact parameter significance, which is shown in Figure 4.25. A fair agreement even in the

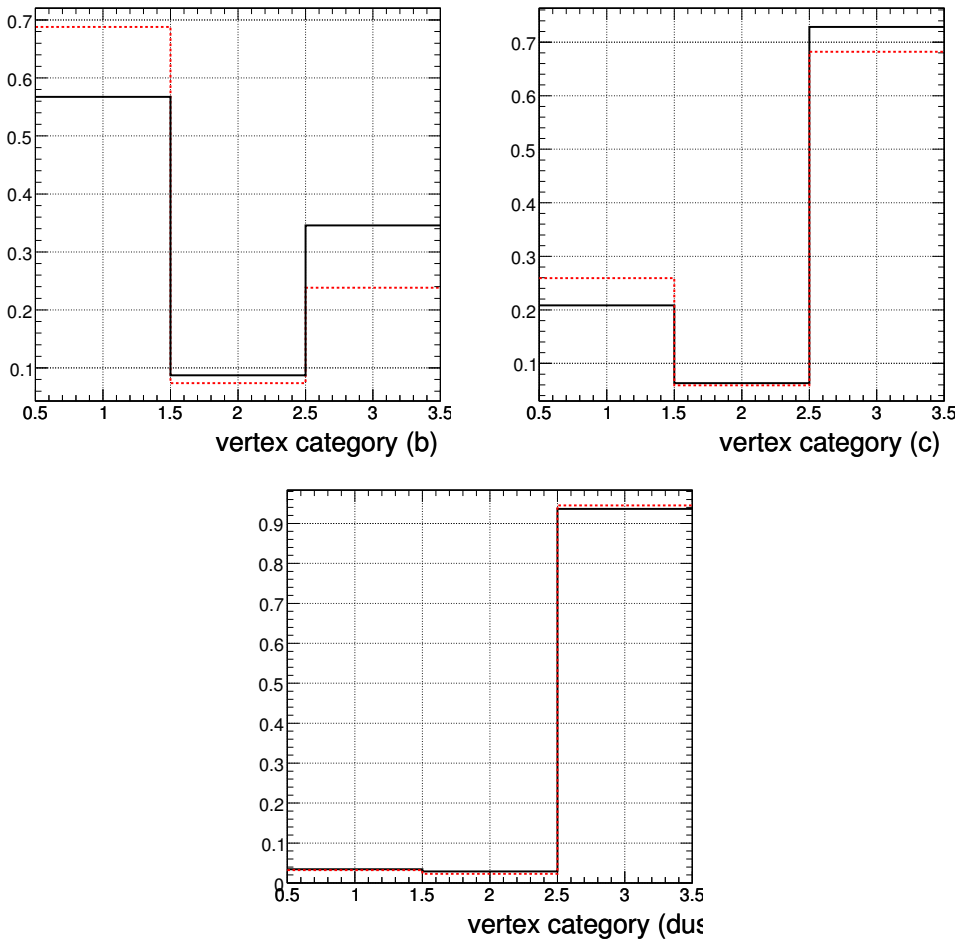

Figure 4.24: Distribution of the vertex categories for ORCA (solid line) and FAMOS (dashed line). The top left plot shows b-jets only, the top right c-jets and the plot on the bottom displays uds- and gluon-jets.

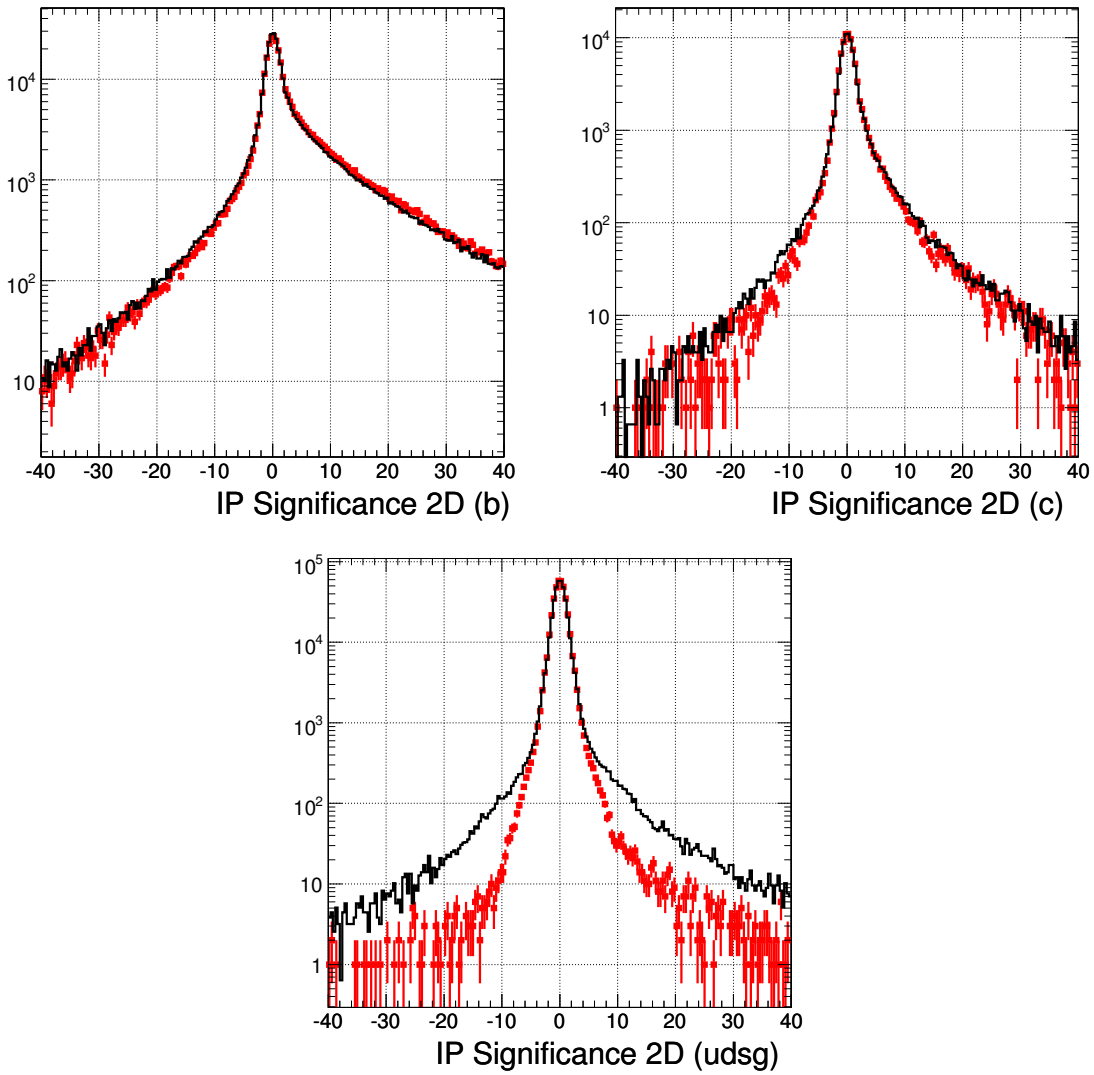

Figure 4.25: Signed Transverse Impact Parameter Significance of all tracks in the jet for ORCA (black solid line) and FAMOS (grey markers). The top left plot shows b-jets only, the top right c-jets and the plot on the bottom displays uds- and gluon-jets.

center of the distribution is found for b- and c-jets. A small systematic shift to positive values is visible in the top left plot. This effect is due to slightly smaller errors of the impact parameters in FAMOS leading to larger significances on the right side of the tail. The left side of the tail, corresponds to cases where the decay vertex seems to be located on the wrong side of the primary vertex caused by badly measured or fake tracks which are less in FAMOS, thus the distribution is shifted to the right also in the negative tail. For light flavour jets (uds- and gluon), however, the distribution in the tails is much narrower in the FAMOS case. Obviously, the impact parameter measurement in ORCA is worse than in FAMOS which is related to effects like multiple scattering or the distribution of material. Again, this behaviour indicates that the b-tagging performance in FAMOS should be better than that in ORCA.

The comparison of the remaining variables used for the calculation of the b-tagging discriminator according to Section 4.2.6, is given in Appendix A. The conclusion is that some of these variables show major discrepancies due to the number of tracks at the primary vertex, while the observables at the secondary vertex are in better agreement, but still show some problems. All deviations without exceptions indicate a too optimistic performance in FAMOS.

The resulting distributions of the b-tagging discriminators for the various jet flavours are compared in Figure 4.26. In the case of b-jets the distribution shows a deficit at a discriminator value of 0.2 and an excess at 0.8 for FAMOS. This behaviour is mostly due to the distribution of the vertex categories, which also shows a deficit at category three and an excess at category one. The b-tagging discriminator has a different behaviour among the vertex categories, in fact, it has a peak at 0.2 in category three and between 0.8 and 1 in category one as shown in Figure 5 in [48]. A similar argumentation holds in the case of c-jets. The third diagram in Figure 4.26 shows a deficit in the range between 0.7 and 1, but a surprising excess in the last bin which can be explained by gluons splitting into $b\bar{b}$ faking real b-jets.

The distribution of the discriminator confirms the expectation that the distribution is shifted to higher values in case of b- and c-jets and to lower values in case of light flavour jets. The final performance plots in terms of misidentification rates versus b-tagging efficiencies are given in Figure 4.27. Obviously, the performances disagree by some factors. In case of light flavour (uds) jets, the misidentification rate is 1% in ORCA and 0.25% in FAMOS at a b-efficiency of 50%. In case of charm jets, the misidentification rate is 10% and 5% respectively at the same working point. As already expected, the difference is too large to be useful for most analyses like the $t\bar{t}H$ analysis, for instance. It is doubtful that an approach like this, using the algorithms of the full reconstruction tools, could lead to a good agreement between full and fast simulation, since the tagging rates strongly depend on a very fine tuning of the input objects. Small changes in the fast parametrization of the tracks might have large impact on the b-tagging results. Therefore, a number of alternative approaches should be investigated. For instance, a parameterization of the tagging efficiencies themselves, depending on p_t, $|\eta|$, and the jet density, based on the true flavour of the jet and the efficiency values from the full simulation, could give a better agreement. This method might be problematic since the tagging efficiencies depend on a multitude of kinematical and topological parameters.

Another possibility is motivated by Figures 4.24 and 4.26 which show that the differences in the distributions of the discriminator are mostly due to the respective contributions of the three vertex categories as discussed above. Hence, a well defined fraction of events could be removed from category one and added to category three according to the expected population of the various categories. Methods like this should probably be applied as long as the reasons for the discrepancies discussed in Appendix A are not fully understood.

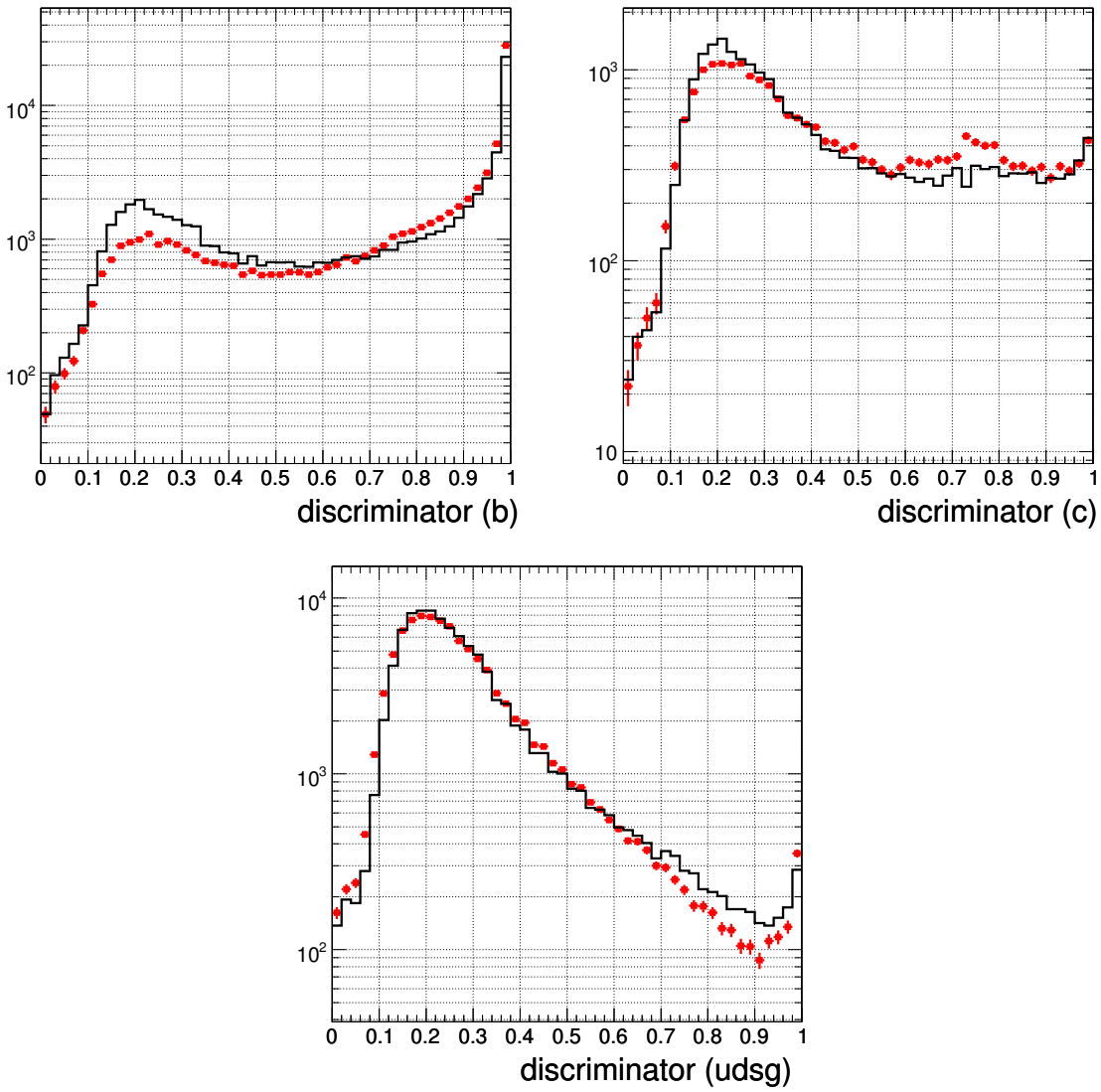

Figure 4.26: Distribution of the b-tagging discriminator for ORCA (black solid line) and FAMOS (grey markers). The top left plot shows b-jets only, the top right c-jets and the plot on the bottom displays uds- and gluon-jets.

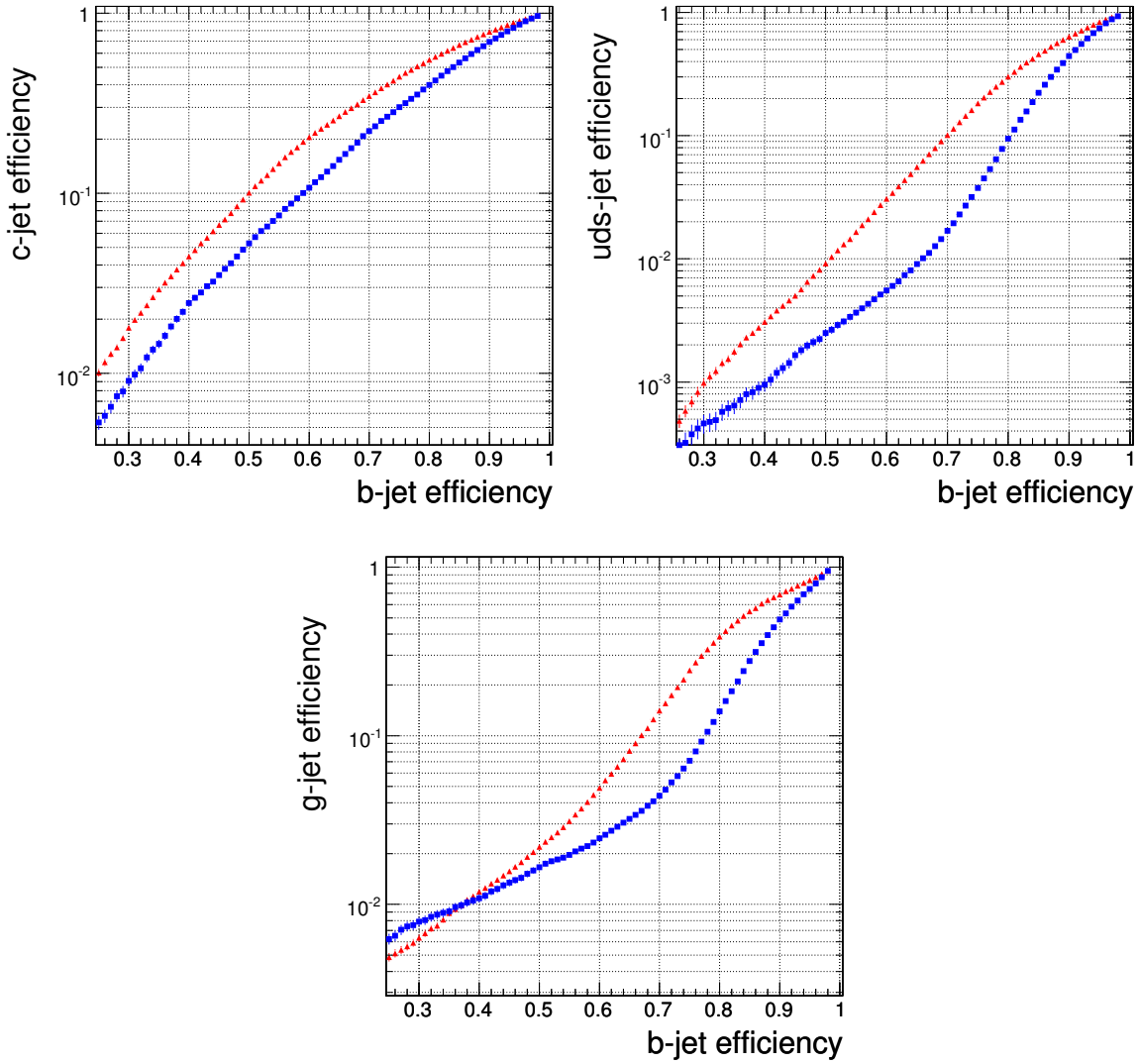

Figure 4.27: Misidentification rate versus b-tagging efficiency for ORCA (triangles) and FAMOS (boxes). The top left plot shows c-jets only, the top right dus-jets and the plot on the bottom displays gluon-jets. The plots are obtained using the "physics" definition of the true jet flavour as explained in Section 4.2.6.

4.5 PAX

The Physics Analysis Expert toolkit (PAX) is a collection of classes aimed to assist in the fourvector reconstruction step of a particle physics analysis. The project has been started at the University of Karlsruhe and has initially been introduced at the CHEP'03 conference [54] followed by other conferences [55, 56] and publications [57]. Meanwhile the development is distributed over a number of institutions, including the RWTH Aachen and the University of Hamburg, Germany.

The project has been motivated by experiences collected with previous analysis packages like H1PHAN [58] of the H1 experiment and ALPHA [59] of the ALEPH experiment, in particular:

- users have been able to quickly answer physics questions
- protection of the physics analysis code against changes in the detector reconstruction layer

Furthermore, PAX tries to face the challenge of dealing with the enormous event complexity of future hadron collider machins with up to 20 simultaneous collisions and large event sizes.

PAX is implemented in the C++ programming language following object oriented design principles. It provides a collection of container classes to manage event interpretations, fourvector arithmetics and the combinatorial task of reconstructing decay trees in different ways. Meanwhile, also a graphical user interface, which enables easy browsing of physics objects, has been developed.

4.5.1 PAX Class Structure

In order to enable fourvector arithmetics, it is desirable to use already well established fourvector implementations like the *TLorentzVector* of ROOT [60] or the *HepLorentzVector* of CLHEP [61]. As shown in Figure 4.28 the user has the choice of selecting either one as base class for the *PaxFourVector*. The *PaxFourVector* extends the functionality of the chosen base classes by some useful datamembers and methods, like charge, particle ID and relations to other PAX objects, like begin- and end-vertices. It has also the ability to store an arbitrary number of user-defined floating point values in its "user record". Another important instrument is the possibility to associate pointers of arbitrary type (e.g. detector object) with any PAX physics object. This way the information about the original detector information can be percolated through the whole analysis and can always be accessed.

A physical vertex is represented by the *PaxVertex*, which is designed in a similar way as the *PaxFourVector*. Instead of a Lorentz-Vector, the *PaxVertex* is based upon a three-vector. Also in this case, it is the choice of the user whether to use the *TVector3* of ROOT or the *Hep3Vector* of CLHEP.

In order to construct decay trees, the *PaxFourVector* and *PaxVertex* can be connected using the functionality of the *PaxRelationManager* as displayed in Figure 4.29. The relation manager follows the so called "Mediator" design pattern [62] which means that the relations are local in the sense that each object knows its related objects but there is no global map or directory of the relationships. A *PaxVertex* has incoming and outgoing fourvector-relations and each *PaxFourVector* has begin- and end-vertex relations, which facilitates the construction of arbitrary decay trees. The *PaxRelationManager* is also used to record an analysis "history".

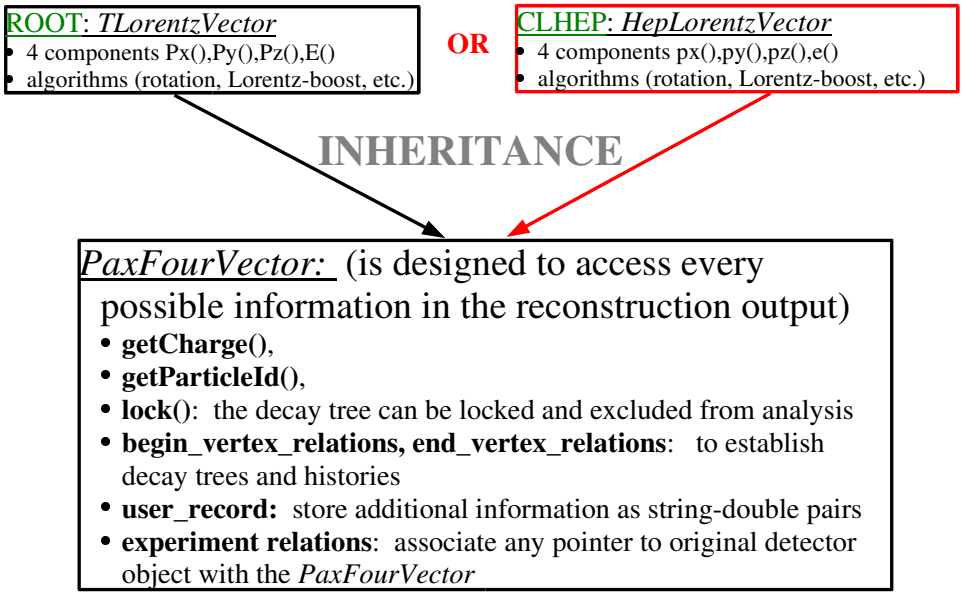

Figure 4.28: Inheritance diagram of the *PaxFourVector*. The functionality of the LorentzVector base class is extended by additional information and functionality.

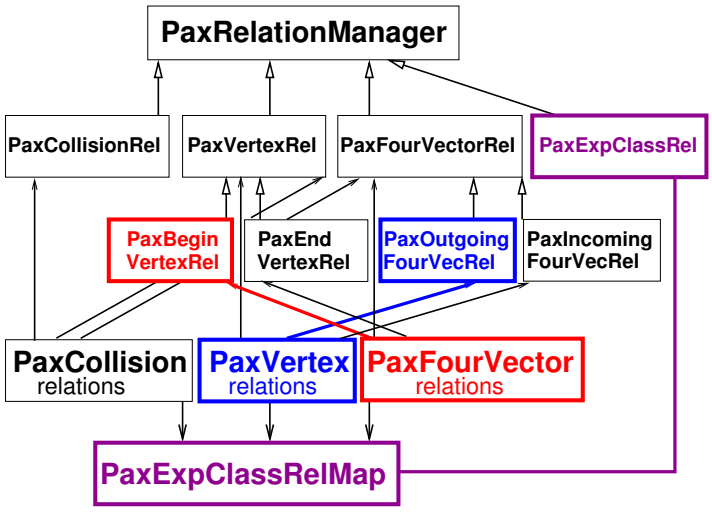

Figure 4.29: Structure of the *PaxRelationManager*. Each PAX object has a relations datamember which inherits from the *PaxRelationManager*.

This means that each copied object keeps a pointer to its original instance. This way, it is always possible to go back and ask for the original properties of an object which might have changed during the analysis.

All the PAX physics objects are being stored in the *PaxEventInterpret* as shown in Figure 4.30. This is a container class that represents one particular interpretation possibility of an event. It takes over the object ownership as soon as the object has been registered into

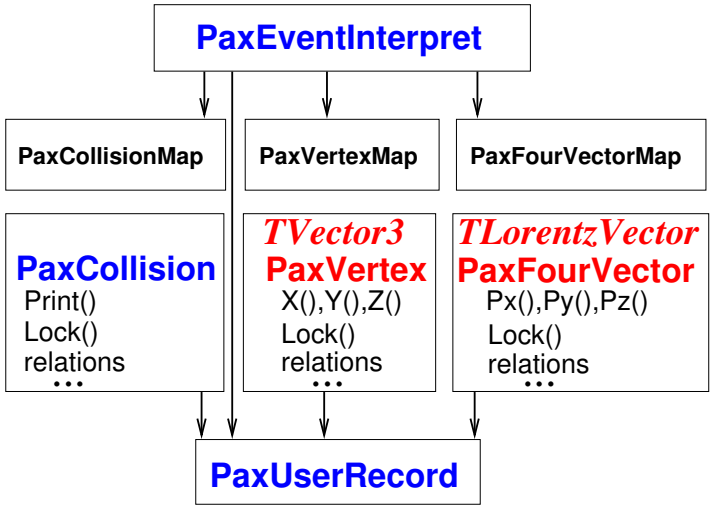

Figure 4.30: Structure of the *PaxEventInterpret* class. Each interpretation possibility is stored in a separate instance of *PaxEventInterpret*.

it. In order to advance the analysis in different directions and to test various hypotheses, the *PaxEventInterpret* can be copied. A copy of a *PaxEventInterpret* is a deep copy, which means that all registered objects are duplicated. Furthermore, the relations of the copied objects are set up correctly in a way so that they stay within the copy. An important feature of the *PaxEventInterpret* is the persistency. An instance of *PaxEventInterpret* can be written to a storage device and read back into memory. During the storage procedure, all the contained objects, including the relations between them, are persistent. This way, an intermediate state of the analysis can be written to disk, which can be considered as a sort of "mini" event data model.

The functionality described above is part of the so called "PAX kernel". Additional tools are provided for convenience and are shortly described in the following Section.

4.5.2 Additional Functionality of PAX

The *PaxFactory* is an extension to the kernel that takes care of bookkeeping and management of different event hypotheses. This is facilitated by the *PaxProcess* or *PaxAutoProcess* classes which take care of the evolution of all combinatorial possibilities of reconstructing a decay tree. The rules, according to which these processes are evolved, are defined by a process model which is represented by a *PaxEventInterpret* instance that contains a prototype of the decay chain. Since one event can be interpreted in terms of various process hypotheses (i.e. signal or background processes), the class *PaxProcessFactory* provides storage and easy access to an arbitrary number of processes (i.e. *PaxProcess* instances). It also performs begin- and end-of-job tasks and copies selected observables of event interpretations to ROOT trees.

The *VisualPax* tool allows to graphically display and modify event interpretations including properties like decay chains of the contained objects. A screenshot of the graphical user interface is shown in Figure 4.31.

Furthermore, a number of interfaces to various data formats (like HEPEVT ntuples) and experiment environments (e.g. CMS and CDF reconstruction software) have been developed and are maintained continuously.

4.5.3 Application of PAX in the $t\bar{t}H$ Analysis

Among the various successfully realized implementations of the $t\bar{t}H$ analysis (described in Chapter 5) is an implementation using the PAX toolkit. The combinatorial task of combining the various detector objects, like jets, missing transverse energy and leptons, as indicated in Figure 4.32, can become quite complicated. Even in the ideal case of exactly four b-jets and two light flavour jets, assuming perfect b-jet identification, there are 24 different possibilities of reconstructing the decay chain. The longitudinal momentum of the neutrino has to be calculated using a W mass constraint since the measurement of the missing energy allows only the determination of transverse components. This leads to a quadratic equation with two solutions and therefore to two possible interpretations. Moreover, the four b-jets have to be assigned to the top quarks and the Higgs boson. In the case of realistic b-tagging and a higher number of reconstructed jets, which might stem from initial and final state radiation, the number of possible combinations increases quickly.

Each of the possible interpretations is constructed and the hypothetical decay tree is calculated. For each possibility, a separate instance of *PaxEvenInterpret* is used. After this, the probability of each interpretation to be the correct one is calculated using a likelihood method. This method makes use of kinematic properties like the top masses and the hadronic

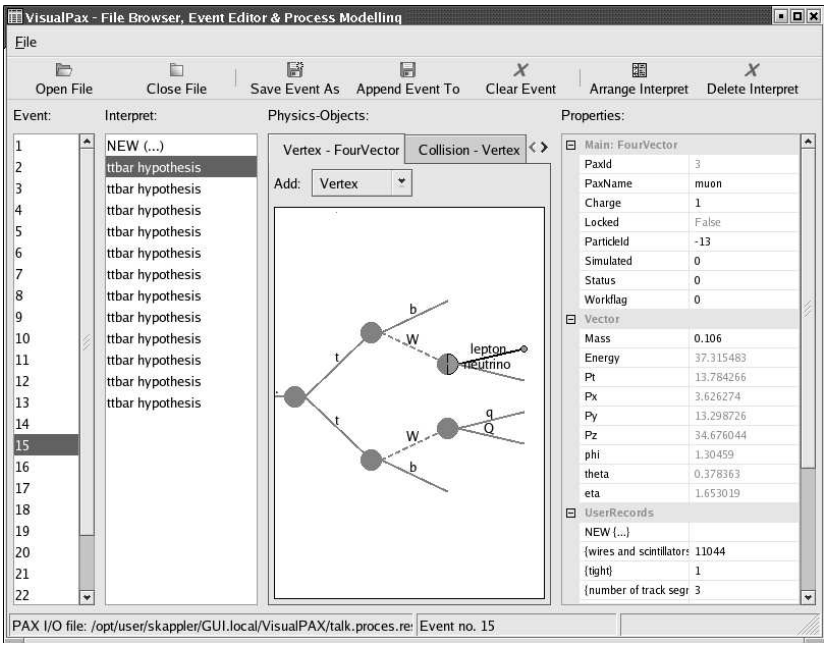

Figure 4.31: Graphical user interface of *VisualPax* and an example of a visualized $t\bar{t}$ decay tree.

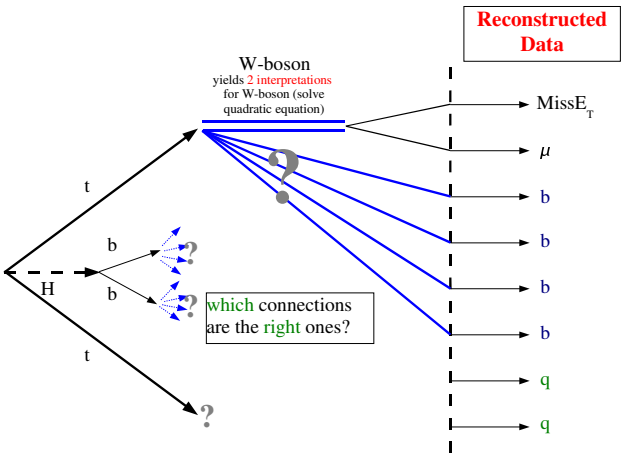

Figure 4.32: Combinatorial possibilities of reconstructing the $t\bar{t}H$ decay tree. In the ideal case, there are 24 possibilities.

W mass, as well as angles between the top quark and its decay products. Eventually, the best one is kept and used to produce the final results.

4.6 The LHC Computing Grid

The LHC Computing Grid (LCG) [63] is one of the key components for a successful accomplishment of the CMS experiment and the other experiments at the LHC. At modern hadron colliders, experiments have to deal with tremendous particle production rates and event sizes that have never been reached up to date. The demand on computing resources and mass storage systems has significantly increased and will continue to increase in the future. To give an example, the CMS detector has a collision rate of 40 MHz. With an expected event size of around 1.5 MByte, this would give a data rate of 40 TByte per second. This enormous information flow is filtered by several trigger levels, described in Section 2.2 and 4.2.1, and the final rate will be at the order of 100 Hz or 150 MByte per second, which is still an impressive data rate. In fact, there will probably be no place in the world where more data will be produced. Furthermore, these large amounts of information have to be processed and analysed by thousands of scientists which are distributed over almost all continents.

The analysis of the data, including comparisons with theoretical simulations would require about 100,000 CPUs and a storage amount of 15 PetaByte per year. The approach of providing these resources at a centralized computing center would be the traditional choice, but in case of the LHC, a novel, globally distributed model for data storage and analysis -the computing grid- has been chosen. The benefits of the distributed approach are:

- Costs for maintenance and upgrades are easier to manage in the context of the participating national organisations, which keep the responsibility for the operation and support of the local facilities.

- Single points of failure, like e.g. power cuts, are excluded in a distributed environment. Reassignment of computing tasks and multiple copies of frequently used data samples facilitate load balancing. User support is available in the same time zone.

The distributed approach does also have drawbacks like:

- Network bandwidth usage will be very high since the amount of data that has to be copied is large.
- Hardware will be heterogeneous.
- Coherence of software versions has to be ensured.

The LCG project adresses these challenges, provides solutions in the form of software products, and deploys the necessary middleware.

4.6.1 Tiered Architecture

The geographically distributed computing system of CMS is divided into four tiers, as indicated in Figure 4.33. The CMS experiment itself is called the "Tier 0" and it distributes its data to a number of "Tier 1/2/3" centers as shortly described below:

- The "Tier 0" center at CERN receives the raw data from the CMS Data Acquisition System, creates one archived copy of the data and performs a first reconstruction pass. It schedules and performs the data allocation for the Tier 1 centers.
- There will be about six "Tier 1" centers, distributed among the larger member states. They accept the data distributed by the Tier 0 center and provide data archiving, data access, reconstruction and analysis services. In general, the Tier 1 centers perform priority tasks like processing (skimming, calibration) of experiment data and preparation of higher level objects for Tier 2 centers.
- "Tier 2" centers have a more flexible architecture that can be managed by smaller organisations like a University Institute. These Tier 2 centers accept preprocessed data from the Tier 1 centers and provide physics analysis services that can be used directly and interactively by the physicists via batch submission systems.

To ensure the full funcionality of the tiered architecture at the time of the experiment's startup, so called data challenges, operating with a part of the expected data flow, are conducted regularly.

4.6.2 LCG Components

Currently, the software components of the LCG are still under heavy development. The next generation of grid software will be called "gLite" and is being developed by the EGEE II (Enabling Grids for E-sciencE) project [65], funded by the European Commission. The EGEE II project followes up the EGEE and EDG (European Data Grid) projects which have been successfully completed in 2006 and 2004, respectively. The key components are already usable and have been deployed on a large number of sites and have also been used in various data challenges and physics analyses. Also the analysis presented in this thesis made use of the grid infrastructure.

Some of the most important components of the current status of the grid infrastructure are explained in the following by means of a concrete example of an analysis job submission as outlined in Figure 4.34.

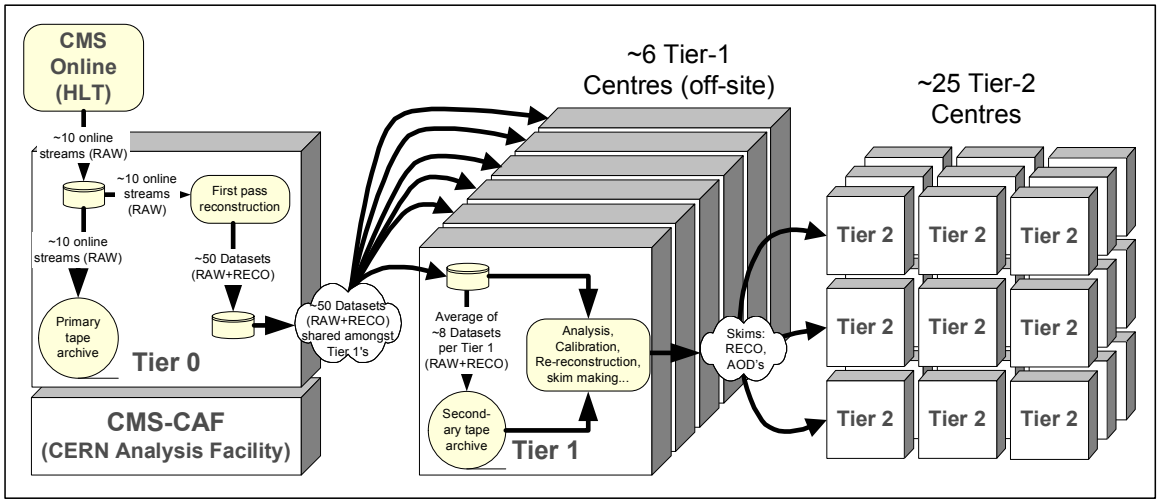

Figure 4.33: The tiered architecture of the CMS data distribution. The Tier 0 center is located at the experiment and receives the data stream directly from the CMS Data Acquisition System. It manages the distribution of selected data streams among the Tier 1 centers. The Tier 2 centers provide interactive services for the physicists. [64]

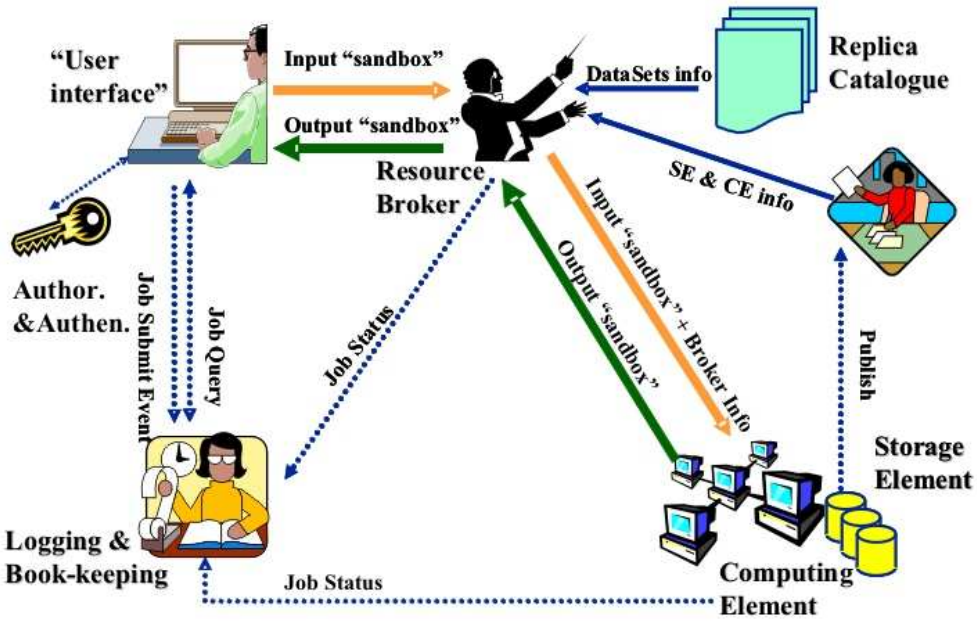

Figure 4.34: Schematic view of the interaction of the various grid services.

The "User Interface" (UI) is the entry point for the user who has a local account on this kind of machine. The user has to create a proxy certificate on this server in order to identify and authenticate himself, using his unique personalized grid certificate. The user creates and submits jobs using the user-space grid tools via the command line or graphical interfaces as documented in the according user guides and manuals [66]. The description of the computing job requirements are specified in a text file using the "job description language" (JDL) [67, 68]. In the JDL file, the details about required CPU usage, software versions, data access and the so called "sandboxes" are specified. The sandbox is a collection of small files needed by the job, like a batch script or configuration files, which is shipped together with the job.

After submission, the job is being transferred to the "Resource Broker". According to the specifications in the JDL file, the Resource Broker tries to find the optimal location for the job execution. The Resource Broker has information about all grid sites and their respective workload, which is matched to the JDL specifications following certain algorithms. The Resource Broker is a service with is being provided by some organisations like CERN or DESY, but there is no necessity for a grid site to provide its own Resource Broker.

After successful match-making, the job is forwarded to the "Computing Element" (CE) of the chosen grid site. Each grid site has to provide a Computing Element which acts as a sort of gateway or interface between the grid and the local computing center. The CE accepts the job and forwards it to the batch system, which in turn forwards it to the "Worker Node" where the job is finally executed.

After successful execution, the job traverses the whole chain backwards until the user queries the Resource Broker in order to collect the output sandbox. During this procedure the job status is regularly communicated to the Resource Broker and can be queried by the user.

The "Storage Element" (SE) plays a special role during job execution. The SE can be considered as a gateway to some storage space, in analogy to the CE. The SE hides the details of the storage area and provides space for experiment data as well as temporary files of analysis jobs. In principle, the Resource Broker is able to recognize the required data files of a job, which are specified in the JDL file, and it distributes the job in a way that it ends up "close" to the SE which is hosting the respective file, i.e. with direct access and without the need to copy the file. In reality, however, the concept of resource broking based on datasets is still problematic as described in the following Section.

The CMS Solution

Beginning with 2003 the CMS experiment started a large Monte Carlo production and physics analysis campaign with the goal to produce large amounts of fully simulated event data to be analyzed for the various Technical Design Reports [41, 1]. Since the grid tools have not yet been in a state where a reliable and efficient remote production would have been possible, the production has been distributed to the various Tier 1 sites and the computing jobs have been submitted to the local batch systems by the responsibles at the respective sites. The event data have been stored locally on the Tier 1 centers while the meta data, like details about production and software setup have been stored centrally in the so called RefDB [69].

Since resource brokering based on physics datasets was not possible, and is still problematic at the time of writing, CMS was forced to develop a temporary solution in order to enable the collaboration's physicists to analyze the large amounts of produced data. In the beginning, this was not more than an incoherent set of webpages, created by the local

production crew, that displayed some information about the available datasets, their size, location and instructions about how to access the datasamples. The situation at the different sites was similar, but there have been significant differences, e.g. in the storage systems and layout of the POOL catalogues [70], which was enough to overstrain the user's patience.

Remedy came from the deployment of the so called "PubDB". This database is situated directly at the site which hosts the respective data in order to make sure that the PubDB is always up to date and synchronized with the provided datasets. The centrally managed RefDB has only a link to the respective PubDB for each dataset. Figure 4.35 shows a schematic view of this concept. The information in the PubDB is accessed from the outside via the

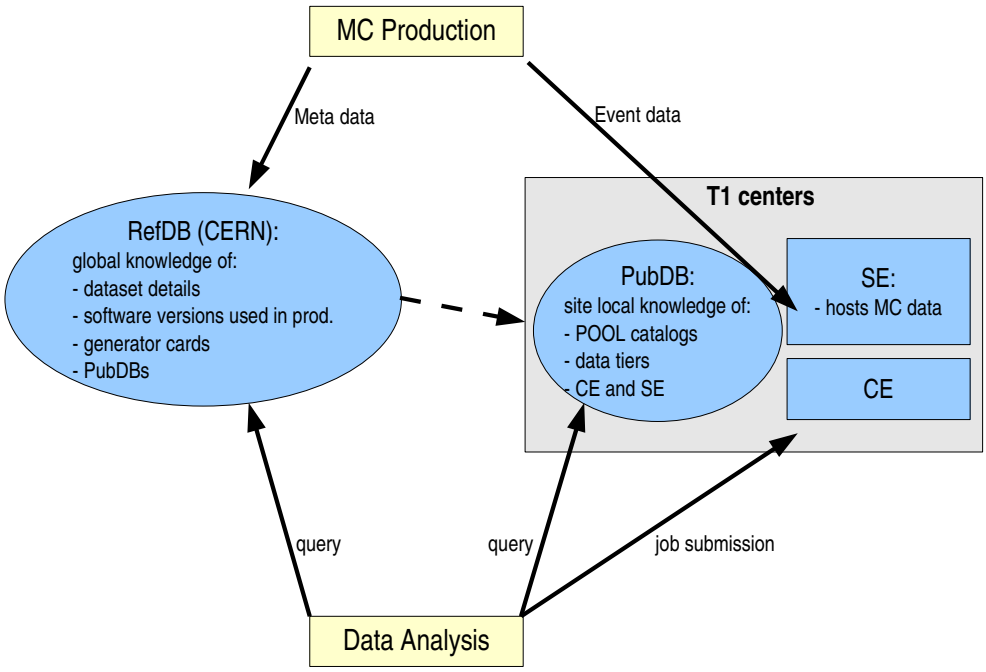

Figure 4.35: Schematic view of the role of "RefDB" and "PubDB" in the CMS production and physics analysis system.

http protocol. PubDB itself uses PHP (PHP: Hypertext Preprocessor) [71] scripts on the webserver and dynamically creates websites displaying the content. The PubDB websites are browsable which ensures that one can get a quick overview of the content. The information from PubDB is also provided in simple machine readable text to enable software tools to use the information in order to automatically create and configure analysis jobs as described below. The database backend of PubDB is implemented as a MySQL [72] server.

To facilitate the submission of physics analysis jobs, a tool called "CRAB" (CMS Remote Analysis Builder) [73] has been developed which takes care of the unconvenient formalities of the dataset discovery and analysis job setup. As indicated in Figure 4.35 this tool queries the

RefDB for the existence of a desired dataset. According to the RefDB entry, CRAB looks up the details in the proper PubDB at the according site which provides the dataset. Using this information, CRAB is able to create a complete set of configuration files for job submission and execution. Based on a local (on the UI) CMS software installation, CRAB finds and packs the necessary libraries and binary files into a bundle which is shipped in the sandbox. CRAB edits the analysis job configuration and applies the necessary changes to enable the job to run on the remote site. Additionally, a JDL file is created by CRAB which makes sure that the job finds its way through the grid to the location of the datafiles.

CRAB is also able to perform a splitting of jobs based on the desired number of events per job and the total number of available events. It takes care of job submission, monitoring and collection of the output.

This way, CMS was and still is able to perform remote physics analyses, even though this is not the originally foreseen clean grid concept.

Chapter 5

Study of $t\bar{t}H$ with $H \rightarrow b\bar{b}$ at CMS

The investigation of the $t\bar{t}H$, $H \rightarrow b\bar{b}$ discovery potential has a long standing history. The first publications within the CMS context by V. Drollinger and Th. Müller [74, 75] showed that this channel holds promise for an observation. Various major advancements of the CMS software and reconstruction methods have taken place since then. In the following sections, the current status is presented and the differences to previous results are investigated.

The analysis presented in this thesis is based on studies that have been performed during the year 2005 and beginning of 2006, when the CMS collaboration was in the process of completing its Physics Technical Design Reports (PTDR) [1, 41]. The outcome of this effort was a CMS Note [2] which is also the reference for all the analysis results presented in the following. In the meantime some improvements in b-tagging and optimizations of analysis techniques have been achieved, this will be pointed out explicitly in the respective sections. Even though the improvements sometimes required a significant amount of effort, the overall picture of this analysis did not change dramatically.

In the detailed presentation of the analysis methods and results in Sections 5.4 and 5.5, the semileptonic decay channel, in which one of the W bosons decays into an electron or muon and its corresponding antineutrino, while the second W boson decays hadronically, has been considered. This channel has the highest potential for an observation due to an optimal compromise between branching ratio and contribution from background events. Although the all-hadron channel has a branching ratio of 49%, it is difficult to observe, because of the large QCD background. About 28% of the events have a semileptonic decay which allows to trigger on the clean signature of an isolated muon or electron. Finally, some 5% contain two oppositely charged leptons, the di-lepton channel, which has a clean signature of two isolated leptons, but which does not allow to reconstruct the top masses unambiguously because of the two neutrinos. The remaining cases correspond to tau decays, which are difficult to distinguish owing to the complexity of the tau decay modes. In fact, these events contribute in small parts to the other channels.

A brief summary of the results for the all-hadron and di-lepton channels is given in Section 5.7.

5.1 Introduction

The Higgs boson decay channel $H \rightarrow b\bar{b}$ is the dominant one in the Standard Model up to a mass $m_H < 135$ GeV/c^2 as shown on the left side of Figure 3.2. The direct Higgs production

via gluon fusion $gg \to H$ has the largest production cross section as indicated on the right side of Figure 3.2, but this mode is impossible to detect because of the huge QCD cross section for $b\overline{b}$ production and the broad resolution of the invariant Higgs mass which does not allow to identify a narrow mass peak. The Higgs production in association with $t\overline{t}$, whose Feynman diagrams are shown in Figure 5.1, holds more promise because of lower background rates and

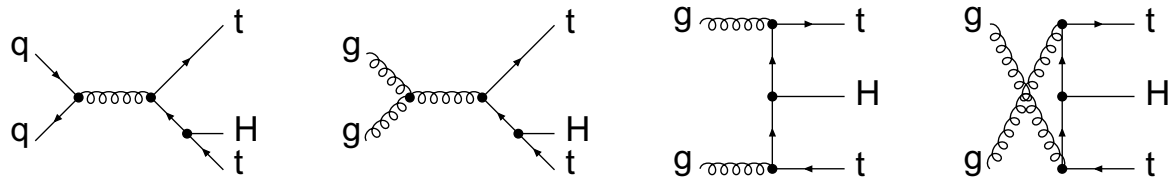

Figure 5.1: Leading order Feynman diagrams of $t\overline{t}H$ production processes.

because the $t\overline{t}$ system provides clear signatures and resonances that can be used for an event identification. Another advantage of this special production and decay mode might be the potential of the measurement of the combined t-H, H-b Yukawa coupling.

While next-to-leading order calculations for the $t\overline{t}H$ signal processes are available and give a correction faktor (k-factor) of ~ 1.2 at the LHC [76, 77], no NLO calculations for the background processes $t\overline{t}N$j with $N >= 2$ have been completed at the time of writing of this thesis. Therefore, a large theoretical systematic uncertainty of the order of roughly 20% [78] has to be assumed.

The top quark decays almost exclusively into $t \to Wb$ which leads to a total number of four b-jets that can be used to suppress "reducible" backgrounds stemming from events with less than four b-jets. In addition, the final state of a semileptonic $t\overline{t}H$ event, which is shown in Figure 5.2, consists of two light flavour jets, one charged lepton and a neutrino which emerges as missing transverse energy. Furthermore, additional jets produced by initial and final state

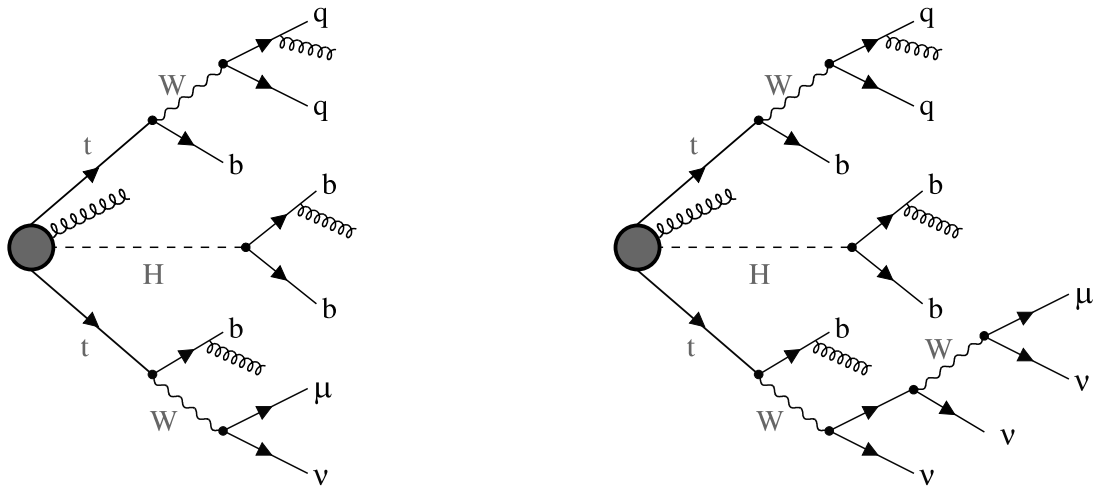

Figure 5.2: Final state of a semileptonic $t\overline{t}H$ event in case of a muonic W boson decay $W \to \mu\nu$. The diagram on the right side shows the same final state but with an intermediate τ decay, which is a priori not distinguishable from the detector signal produced by the final state on the left side.

radiation (ISR and FSR) occur. This complexity of the final state shows that all detector components are involved in the reconstruction of the $t\bar{t}H$ system. The results of the study will also show that this channel acts as a benchmark for the detector performance because the measurement is very difficult and pushes the analysis methods and detector reconstruction tools to the limits.

All the event samples are processed in full detector simulation including minimum bias and pileup events for a luminosity of $L = 2 \cdot 10^{33}\text{cm}^{-2}\text{s}^{-1}$. Realistic offline reconstruction tools as they currently exist at the time of writing of the CMS Physics TDR have been used. This leads to a substantially more complicated analysis procedure than previous simpler approaches using parameterized detector models, thus pointing out limitations that might not have become apparent before.

The analysis presented in this thesis tries to answer the question of feasibility of the discovery of the $H \rightarrow b\bar{b}$ decay based on the current understanding of things with respect to systematic uncertainties arising from detector effects as well as theoretical knowledge of the underlying physics processes. In both cases, the situation will change as soon as data and control samples arrive. Experiences from other hadron collider experiments show that the availability of a wide array of datasets and control samples allows very precise measurements even in difficult environments. Therefore, the question of how much the detector and theoretical predictions have to improve before a measurement is possible will also be discussed.

5.2 Event Generation and Simulation

5.2.1 Generation of Signal and Background Samples

The identification of the $t\bar{t}H$ signal makes use of the presence of two top quarks and their subsequent decay products. Hence, the most important backgrounds are also associated with $t\bar{t}$ production. The $t\bar{t}$ plus N light flavour jets ($t\bar{t}N\text{j}$) backgrounds turned out to be the most dominant, followed by $t\bar{t}$ plus b-jets and $t\bar{t}Z$ with $Z \rightarrow b\bar{b}$. These backgrounds are studied in detail in the following. Of minor importance are pure QCD multijet events and W/Z plus jets backgrounds, which are relevant for the all-hadron and di-lepton channels, but can be neglected for the semi-leptonic channel[1].

The $t\bar{t}H$ signal samples have been generated for three different Higgs boson masses (115, 120 and 130 GeV/c^2) using the CompHEP [79] generator, version 41.10. In order to simulate parton shower effects and initial and final state radiation, CompHEP was interfaced to PYTHIA [44], version 6.215. On generator level, no cuts have been applied for the signal sample. The next-to-leading-order cross sections and branching ratios for $H \rightarrow b\bar{b}$ for these samples are given in Table 5.1.

The CompHEP generator is not well suited to simulate $t\bar{t}N\text{j}$. CompHEP produces an inclusive event sample and higher order perturbative diagrams are not distinguished from the effects introduced by PYTHIA leading to a significant overestimation of the cross section of $t\bar{t}$ plus jets backgrounds. As an alternative, the $t\bar{t}N\text{j}$ samples are produced with ALPGEN, version 2.05 [80], for five different jet multiplicities ($N = 0, 1, 2, 3$ exclusive and $N \geq 4$ inclusive). ALPGEN applies a "matching" mechanism, in particular all of the matrix elements for $t\bar{t}$ plus N additional hard partons are included and properly combined at each order taking into account the interferences between amplitudes. These are then propagated through

[1] This is demonstrated by a short evaluation in Section 5.6

Table 5.1: NLO signal cross sections and $\mathrm{H} \to \mathrm{b\bar{b}}$ branching ratios for different Higgs mass hypotheses.

m_{H}	115 GeV/c^2	120 GeV/c^2	130 GeV/c^2
σ_{NLO} (pb)	0.747	0.664	0.532
$BR(\mathrm{H} \to \mathrm{b\bar{b}})$	0.731	0.677	0.525

PYTHIA, version 6.325, which adds parton shower and initial and final state radiation. The resulting events are then checked to see whether the number of hard partons in the final state is indeed N and not greater than N for exclusive samples. Events with more jets can occur as a result of the high energy extremes of the parton shower simulation in PYTHIA. This way it is possible to get a set of event samples with separated jet multiplicities. The following generator level cuts have been applied on $\mathrm{t\bar{t}}N\mathrm{j}$ events:

$$p_t(\mathrm{j}) > 20\mathrm{GeV/c}, \qquad |\eta(\mathrm{j})| < 5, \qquad \Delta\mathrm{R}(\mathrm{j}_1, \mathrm{j}_2) > 0.7,$$

where j_x denotes any of the light flavour extra jets and ΔR is the angular distance between jets in η, ϕ space: $\Delta R(\mathrm{j}_1, \mathrm{j}_2) = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$. The resulting leading order cross sections after all generator cuts for these samples are given in Table 5.2.

Table 5.2: LO ALPGEN cross-sections for the different jet multiplicities of $\mathrm{t\bar{t}}N\mathrm{j}$ after all generator cuts.

	exclusive $\mathrm{t\bar{t}}$+0j	excl. $\mathrm{t\bar{t}}$+1j	excl. $\mathrm{t\bar{t}}$+2j	excl. $\mathrm{t\bar{t}}$+3j	inclusive $\mathrm{t\bar{t}}$+4j
σ_{LO} (pb)	190	170	100	40	61

Historically, the analysis has been developed using an inclusive $\mathrm{t\bar{t}}$2j sample produced with CompHEP, because ALPGEN became available in the CMS framework only late in 2005. Using this sample, the impact of this background was much worse, even though $\mathrm{t\bar{t}}$0j and $\mathrm{t\bar{t}}$1j were not included. A detailed comparison of the $\mathrm{t\bar{t}}N\mathrm{j}$ background for CompHEP and ALPGEN is given in Section 5.2.4.

The same considerations are valid for the $\mathrm{t\bar{t}}$ plus b-jets background. But since this background is not dominant in comparison with $\mathrm{t\bar{t}}$ plus N light flavour jets ($\mathrm{t\bar{t}}N\mathrm{j}$), the older and conservative sample produced with CompHEP has been used for this analysis. This way the $\mathrm{t\bar{t}b\bar{b}}$ background is overestimated, but is still less dominant than $\mathrm{t\bar{t}}$ plus light flavour backgrounds. The following generator level cuts have been applied on $\mathrm{t\bar{t}b\bar{b}}$ events:

$$p_t(\mathrm{b}) > 15\mathrm{GeV/c}, \qquad |\eta(\mathrm{b})| < 3, \qquad \Delta\mathrm{R}(\mathrm{b}_1, \mathrm{b}_2) > 0.3.$$

The difference to the cuts on the ALPGEN samples has historical reasons. The higher p_t cut in ALPGEN reduces the generation inefficiency significantly. The effective cross sections before and after all generator level cuts for the background generated with CompHEP are listed in Table 5.3. The table shows that the generator preselection efficiency ϵ for both CompHEP backgrounds is around $\epsilon = 0.86$.

Table 5.3: Leading order CompHEP cross-sections of the considered background processes before and after the generator filters.

	$t\bar{t}b\bar{b}$	$t\bar{t}Z$
σ_{LO} (pb)	3.28	0.65
$\sigma_{LO} \times \epsilon$ (pb)	2.82	0.565

For all samples, the top mass has been assumed to be 175 GeV/c^2. For the CompHEP samples, CTEQ4L [81] parton distributions have been applied, while CTEQ5L [82] has been used for the ALPGEN samples.

For completeness, the total number of generated, simulated and analyzed events of all signal and background samples, including the expected number of events corresponding to an integrated luminosity of 60 fb^{-1} is given in Table 5.4. The last column gives a scaling factor which has to be applied to the analysis results in order to obtain the final event yields, for instance the number of remaining events after event selection. The number of remaining events after event selection can be very small for some of the $t\bar{t}N$j samples, and the scaling factors of these samples are large at the same time. In the case of very tight selection cuts, this might lead to statistical problems. Table 5.4 shows that the available number of Monte Carlo events is certainly sufficient for the signal and $t\bar{t}b\bar{b}$ and $t\bar{t}Z$ background samples.

The amount of available Monte Carlo Statistics might seem unsatisfactory, but the analysis conclusions of Section 5.5 are found to be quite stable within certain bounds. The number of simulated events is not a matter of choice since large amounts of computing resources are required. The exact number of events has been decided collectively within the CMS experiment. The somewhat odd numbers of generated and analyzed events is a result of instabilities in the computing and software environment, e.g. crashing jobs or corrupt files.

One possibility to get rid of the problem of lacking Monte Carlo statistics is the applica-

Table 5.4: Total number of generated, analyzed and expected events corresponding to an integrated luminosity of 60 fb^{-1}. The effective signal cross sections include the branching ratio for H $\rightarrow$ b$\bar{b}$ and the branching ratio (28%) for semileptonic W boson decays (μ and e), while the cross sections for the backgrounds refer to fully inclusive samples after all generator preselection cuts.

sample	eff. cross sec.	expected # ev.	gen. and ana. # ev.	scaling factor
$t\bar{t}H$ ($m_H = 115$ GeV)	0.153 pb	9180	55395	0.16572
$t\bar{t}H$ ($m_H = 120$ GeV)	0.126 pb	7560	191133	0.03955
$t\bar{t}H$ ($m_H = 130$ GeV)	0.078 pb	4692	44595	0.10521
$t\bar{t}$0j	190 pb	$11.4 \cdot 10^6$	98578	115.64
$t\bar{t}$1j	170 pb	$10.2 \cdot 10^6$	1297064	7.86
$t\bar{t}$2j	100 pb	$6 \cdot 10^6$	827615	7.25
$t\bar{t}$3j	40 pb	$2.4 \cdot 10^6$	108778	22.06
$t\bar{t}$4j	61 pb	$3.66 \cdot 10^6$	114054	32.09
$t\bar{t}b\bar{b}$	2.82 pb	169200	384407	0.4402
$t\bar{t}Z$	0.565 pb	33900	94706	0.3579

tion of a fast detector simulation. Unfortunately, the b-tagging performance, especially the light flavour misidentification rate in the fast simulation, described in Section 4.4.1, does not reproduce the performance of the full simulation well enough and can therefore not be used in this channel which primarily depends on the light flavour misidentification rate.

5.2.2 Reconstruction of Generator Parton Kinematics

In order to get an idea of the behaviour of signal and backgrounds and to estimate the expected event reconstruction performance, it is useful to study the behaviour of kinematic properties of the simulated events at generator level. For this purpose, the generator output has to be deciphered in a procedure which is not always unambiguous. An example for a generator listing and an explanation of the methods to reconstruct the primary partons is given in Appendix B.

The invariant top quark and W boson masses in the $t\bar{t}H$ sample, that have been reconstructed this way, are shown in Figure 5.3. The plots reproduce a Breit-Wigner distribution

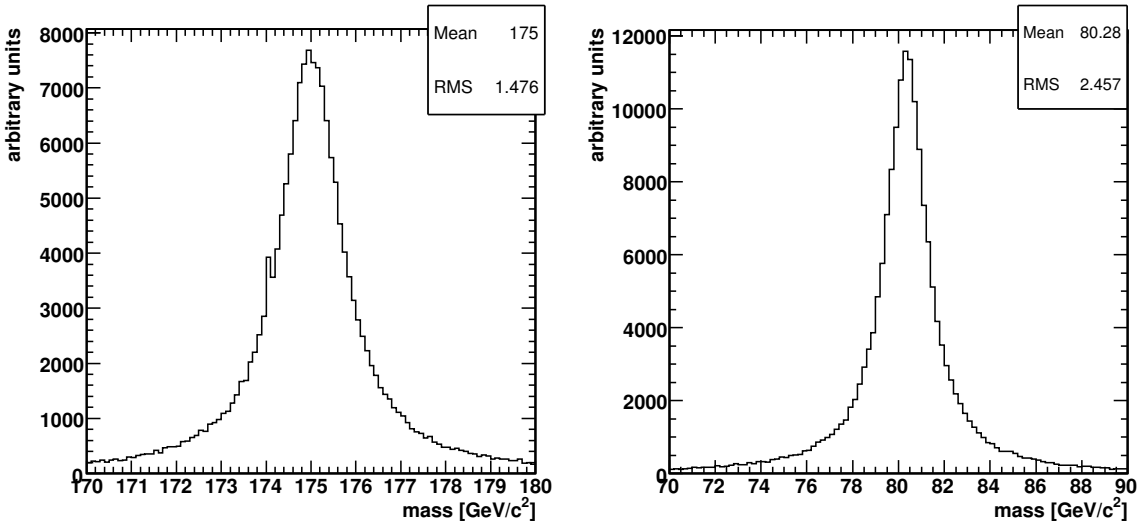

Figure 5.3: Distribution of the invariant masses of the generated top quark (left) and W boson (right). The plots show the hadronically decaying branch, but the distributions look identical in the leptonic case.

with the proper width and mean values. The invariant $b\bar{b}$ mass, i.e. the Higgs mass, is not displayed, since it simply shows a sharp peak at the generated m_H value of 120 GeV/c^2 in this case.

The generated distributions of the transverse momenta of the Higgs boson and the top quark are shown in Figure 5.4.

Of major interest for the efficiency of the $t\bar{t}H$ analysis is the distribution of the transverse momenta of the six signal partons which is shown in Figure 5.5. Ideally, these partons finally emerge in the detector as reconstructed jets, but several effects like parton showering, hadronization and detector resolution obfuscate this image. Jet reconstruction in the CMS experiment is only possible above a certain p_t threshold of approximately 15 to 20 GeV/c. Figure 5.5 shows that a significant amount of the signal partons have values below this

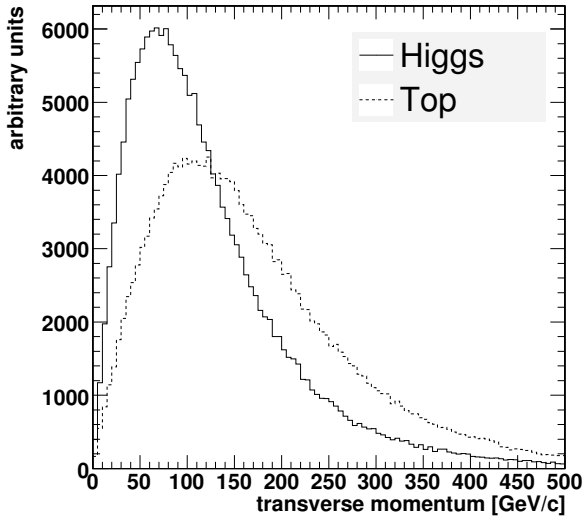

Figure 5.4: Distribution of the generated transverse momenta of the Higgs boson (m_H = 120 GeV/c^2) and hadronically decaying top quark. The case of the leptonically decaying top quark looks identical.

threshold. Especially in the case of the sixth parton in Figure 5.5, the peak of the distribution is below the reconstruction threshold. In addition, the systematic uncertainties connected with the jet energy scale are large at low p_t. This behaviour demonstrates the enormous challenge, which is connected with this analysis, since the reconstruction tools have to be used at the limits of their capabilities. The fraction of remaining signal events in dependence on the cut on the generated transverse momentum of the partons is shown in Figure 5.6. It is visible that already 50% of the events are cut away with a minimal requirement of $p_t > 20$ GeV/c.

Therefore, the reconstruction efficiency does not only strongly depend on the choice of the cut on the transverse jet momenta, but also the systematic error due to the jet energy scale is large, since small shifts in the energy scale might have a strong impact on the reconstruction efficiency.

5.2.3 Simulation and Digitization

The simulation of the interaction with the detector has been performed according to Section 4.1 with "cmsim", version 133, based on GEANT3, in case of the CompHEP samples. The ALPGEN samples have been produced with OSCAR, version 3.9.8, based on GEANT4. This separation has historical reasons. Originally, all samples used in the $t\bar{t}H$ analysis have been simulated with cmsim. Only late in 2005, the ALPGEN samples became available and have been simulated with the latest version of the CMS detector simulation programs. This mixing of different versions is legitimate since the performance of OSCAR and cmsim have been validated to give similar results.

The response of the detector electronics has been simulated with ORCA version 7.6.1 in case of the CompHEP samples and version 8.13.1 in case of the ALPGEN samples. An overview of the used software versions is given in Table 5.5.

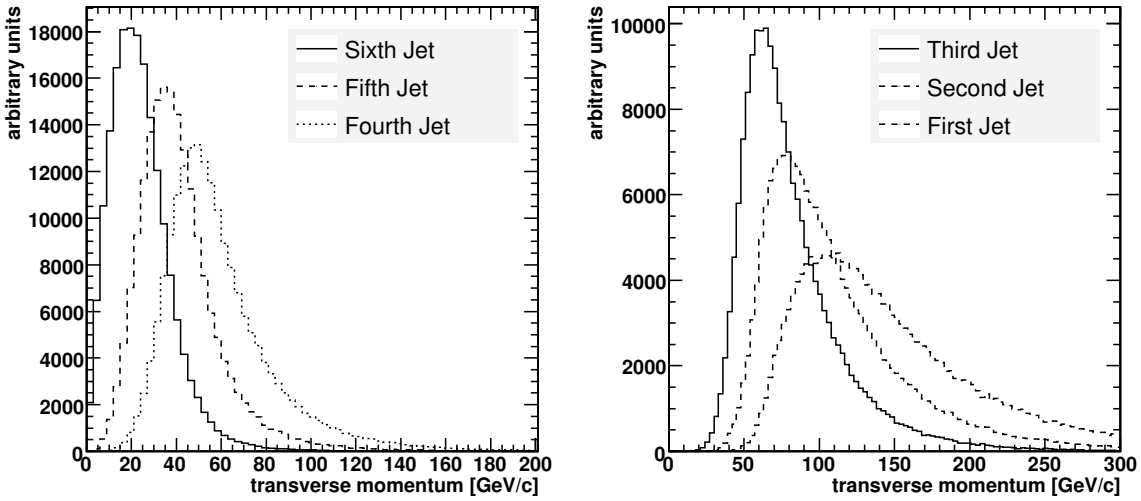

Figure 5.5: Distribution of the generated transverse momenta of the six leading signal partons, sorted in decreasing order of p_t.

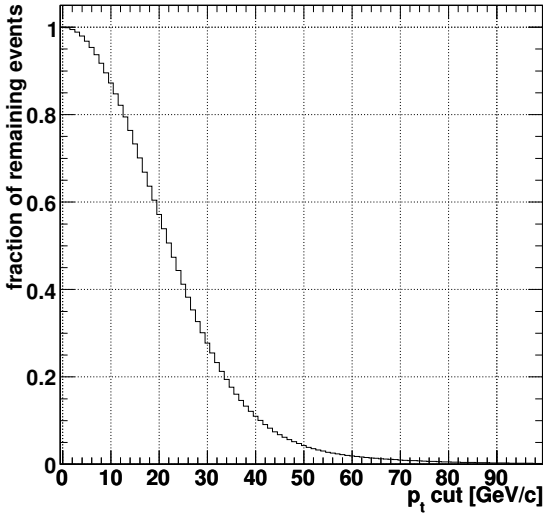

Figure 5.6: Fraction of remaining events in dependence on the cut on the generated transverse momentum of the signal partons.

Table 5.5: Generation parameters used for production of signal and background datasets.

Channel	Generator	PDF	Detector Simulation	Digitization
$\mathrm{t\bar{t}H}$	CompHEP + PYTHIA 6.215	CTEQ4L	CMSIM 133 (GEANT 3)	ORCA 7.6.1
$\mathrm{t\bar{t}b\bar{b}}$	CompHEP + PYTHIA 6.215	CTEQ4L	CMSIM 133 (GEANT 3)	ORCA 7.6.1
$\mathrm{t\bar{t}Z}$	CompHEP + PYTHIA 6.215	CTEQ4L	CMSIM 133 (GEANT 3)	ORCA 7.6.1
$\mathrm{t\bar{t}Nj}$	ALPGEN 2 + PYTHIA 6.325	CTEQ5L	OSCAR 3.9.8 (GEANT 4)	ORCA 8.13.1

5.2.4 Comparison of CompHEP and ALPGEN for the $t\bar{t}$ Plus Jets Background

Originally, the $t\bar{t}jj$ background was generated with CompHEP and proved to be very difficult to suppress. Since processes with extra jets in the final states are better described by ALPGEN for reasons explained in section 5.2.1, this generator has also been used, and the comparison of ALPGEN and CompHEP generated events is presented in this section.

The main feature of ALPGEN version 2 is the matching procedure introduced in the Matrix Element (ME) to Parton Shower (PS) interfacing. The parton-shower matching criteria avoid double counting due to the fact that initial and final state radiation are added by the PS generator on top of the extra jets already described at parton level. If no matching is applied from ME to PS generator, a significant overestimation of the rate of extra-jet production occurs.

This proper treatment of the parton shower matching is responsible for the reduction of the cross section of these background sources by more than a factor of two. In addition, the ALPGEN matching procedure allows the kinematics of the extra jets to be better described. Actually, a PS generator provides a more reliable description of extra jets with low transverse momentum, while a ME generator is more suitable to describe extra jets in the higher region of the p_t spectra of jets.

A direct comparison of the CompHEP and ALPGEN samples has to be taken with care, because of a different Q-scale and different PDFs (CTEQ4L and CTEQ5L). As mentioned in Section 5.2.1, exclusive samples of $t\bar{t}$ with exactly one, two and three extra jets, respectively, and an inclusive $t\bar{t}$ sample with at least four extra jets have been generated with ALPGEN to be compared with the inclusive $t\bar{t}jj$ CompHEP sample.

The effective cross sections on generator level after applying similar kinematical cuts are shown in Table 5.6. Taking the two-jet and the higher jet multiplicities together, the effective

Table 5.6: Comparison of the effective cross sections of the inclusive CompHEP $t\bar{t}jj$ sample and the exclusive ALPGEN samples after application of the same kinematical cuts: $p_t > 20$ GeV/c, $|\eta| < 3$, $\Delta R(j,j) > 0.7$. These cuts systematically reduce the effective cross sections listed in Table 5.2.

CompHEP $t\bar{t}jj$	330pb
ALPGEN exclusive $t\bar{t}1j$	120pb
ALPGEN exclusive $t\bar{t}2j$	73pb
ALPGEN exclusive $t\bar{t}3j$	32pb
ALPGEN inclusive $t\bar{t}4j$	51pb

ALPGEN cross section is still a factor of two smaller than the CompHEP cross section, where the CompHEP sample represents an inclusive two-jet or higher multiplicity sample.

Both the CompHEP and ALPGEN samples have been simulated and reconstructed using the setup described in Sections 5.2.3 and 5.3. Some reconstructed values are compared in the following.

Table 5.7 shows the event selection efficiency for a very basic choice of the selection cuts, i.e. High-Level Trigger for single muons, p_t cuts and b-tagging cuts. Figure 5.7 shows the number of reconstructed jets which pass a p_t cut of 20 GeV/c, while Figures 5.8 and 5.9 show

Table 5.7: Comparison of the event selection efficiency after application of the High-Level Trigger (HLT) for single muons and cuts. The cuts are subsequently applied from left to right. $disc > 0.7$ means that a cut of 0.7 is applied on the b-tagging discriminators of the first four jets.

	HLT	6 jets w. $p_t > 20\mathrm{GeV}/c$ and $\eta < 2.4$	4 b-jets w. $disc > 0.7$
CompHEP $\mathrm{t\bar{t}jj}$	18%	8.3%	0.05%
excl. ALP. $\mathrm{t\bar{t}1j}$	14%	2.2%	0.008%
excl. ALP. $\mathrm{t\bar{t}2j}$	14%	4.7%	0.019%
excl. ALP. $\mathrm{t\bar{t}3j}$	14%	8.3%	0.038%
incl. ALP. $\mathrm{t\bar{t}4j}$	13.4%	11.2%	0.13%

the spectra of the transverse momenta of the six leading jets.

5.3 Reconstruction of Basic Detector Objects

The following sections summarize the setup used for reconstruction of high level physics objects, like leptons and jets. Significant effort has been invested in the determination of the optimal configuration of the reconstruction algorithms. In case of the b-tagging algorithms, some dedicated improvements for the $\mathrm{t\bar{t}H}$ analysis have been introduced which has already been discussed in detail in Section 4.2.7.

5.3.1 High Level Trigger

The single lepton triggers as described in Section 4.2.1 have been found to be a good choice. The rest of the event selection beyond lepton selection is better performed offline in order to have more control over the discarded and accepted events.

The p_t threshold for muons in the HLT is 19 GeV/c and for electrons 26 GeV/c. The trigger efficiencies for signal and backgrounds are listed in Table 5.8.

For the sake of completeness, the efficiencies for the di-lepton and all-hadron channels, which are applying a different trigger setup are also given in Table 5.8. The di-lepton channel uses the single electron, single muon and single tau triggers in "OR" logic. The setup is the same as for the semileptonic channels except for the p_t threshold which is lowered to 15 GeV/c. For the "Jets" trigger, which is used in the all-hadron channel, the single jet, 3-jet and 4-jet triggers are combined, using E_t thresholds of 572, 195 and 80 GeV, respectively.

It should be noted that in a mature experiment, the trigger efficiencies can be expected to be much higher, especially if the single lepton triggers are combined with other physics objects, like missing energy or b-tagged jets. Up to 90% trigger efficiency can probably be expected.

5.3.2 Muon Reconstruction

Muons are reconstructed making use of the muon system and the tracker as described in Section 4.2.2 and Reference [83]. For this analysis, muons stemming from W boson decays have to be identified and separated from muons originating from other sources, in order to

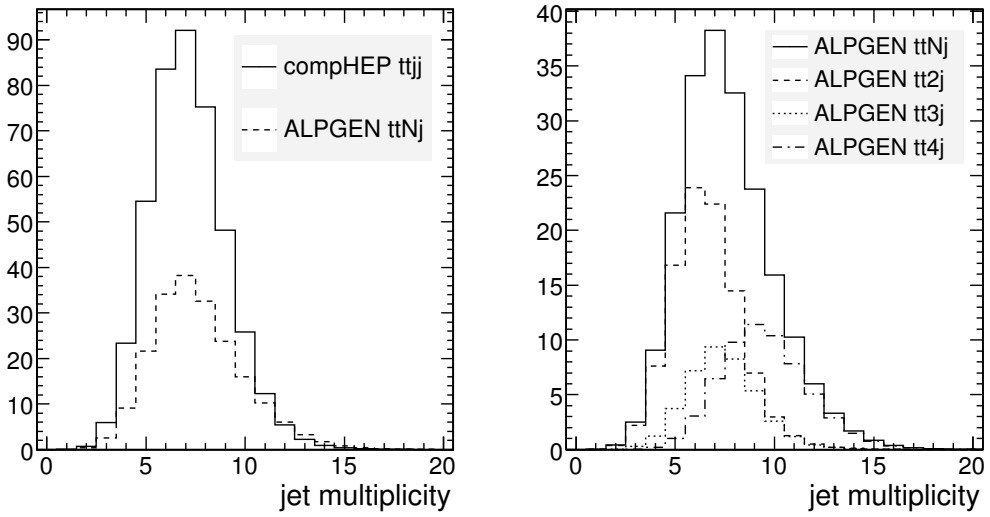

Figure 5.7: Comparison of the number of reconstructed jets above a p_t threshold of 20 GeV/c. $t\bar{t}$Nj represents the sum of the four ALPGEN multiplicities $t\bar{t}$1j, $t\bar{t}$2j, $t\bar{t}$3j and $t\bar{t}$4j. The units on the vertical axis are normalized to the CompHEP cross section, thus the relative contributions of the several multiplicities are reflected correctly. The left plot shows the CompHEP sample and the ALPGEN sample, where the $t\bar{t}N$j multiplicities are combined. The right plot shows the breakdown of the several multiplicities contributing to $t\bar{t}N$j .

Table 5.8: Signal and background efficiencies of the High Level Triggers, including the Level-1 Trigger selection. All the background efficiencies are defined with respect to inclusive background samples containing all top decay modes. The signal efficiencies for the single muon and single electron triggers are defined with respect to exclusive signal samples, containing only the respective semileptonic decay, while the "Jets" trigger refers to an exclusive hadronic sample. The $t\bar{t}$H efficiency for the Single e OR μ OR τ trigger is defined with respect to a sample, containing at least one leptonic top decay. The numbers are given in percent.

	Single μ (%)	Single e (%)	Single e OR μ OR τ (%)	Jets (%)
$t\bar{t}$H	63.5	52.4	76.7	24.9
$t\bar{t}b\bar{b}$	19.0	16.1	83.6	18.3
$t\bar{t}$1j	13.9	11.3	53.0	2.9
$t\bar{t}$2j	14.0	11.1	59.8	6.2
$t\bar{t}$3j	14.0	11.1	68.5	11.4
$t\bar{t}$4j	13.4	11.1	78.6	31.4
$t\bar{t}$Z	20.4	18.8	84.4	25.3
QCD 120-170 GeV/c	0.08	0.8	4.3	1.7
QCD >170 GeV/c	0.07	2.1	4.4	10.3

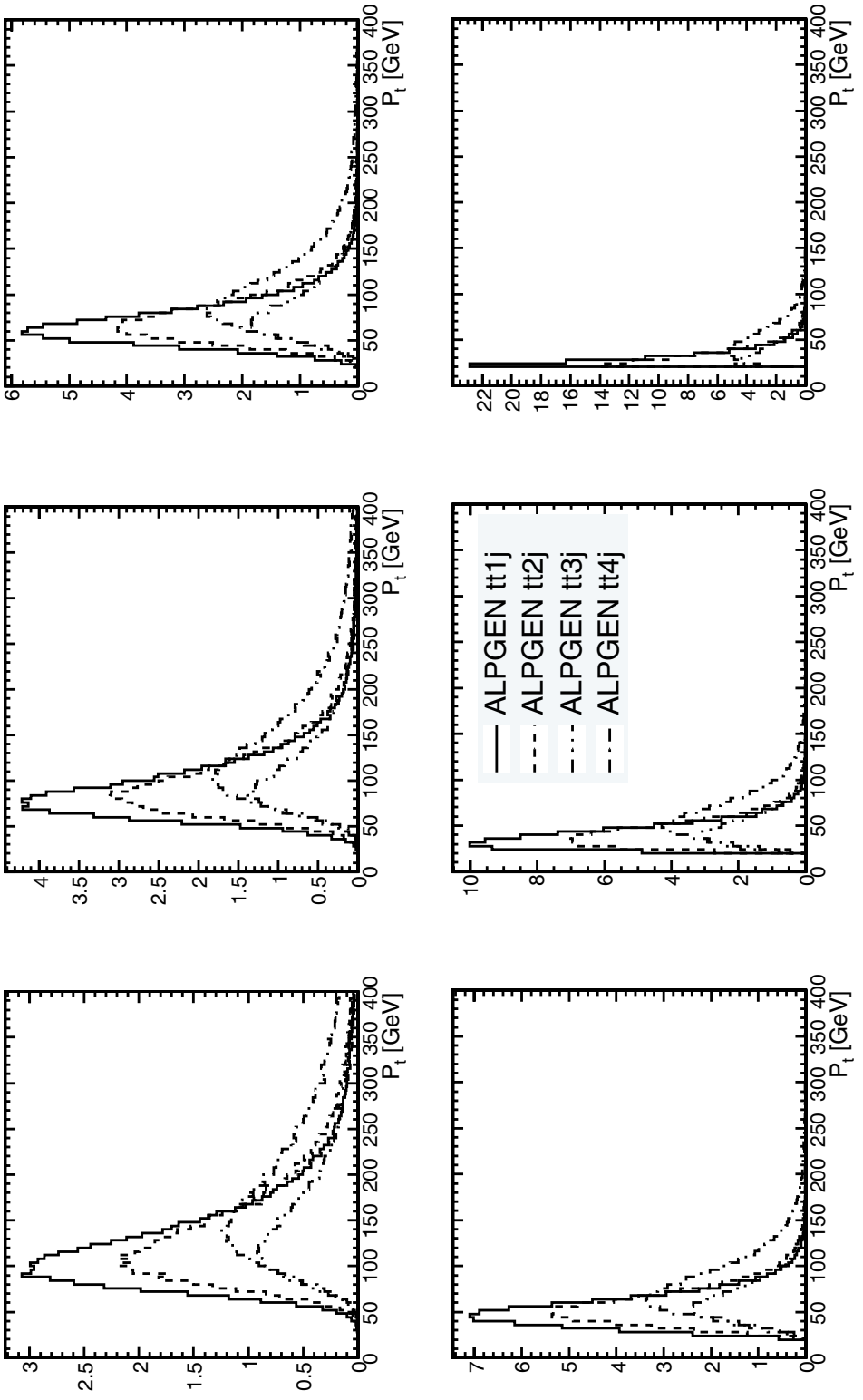

Figure 5.8: Comparison of the transverse momenta of the six leading jets for ALPGEN. The histograms are in order of decreasing p_t from top left to bottom right from the sideways perspective. The units on the vertical axis are normalized to the CompHEP cross section, thus the relative contributions of the several multiplicities are reflected correctly.

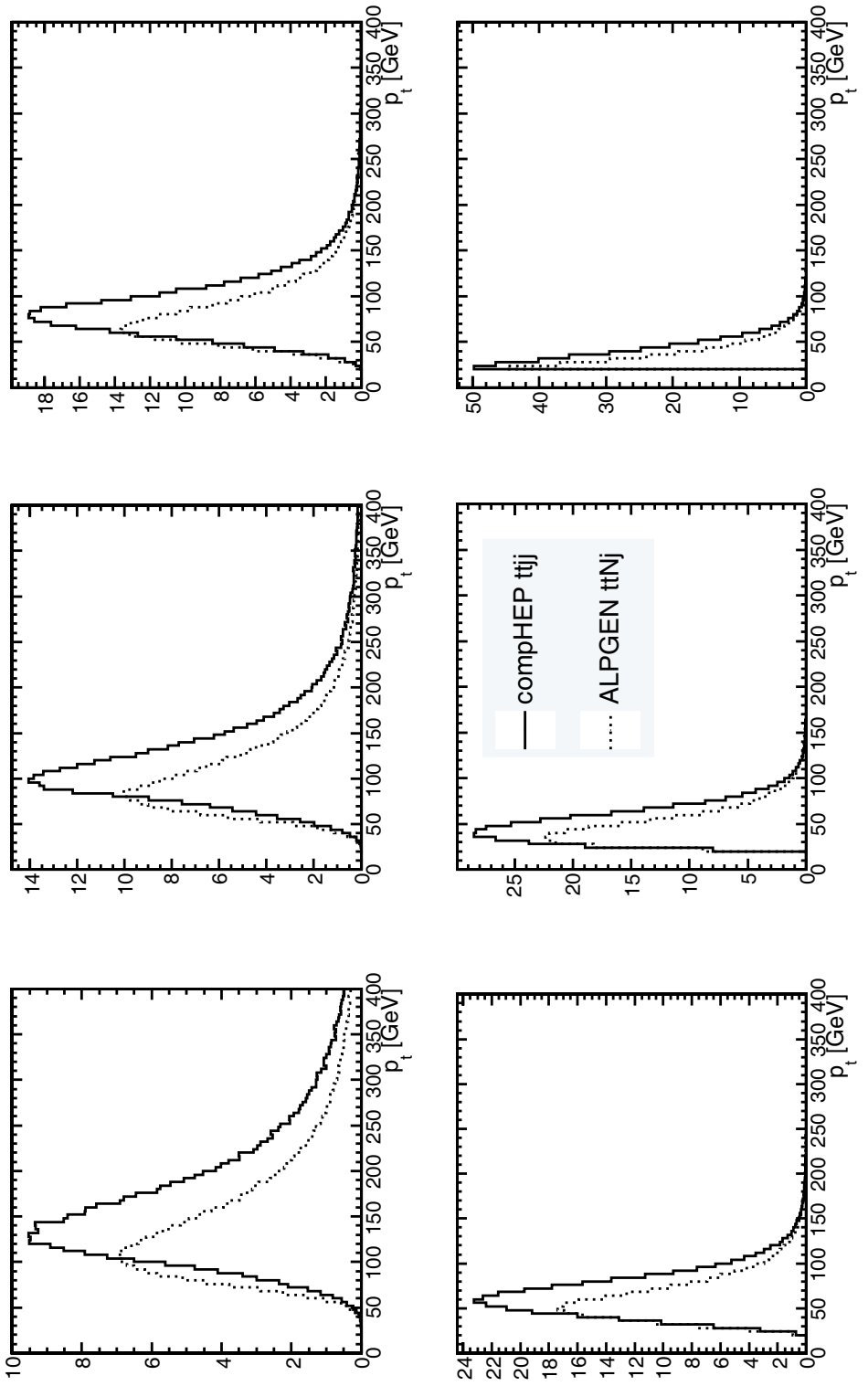

Figure 5.9: Comparison of the transverse momenta of the six leading jets. The histograms are in order of decreasing p_t from top left to bottom right from the sideways perspective. $t\bar{t}$Nj shows the sum of the four ALPGEN multiplicities $t\bar{t}$1j, $t\bar{t}$2j, $t\bar{t}$3j and $t\bar{t}$4j, while $t\bar{t}$jj represents the inclusive CompHEP sample. The units on the vertical axis are normalized to the CompHEP cross section, thus the relative contributions of the several multiplicities are reflected correctly.

properly reconstruct the decay chain. Other sources include muons from pileup and b-hadron decays, which are very frequent since there are four b-jets in a $\mathrm{t\bar{t}H}$ event. This explains the average number of about 3.5 muons per event, as shown in Figure 5.10. The separation

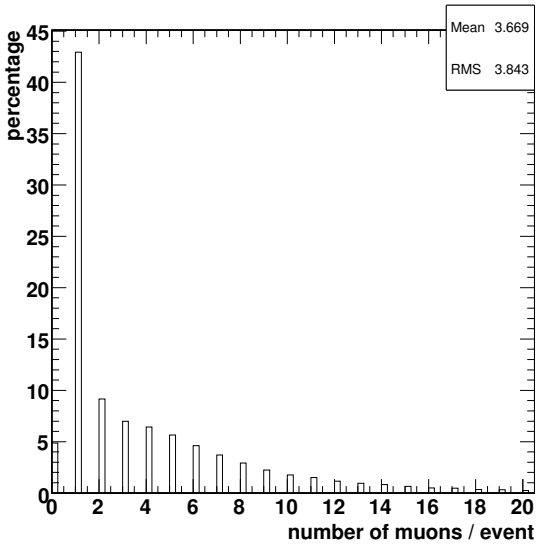

Figure 5.10: Number of reconstructed muons per event in semileptonic $\mathrm{t\bar{t}H}$ events.

between "signal" muons from W boson decays and muons from other sources ("background") is accomplished by constructing a discriminator which is based upon the following observables:

- Transverse momentum, p_t
- Track isolation, $IsoTk$, making use of the $\sum p_t$ of tracks inside a cone around the muon, as explained in Section 4.2.2
- Calorimeter Isolation, $IsoCalo$, making use of the $\sum E_t$ of calorimeter energy deposits in a cone around the muon, as explained in Section 4.2.2
- Impact Parameter Significance, $S_{ip} = d/\sigma_d$

The p_t variable is motivated by the fact that fake and pileup muons tend to have low transverse momenta. The isolation criteria are powerful because muons from W boson decays are not accompanied by any jets, in contrast to muons from b-decays. Also the impact parameter significance has the potential to suppress muons from b-decays since b-hadrons have a lifetime which is long enough to be able to cover a significant distance.

The Probability Density Functions (PDFs) associated with these observables are shown in Figure 5.11. These distributions are obtained by matching to generated muons, i.e. the reconstructed muon which is closest to the generated muon of the W boson decay (in ΔR disctance) is considered to be the "signal" muon. These PDFs are then combined into a likelihood ratio

$$L = \Pi_i \frac{P_i^{sig}(x_i)}{P_i^{sig}(x_i) + P_i^{bkg}(x_i)}, \tag{5.1}$$

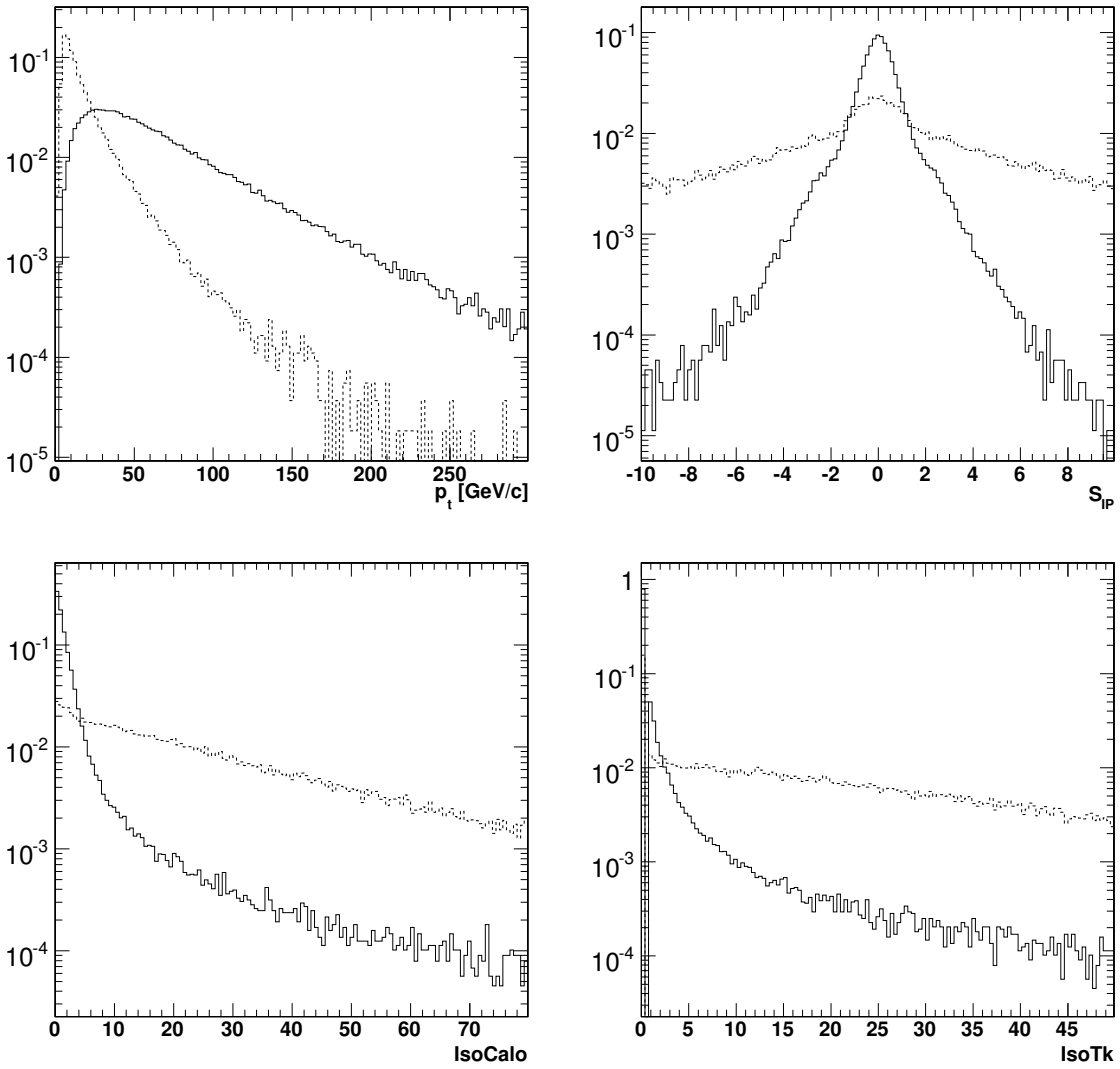

Figure 5.11: Probability Density Functions of the observables used for the muon selection. The black line refers to muons from the W boson decay and the dashed line to muons from other sources. From top left to bottom right: p_t, S_{ip}, $IsoCalo$, $IsoTk$.

where i denotes the observable (p_t, S_{ip}, $IsoCalo$, $IsoTk$), while $P_i^{sig}(x_i)$ or $P_i^{bkg}(x_i)$ denotes the probability to observe the value x_i in case of the "signal" or "background" muon distributions, respectively. The resulting distribution of the likelihood ratio and the performance of the muon selection are shown in Figure 5.12. It is visible that a rate of only 1% of wrong

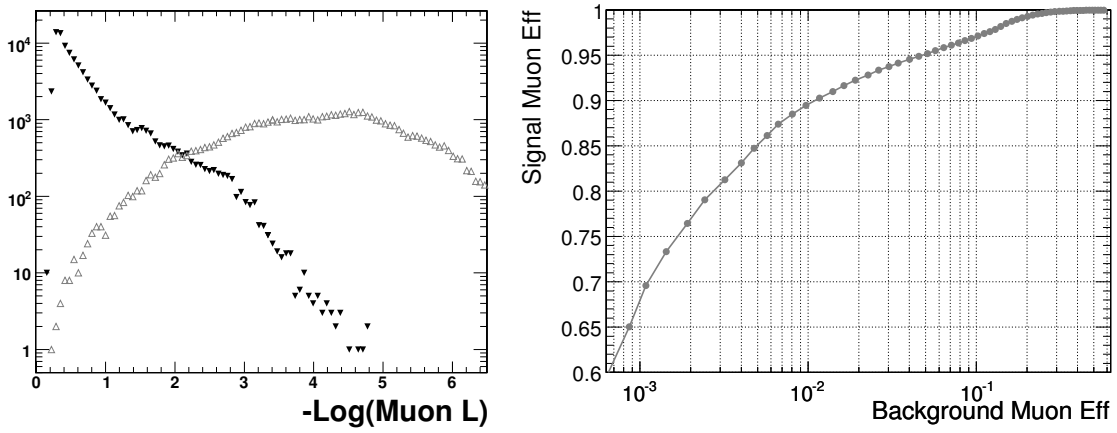

Figure 5.12: On the left: Distribution of $-Log$ of muon likelihoods. The black triangles refer to *signal* muons from W boson decays and the grey triangles to muons from other sources, as defined in the text. On the right: Performance of the muon likelihood discriminator for the $\mathrm{t\bar{t}H}$ channel.

selections is obtained at a signal muon selection efficiency of 90%.

The muon likelihood does not only facilitate the selection of the correct signal muon, it also has the power of suppressing the QCD background. Figure 5.13 shows the signal ($\mathrm{t\bar{t}H}$) selection efficiency versus the QCD background selection efficiency, where the QCD sample has been generated with $\hat{p}_t > 170$ GeV/c. The red star in this figure represents the HLT efficiency. The black line shows that the QCD selection efficiency can be reduced by a factor of about 3 (from 0.06% to 0.02%) at a minimal reduction of the signal selection efficiency (from 63% to 60%).

The muon likelihood is also used for the task of vetoing muons in the electron channel. The choice of the working point for the muon selection and double muon veto is discussed in Section 5.4.1.

The resolution of the inverse transverse momentum $1/p_t$ of muons is defined as

$$\frac{1/p_t^{rec} - 1/p_t^{gen}}{1/p_t^{gen}} \ , \tag{5.2}$$

where p_t^{gen} denotes the generated value of the transverse momentum and p_t^{rec} the reconstructed value. The left plot in Figure 5.14 shows the resolution in the case where the muon has been selected using the described likelihood method. This plot looks almost identical in the case where the muon resolution is obtained using a ΔR matching to the generated muon. The difference between these two methods of the muon selection is shown in the right plot of Figure 5.14. This plot has been obtained by exchanging the p_t^{gen} value in Equation 5.2 by the p_t^{rec} value of the reconstructed muon that matches best the generated muon. The large peak at 0 and the small tails and RMS confirm the small value of less than 1% of wrong muon selections, i.e. that the muon selected with the likelihood method is the same as the one selected by angular matching to generator muons.

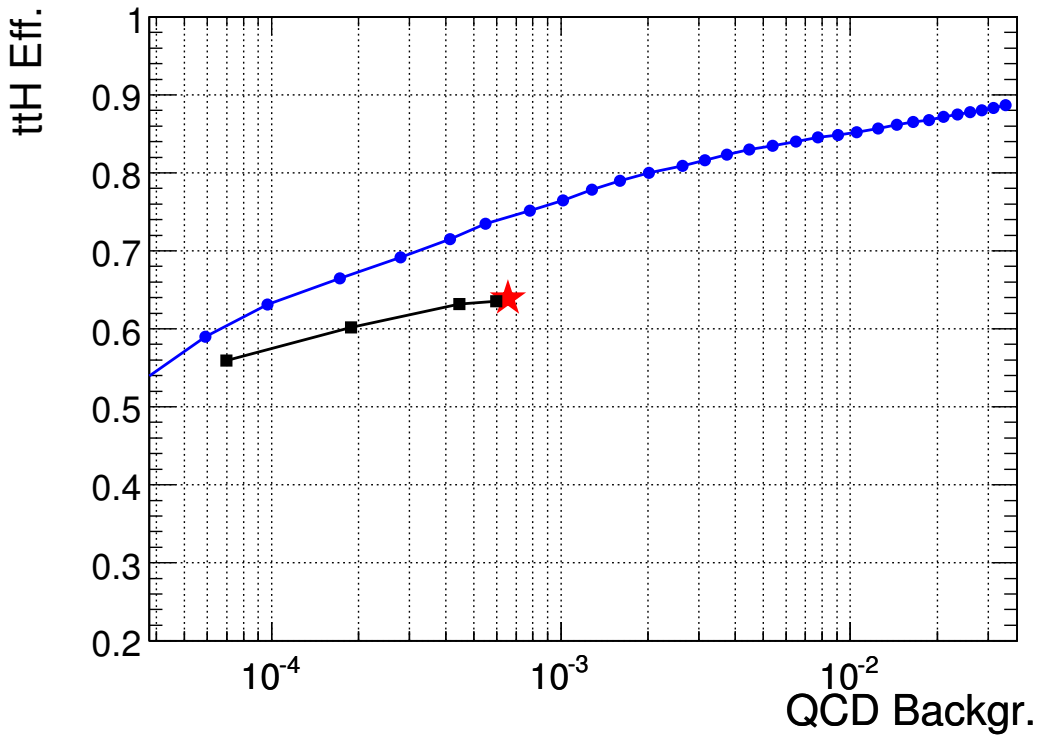

Figure 5.13: $t\bar{t}H$ signal efficiency versus QCD ($\hat{p}_t > 170$ GeV/c) efficiency. Circles: Likelihood performance without HLT selection; Star: HLT selection; Squares: Likelihood performance after the HLT selection.

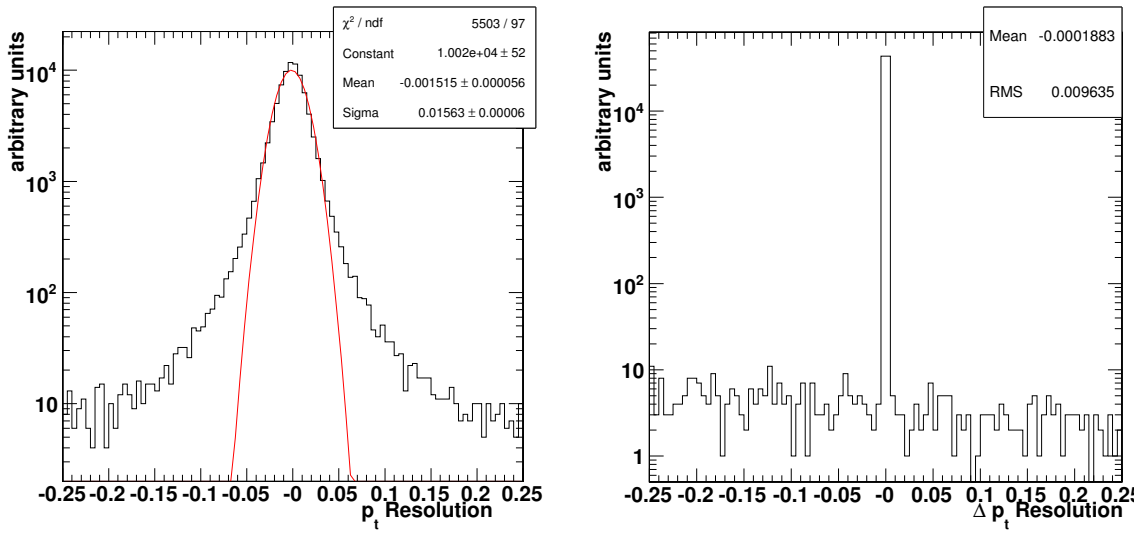

Figure 5.14: On the left: Resolution (defined according to Equation 5.2) of the transverse momentum of muons that have been selected with the likelihood method. On the right: Resolution difference between the selection with the likelihood method and angular matching to generator muons.

The resolution of 0.015 is obtained by a gaussian fit to the resolution distribution and is in perfect agreement with the values for muons between 10 and 100 GeV/c, as quoted in Section 4.2.2 and [41].

5.3.3 Electron Reconstruction

Electrons are reconstructed using the electromagnetic calorimeter in combination with the tracker as described in Section 4.2.3 and Reference [84]. As in the case of muons, the average number of electrons per semileptonic $\mathrm{t\bar{t}H}$ event is larger than two. Figure 5.15 shows the

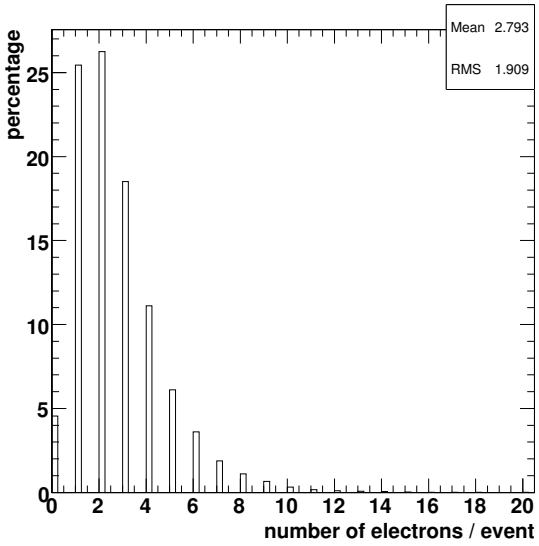

Figure 5.15: Number of reconstructed electrons per event in semileptonic $\mathrm{t\bar{t}H}$ events.

multiplicity of reconstructed electrons per event. In order to distinguish electrons from W boson decays from other electron sources and to suppress backgrounds, a likelihood ratio method has been developed analogously to the muon case. In principle, similar arguments as for the muon selection can be asserted for the electron selection. But since electrons have a different behaviour in the detector and since their measurements imply different methods, the observables are not the same.

In the electron case, isolation is defined my means of two variables. The first is the sum of transverse momenta of tracks inside a cone of radius $\Delta R = 0.3$ around the electron's direction as defined by the calorimeter. A veto cone of radius $\Delta R = 0.015$ around the electron's direction is defined in order to exclude the electron energy from this sum. The second variable is the distance (in ΔR) between the electron and the closest track outside the veto cone. Some additional variables, also making use of the hadronic calorimeter, are defined, which leads to the following five observables that are used in the electron likelihood:

- $\sum p_t$ of tracks around the electron, p_t^{Iso}
- Distance to closest track, $\Delta R(electron, track)$
- Transverse momentum of the electron, p_t
- Ratio of cluster energy to track momentum, E/p

- Ratio of hadronic to electromagnetic energy of the cluster, H/E

The distributions of these variables are displayed in Figure 5.16, where the "signal" electrons are represented by solid lines and electrons from other sources by dashed lines. These distributions are normalized and thus represent the probability density functions used in the likelihood ratio. The construction of this likelihood ratio follows Equation 5.1 in the same way as in the muon case. The resulting distribution of the $-Log$ likelihood and the performance are shown in Figure 5.17.

The resolution of the transverse momentum p_t of electrons is defined identically as the muon p_t resolution according to Equation 5.2 and is displayed on the left side of Figure 5.18. As in the muon case, the difference between the resolution obtained with angular matching and electrons selected with the likelihood method is very small as shown on the right side of Figure 5.18. However, the p_t resolution of electrons is asymmetric. The reconstructed momentum is much smaller than the generated momentum. This behaviour is due to radiation effects like Bremsstrahlung that are not corrected for.

5.3.4 Jet Reconstruction

Motivated by the studies on generator level presented in Section 4.3 which suggest a cone radius of 0.4 for the Iterative Cone algorithm, several radii have been used on reconstructed calorimeter towers in case of the semileptonic $t\bar{t}H$ analysis. A value of 0.5 was found to deliver a useful performance. This is also the setup recommended by the guidelines for the CMS Physics TDR analyses [85]. These guidelines suggest to use the MCJet calibration functions, which have been adopted for this analysis. All the jet energies quoted in the following sections are calibrated energies.

In case of the all-hadron channel, a smaller cone radius is expected to give better results which has been confirmed by a separate optimization described in Section 5.7.

Since electrons deposit their energy predominantly in the electromagnetic calorimeter and since the jet finding algorithms use both, electromagnetic and hadronic calorimeter towers, as input, electrons are reconstructed as jets. This is in principle a good thing, because electrons are mostly part of the decay chains in jets. For example, 20% of all b-jets have one or more leptons and the lepton energy must be counted together with the hadronic energy. A problem arises in the case where the isolated signal electron from the W boson decay (see Section 5.3.3) is counted as a jet. In this case the electron is double-counted because it is already reconstructed separately by the electron reconstruction. Therefore, the jet produced by the isolated signal electron has to be removed from the list of jets for the subsequent analysis. Figure 5.19 shows the angular distance (in ΔR) between the signal electron and the closest or second closest jet, respectively. Here, the signal electron has been identified by two different methods, once by angular matching to the generated electron, and second by a realistic selection using the electron likelihood from Section 5.3.3. The former is shown in Figure 5.19, the latter looks almost identical, as expected, considering the clean lepton identification. The plot clearly shows that there is exactly one and only one jet that has the same direction as the signal electron. There is almost no overlap between the second closest jet and the signal electron. Hence, the simplest method of removing the jet that is closest to the signal electron is enough to get rid of this unwanted additional jet.

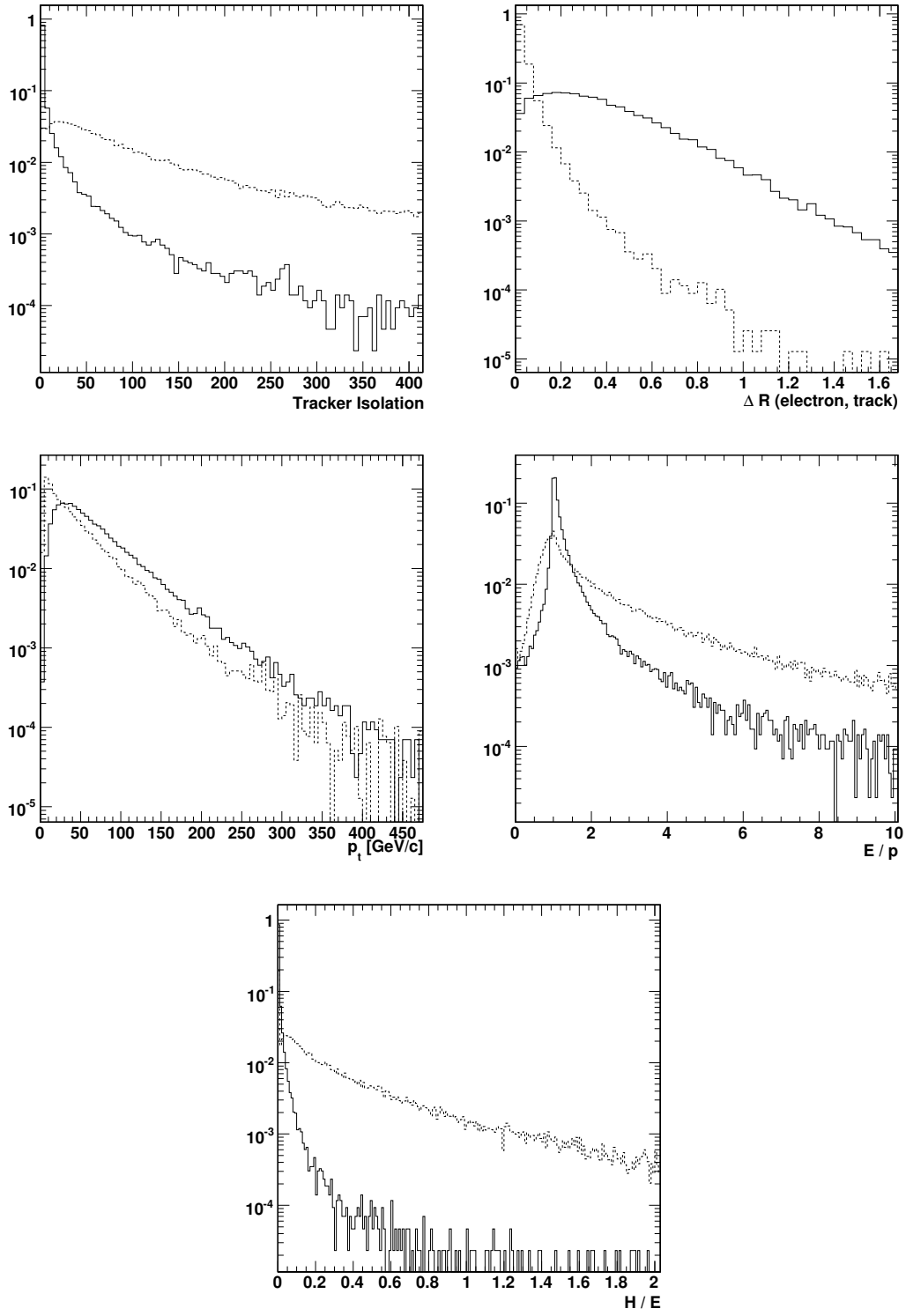

Figure 5.16: Probability density functions of the five observables used to construct the electron likelihood, as discussed in the text. The solid line refers to signal muons from W boson decays and the dashed line to muons from other sources. From top left to bottom: Tracker Isolation P_t^{Iso}, Distance to closest track $\Delta R(electron, track)$, Transverse momentum of the electron p_t, Ratio of cluster energy to track momentum E/p, Ratio of hadronic to electromagnetic energy of the cluster H/E.

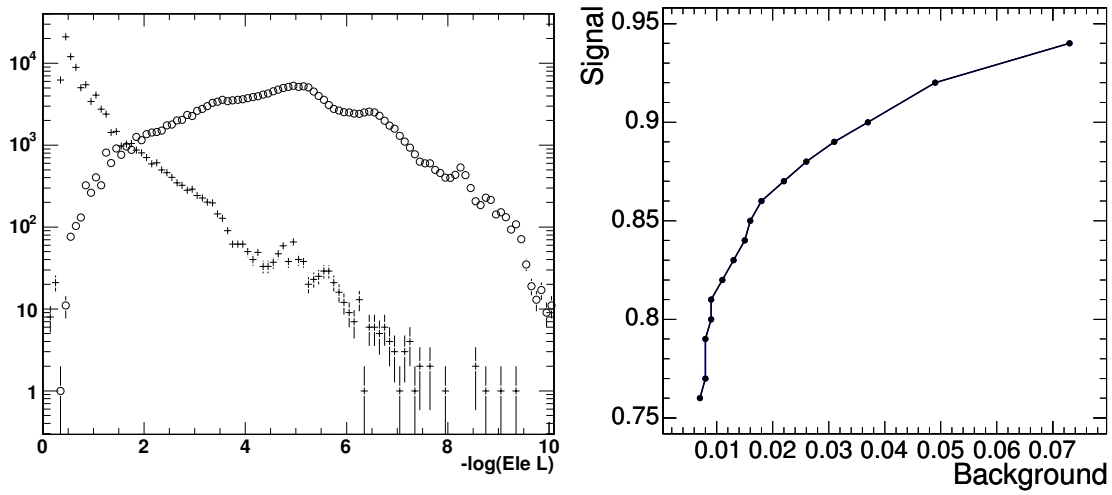

Figure 5.17: On the left: $-Log$ of the electron likelihood distributions for *signal* electrons from W boson decays (crosses) and electrons from other sources (circles). On the right: Performance of the electron likelihood discriminator for the $t\bar{t}H$ channel. [2]

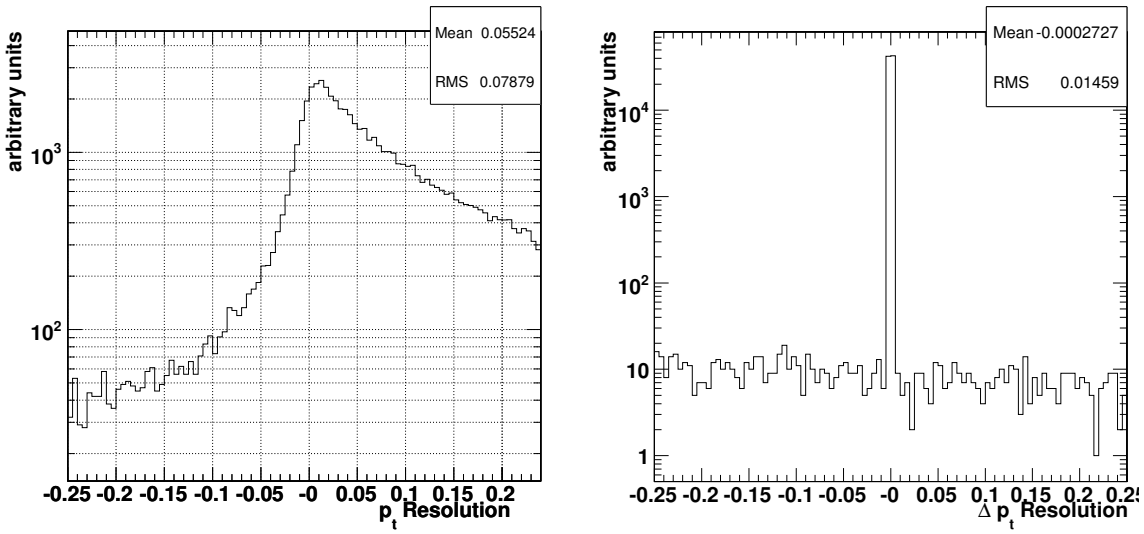

Figure 5.18: On the left: Resolution (defined according to Equation 5.2) of the transverse momentum of electrons that have been selected with the likelihood method. On the right: Resolution difference between the selection with the likelihood method and angular matching to generator electrons.

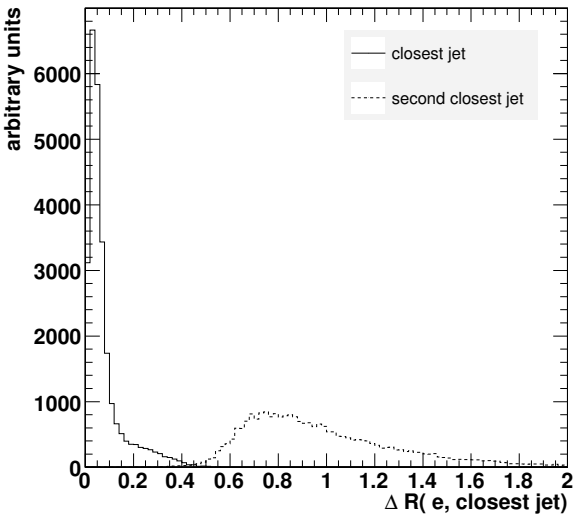

Figure 5.19: Angular distance (in ΔR) between the signal electron and the closest or second closest jet, respectively.

5.3.5 Reconstruction of Missing Transverse Energy

The Missing Transverse Energy E_T is reconstructed using the sum over all electromagnetic and hadronic calorimeter towers, including corrections from jet calibration and muons [86]. Since muons deposit only a small fraction of their energy in the calorimeter (about 4 GeV), they have to be added to the sum of calorimeter towers. This leads to the following equation for E_T:

$$E_T = \sum_i E_T^{tower} - (\sum_j E_T^{RawJet} - \sum_k E_T^{CaliJet}) + \sum_m E_T^{Muon}, \tag{5.3}$$

where the index i runs over all calorimeter towers, while the j and k indices run over uncalibrated ("raw") and calibrated ("cali") jets, respectively. The index m counts all muons.

The corrections due to the jet calibration can be justified by the fact that the energy measurement in the calorimeter underestimates the true energy, which is being corrected by the calibration. Even though the jet calibration corrects also for "out-of-cone" effects, which are double-counted in this case, this correction is necessary, because the latter is a rather small effect.

The resulting resolutions of E_T are shown on the right side of Figure 5.20. In this case the resolution is defined as $E_T^{reconstructed} - E_T^{generated}$, where $E_T^{generated}$ is simply the transverse momentum of the generated neutrino. The "more correct" way of determining the resolution would be to use the total sum of generated stable particles (except neutrinos) as reference instead of just the neutrino, since more than one neutrino could be present. However, in the present case one is interested in how well the neutrino is being represented by E_T. The difference between these two approaches is small, at the order of a few GeV, anyway. Figure 5.20 shows the improvement of the resolution that is achieved by applying the muon and jet corrections following Equation 5.3. The correction using muons improves the E_T resolution by 14.5% while the application of the correction from jet calibrations improves the resolution by another 15%. In order to reject jets with bad reconstruction and calibration reliability, only

jets with $p_t > 15$ GeV/c are used for this correction.

The left side of Figure 5.20 shows the absolute distribution of E_T for semileptonic $t\bar{t}H$ events in the case where both muon and jet corrections are applied. The distribution of the uncorrected E_T looks similar.

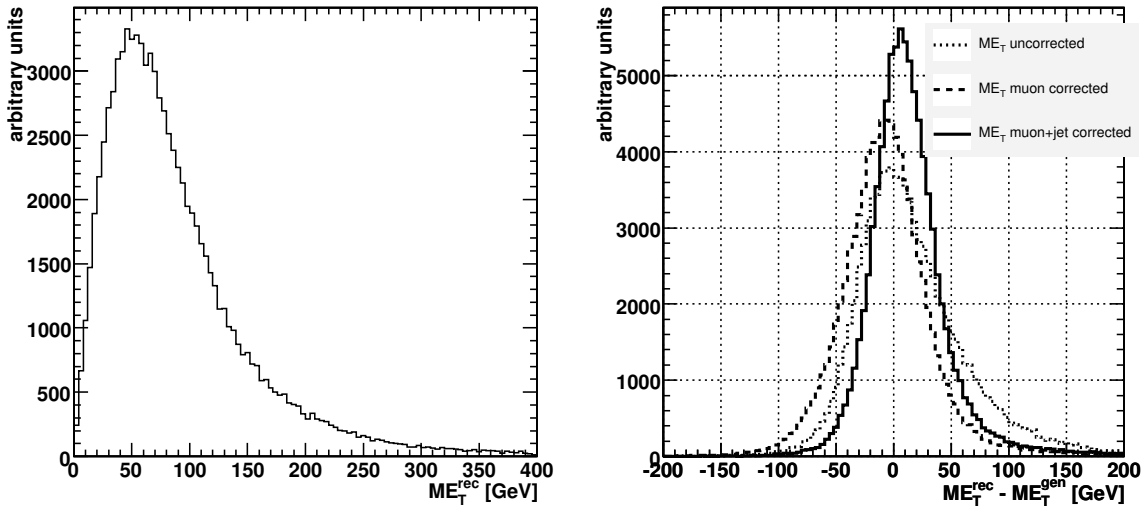

Figure 5.20: On the left: Distribution of E_T including muon and jet corrections in semileptonic $t\bar{t}H$ events.
On the right: Absolute resolution in E_T. The Root Mean Square (RMS) value of the uncorrected E_T resolution is 47 GeV, while the RMS with muon corrections is 40 GeV and 34 GeV with both, muon and jet corrections.

5.4 Event Reconstruction

5.4.1 Optimization of the Preselection

The event preselection fulfills two important tasks. First, the analysis performance is increased significantly since a very large number of background events is already rejected at preselection level and does not have to be considered in the rather CPU intensive construction of the combinatorial possibilities and subsequent likelihood evaluation.

Second, the separation of the four channels, i.e. semileptonic muon, semileptonic electron, all-hadron and di-lepton channels is performed efficiency on preselection level based on the lepton likelihood variables introduced in Sections 5.3.2 and 5.3.3. To enable an easy calculation of the combined significance using all channels together, this preselection is constructed in such a way that the four channels are completely disjoint. For this purpose, the following set of cuts on the likelihood values for the lepton acceptances and vetoes have been agreed upon [2]:

- Semileptonic muon channel: First muon selection $-\log L < 1.2$, second muon veto $-\log L < 1.4$, electron veto $-\log L < 1.2$

- Semileptonic electron channel: First electron selection $-\log L < 1.2$, second electron veto $-\log L < 1.4$, muon veto $-\log L < 1.2$
- Di-lepton channel: First or second muon selection $-\log L < 1.4$, first or second electron selection $-\log L < 1.2$
- All-hadron channel: Electron veto $-\log L < 1.2$, muon veto $-\log L < 1.4$

For instance, a muon selection of $-\log L < 1.2$ means that the lowest likelihood value has to be smaller than 1.2, otherwise the event will be rejected. A second muon veto of $-\log L < 1.4$ means that the event will be rejected if the second lowest likelihood value is smaller than 1.4. This way the four channels are by construction strictly separated without any overlap.

Furthermore, the preselection uses a simple cut on the b-tagging discriminator in order to reject a large number of background events without reducing the signal acceptance too much. This cut is not being optimized at this stage, because in Section 5.4.3 a more advanced likelihood method, combining the probabilities of four b-jets, is introduced and optimized. Figure 5.21 shows the efficiency of accepting events in dependence on the cut on the b-tagging

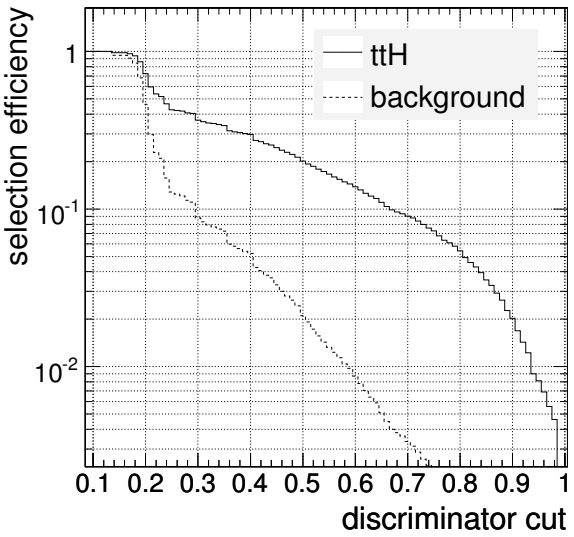

Figure 5.21: Selection efficiency in dependence on the cut on the b-tagging discriminator of the four jets with the highest discriminator values for the $\mathrm{t\bar{t}H}$ signal and all relevant backgrounds ($\mathrm{t\bar{t}}N\mathrm{j}$, $\mathrm{t\bar{t}b\bar{b}}$ and $\mathrm{t\bar{t}Z}$).

discriminator of the four jets with the highest discriminator values for the $\mathrm{t\bar{t}H}$ signal and all backgrounds ($\mathrm{t\bar{t}}N\mathrm{j}$, $\mathrm{t\bar{t}b\bar{b}}$ and $\mathrm{t\bar{t}Z}$). The choice of a cut value of 0.3 seems to be reasonable, because it cuts away a large fraction of background events, which significantly reduces the required amount of CPU time consumption for the further analysis.

In a final step, a preselection based on the transverse momentum p_t of the jets is performed. This has to be done under the application of a b-tagging cut, because the b-tagging performance depends on p_t as discussed in Section 4.2.6. To illustrate this, the event selection efficiency in dependence on the p_t cut is shown in Figure 5.22. This Figure also displays the dependence on the cut on the maximum number of jets (none, 7 or 8), for the $\mathrm{t\bar{t}H}$ signal and the $\mathrm{t\bar{t}4j}$ background. An increasing p_t cut leads to a decrease of jets passing the minimum

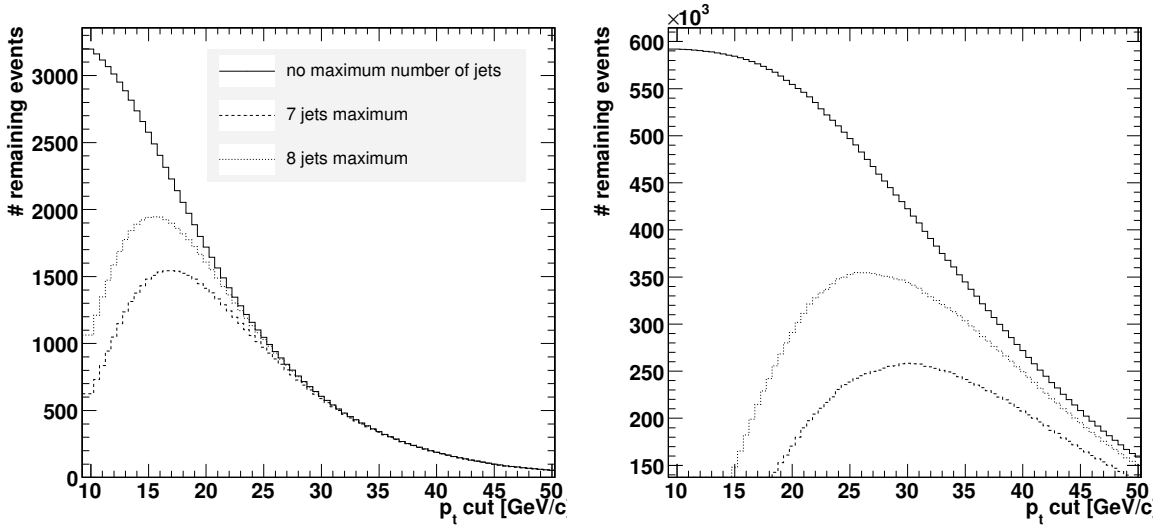

Figure 5.22: Number of selected events in dependence on the cut on jet p_t. On the left: $t\bar{t}H$ signal. On the right: $t\bar{t}4j$ background. The Number of selected events is scaled to an expected number of events after an integrated luminosity of 60 fb^{-1}. No b-tagging cuts are applied. The events are preselected using HLT. The solid line shows the requirement of at least 6 jets, while the dashed lines include also a cut on the maximum number of jets that are passing the p_t cut.

number of 6 jets requirement. A low p_t cut leads to more jets passing the p_t cut and therefore to more rejected events due to the maximum number of 7 or 8 jets, respectively. It is visible that a cut on the maximum number of jets is more effective in the case of the $t\bar{t}4j$ background and can therefore be used to suppress this background. The best working point in terms of signal selection efficiency and rejection of the $t\bar{t}4j$ background is therefore around 16 GeV/c, if a cut on the maximum number of jets is applied. To verify this hypothesis, the purity[2] S/B and significance $S/\sqrt{B}$, where S denotes the number of signal events and B the total number of background events, is shown in Figure 5.23. These plots include all relevant backgrounds ($t\bar{t}N$j , $t\bar{t}b\bar{b}$ and $t\bar{t}Z$). A b-tagging preselection as described earlier in this Section has been applied. The optimal working point is around 20 GeV/c with a 7 jets maximum cut. For the subsequent analysis, a cut of $p_t > 20$ GeV/c has been chosen.

5.4.2 Reconstruction of the Neutrino

Since the neutrino does not interact with the detector, its momentum components have to be determined using missing energy. In the ideal case of only one single neutrino and a perfect energy measurement, the neutrino's momentum would be equal to the missing energy. However, in a hadron collider experiment, the longitudinal component of the missing energy cannot be measured, because the initial state of the interaction is a priori unknown, and the proton's remnants are not accessible since they go down the beam pipe.

[2]Actually, the purity is not defined as S/B but as $S/(S+B)$. In this analysis, the number of signal events is always much smaller than the number of background events, $S \ll B$, so that this makes no difference and S/B can be denoted "purity".

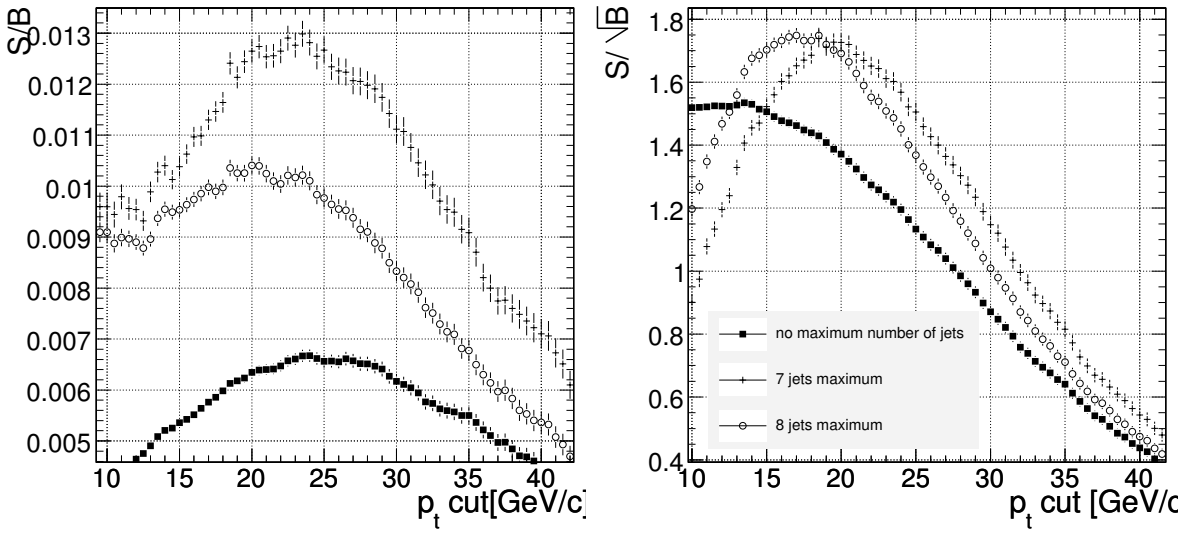

Figure 5.23: Purity S/B (left plot) and significance $S/\sqrt{B}$ (right plot) in dependence on the cut on jet p_t after an integrated luminosity of 60 fb^{-1}. A simple b-tagging cut as discussed earlier in this Section is applied. The events are preselected using HLT. The squares indicate the requirement of at least 6 jets, while the crosses and circles include also a cut on the maximum number of jets that are passing the p_t cut. All relevant backgrounds ($\mathrm{t\bar{t}}N\mathrm{j}$, $\mathrm{t\bar{t}b\bar{b}}$ and $\mathrm{t\bar{t}Z}$) are taken into account.

Therefore, only the transverse components of the missing energy, E_T, are available. It is possible to calculate the longitudinal component by imposing the requirement that the neutrino's plus the lepton's fourvector ($\nu^\alpha + l^\alpha$) have to constitute the W boson's fourvector W^α:

$$W^\alpha = \nu^\alpha + l^\alpha$$

After applying a W mass constraint

$$W^\alpha W_\alpha = m_W^2$$

and assuming the W mass to be the generated W mass of $m_W = 80.45$ GeV/c^2, a quadratic equation is obtained for the longitudinal component of the neutrino's fourvector:

$$(E_l + E_\nu)^2 - (\vec{p}_l + \vec{p}_\nu)^2 = m_W^2$$

$$p_z^\nu = \frac{1}{\left(p_t^l\right)^2}\left(\xi p_z^l \pm \sqrt{\left(\xi p_z^l\right)^2 + \left(E_l p_t^l p_t^\nu\right)^2 - \left(\xi p_t^l\right)^2}\right), \tag{5.4}$$

where

$$\xi \equiv \frac{m_W^2}{2} + p_t^l p_t^\nu \cos\left(\phi_l - \phi_\nu\right).$$

In general, this equation yields two solutions, which are both used as an interpretation possibility in the analysis described in subsequent sections. In 32% of the cases, however, the formula does not give a solution, because of a negative sign under the square root. This

happens in the case when the assumed W boson mass m_W is too far away from the real value, because the width of the W boson and the detector resolutions of E_T and the lepton are neglected. In the cases where Eq. 5.4 does not give a solution, the square root is assumed to be zero. This assumption reduces the p_z^ν resolution only by 6% as shown in Figure 5.24. This figure shows the resolution of the neutrino's z-component in the case where Equation 5.4 gives one or two solutions. If two solutions are found, only the one which is closer to the generated value is taken for this plot. This choice is ligitimate since the goal is to show the decrease of the resolution due to the discussed method. The reduction of the resolution is rather small and comes at the benefit of having a solution for 100% of the events.

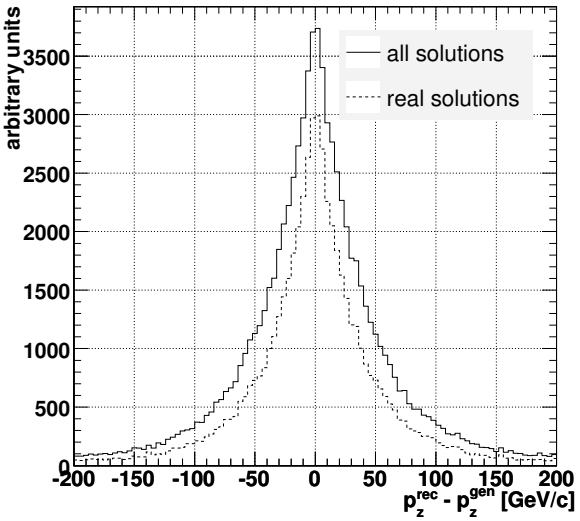

Figure 5.24: Resolution of the neutrino's z-component, i.e. $p_z^{rec} - p_z^{gen}$ for the case, when Equation 5.4 gives one or two solutions (called "real solutions" in the legend and shown as dashed line). The Root Mean Square (RMS) is 54 GeV/c in this case. The distribution represented by the solid line includes also the cases when Equation 5.4 does not give a solution and the square root is assumed to be zero (called "all solutions" in the legend. The RMS is 58.6 GeV/c in this case.

5.4.3 b-Tagging Likelihood

The $t\bar{t}H$ analysis depends primarily on the performance of the identification of the four b-jets. The $t\bar{t}$ plus light flavour jets background has a very large cross section and is being rejected through b-tagging. The improvements of the b-tagging algorithms themselves have been introduced in Section 4.2.7 and the preselection of events based on a simple b-tagging discriminator cut has been discussed in Section 5.4.1. In the present Section, the event selection and background suppression is optimized using a likelihood ratio method that exploits information about the distribution of the b-tagging discriminator of the four jets with the highest discriminator values. This is illustrated in Figure 5.25 which shows the ordered distributions of the b-tagging discriminator values for real b-jets and for non-b-jets, after all preselection cuts.

Analogously to the pairing likelihood, Equation 5.5 in Section 5.4.4, a likelihood ratio

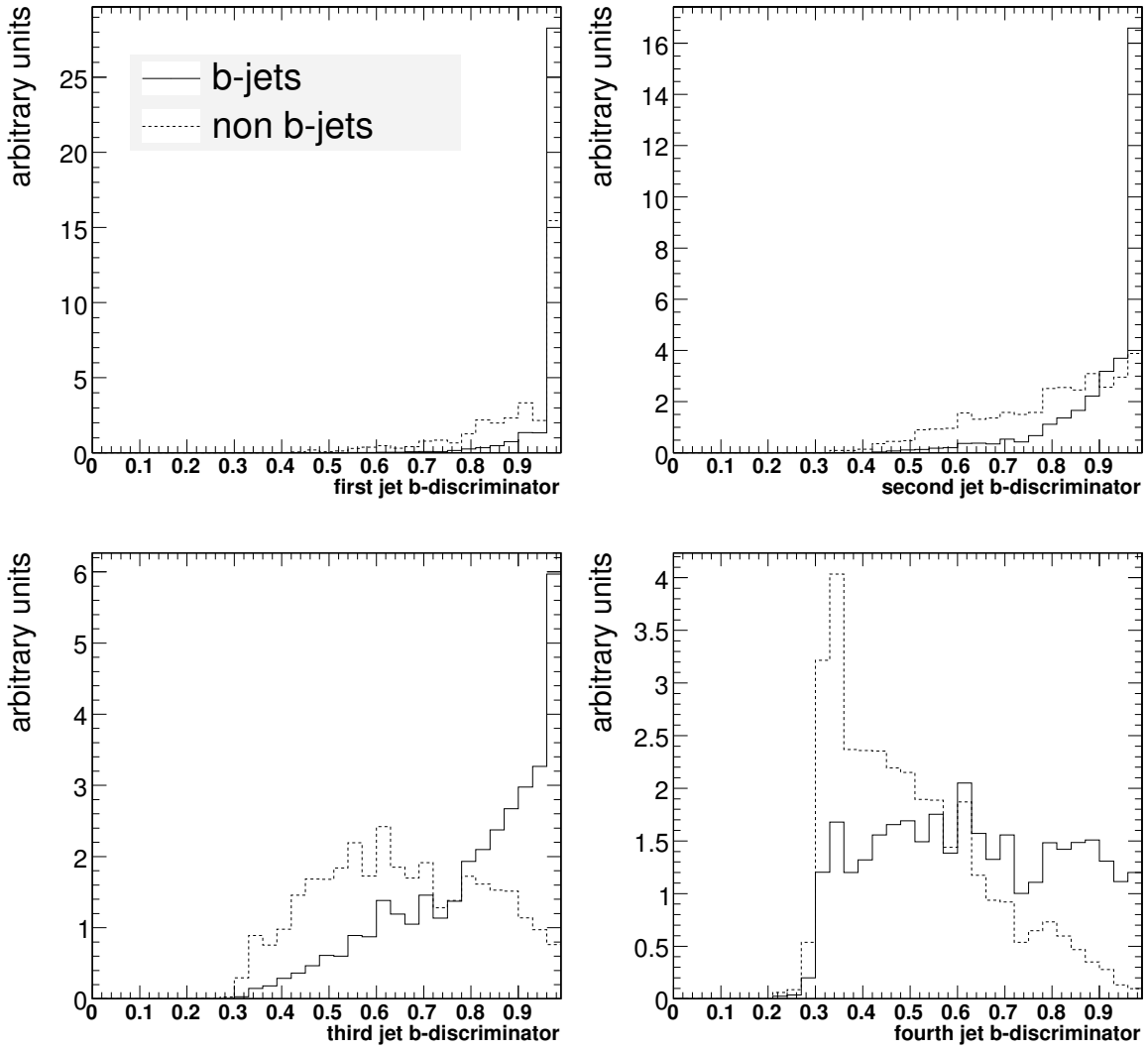

Figure 5.25: Ordered distributions of the b-tagging discriminators of the four jets with the highest values, for b-jets and non-b-jets. The top left plot shows the first jet, while the bottom righ plot shows the fourth jet. All preselection cuts are applied. The statistical fluctuations are a result of the preselection and the small number of non-b-jets in the $\mathrm{t\bar{t}H}$ signal event sample that has been used for these plots.

L_{bTag} is constructed using the distributions in Figure 5.25. This way, the b-tagging information of four jets is combined into one single discriminator L_{bTag}, which simplifies the identification of the optimal b-tagging working point, because there is no need to adjust four b-tagging cuts simultaneously. In addition, also the information about non-b-jets is taken into account, leading to an improved performance of the L_{bTag} method compared to a simple b-tagging discriminator cut. This is illustrated in Figure 5.26, which shows the distribution

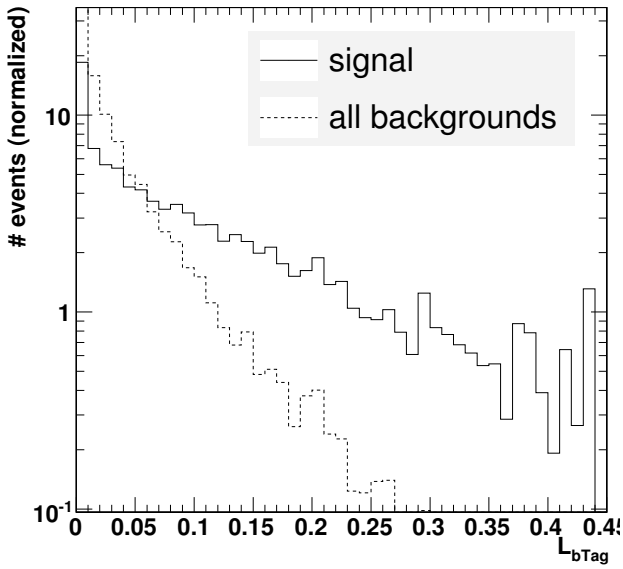

Figure 5.26: Distribution of L_{bTag} for signal and all background events.

of the L_{bTag} variable for $t\bar{t}H$ signal events and all background events. The resulting performances of the L_{bTag} cut in terms of purity S/B and significance $S/\sqrt{B}$, compared to a simple sliding cut on the b-tagging discriminator are shown in Figure 5.27 and 5.28, respectively.

It is visible that the L_{bTag} method reaches significances of about 2.5, while the discriminator cut stays below 2.4. This corresponds to an increased performance of about 8%.

It should be noted that there is an alternative approach to this b-tagging likelihood method: instead of the ordered b-discriminator distributions of b-jets and non-b-jets in $t\bar{t}H$ signal events only, it is also possible to compare the ordered b-discriminator distributions in $t\bar{t}H$ signal to $t\bar{t}N$j background events. It has been verified that the distributions in this case look similar and that the results do not differ significantly.

5.4.4 Jet Pairing Likelihood

One of the challenges of this analysis is the identification of the two jets from the Higgs boson decay which has to be performed in an environment with at least 6 jets and their according b-tagging probabilities. One reason why the search for $H \rightarrow b\bar{b}$ is being carried out in association with $t\bar{t}$ is the advantage of the availability of the signature of two top quarks. The reconstruction of the $t\bar{t}$ system is facilitated by the presence of four resonances, the two top quarks and the two W bosons, that can be exploited in order to identify the correct jet assignments.

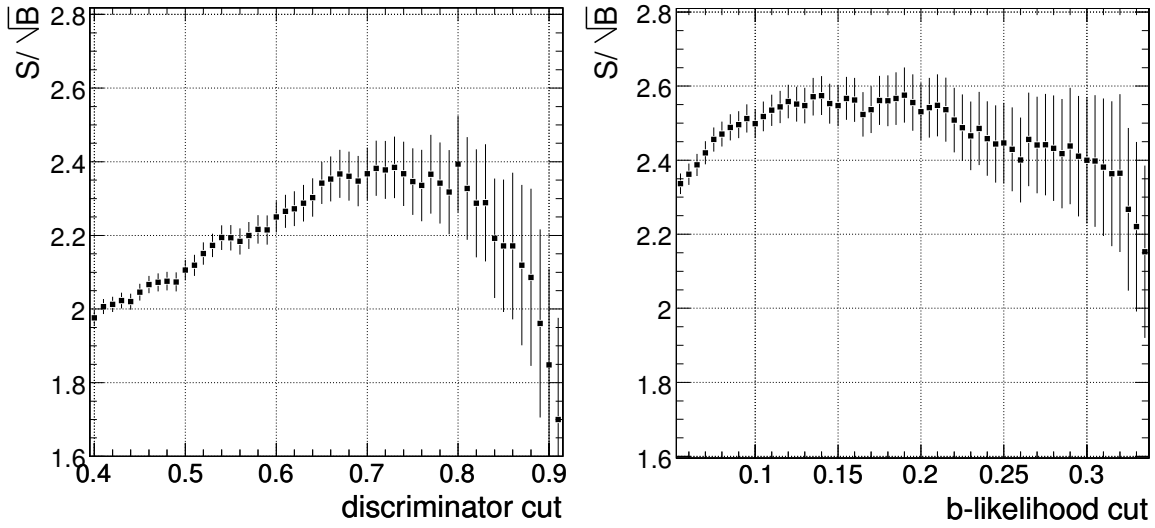

Figure 5.27: Significance $S/\sqrt{B}$ in dependence on b-tagging cuts. On the left: A simple sliding cut on the b-tagging discriminators of the four jets with the highest discriminator values. Four jets have to pass this cut. The cut is the same for all jets. On the right: Cut on the L_{bTag} likelihood ratio. The $\mathrm{t\bar{t}H}$ signal sample that has been used for these plots assumes a Higgs boson mass of $m_H = 120$ GeV/c^2. All relevant backgrounds ($\mathrm{t\bar{t}}N\mathrm{j}$, $\mathrm{t\bar{t}b\bar{b}}$ and $\mathrm{t\bar{t}Z}$) are taken into account.

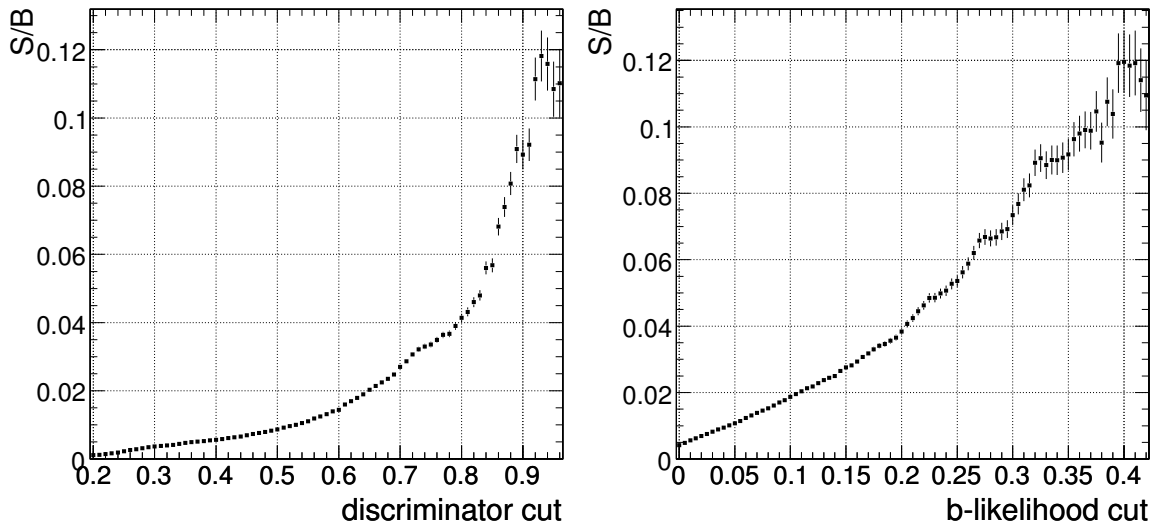

Figure 5.28: Purity S/B in dependence on b-tagging cuts. Otherwise identical to Figure 5.27.

In addition, some kinematic observables, like the angle between the W boson and the according top quark as well as b-tagging information can be used to fully identify the $t\bar{t}$ system. After the reconstruction of the $t\bar{t}$ system, the two remaining b-jets are assigned to the Higgs boson.

To summarize, the following observables are used for the identification of the $t\bar{t}H$ system:

- Invariant mass of the hadronically decaying top quark, m_{tHad}
- Invariant mass of the leptonically decaying top quark, m_{tLep}
- Invariant mass of the hadronically decaying W boson, m_W
- ΔR distance between the b-jet of the hadronically decaying top quark and the W boson, $\Delta R(\mathrm{b}, \mathrm{W}_{had})$
- ΔR distance between the b-jet of the leptonically decaying top quark and the lepton, $\Delta R(\mathrm{b}, \mathrm{l})$

The mass of the leptonically decaying W boson cannot be used because it has a fixed value which is used for the determination of the longitudinal momentum of the neutrino as described in Section 5.4.2. This is also the reason, why the angular distance in the leptonic case is calculated using the lepton and not the W boson fourvector.

The usage of the ΔR distance between the b-jet and the W boson stemming from the top quark decay is motivated by the transversal boost of the decaying system. The transverse momentum of the top quark is shown in Figure 5.4 and has a mean value of $\overline{p}_t^{top} = 165.8$ GeV/c leading to a mean value of $\overline{\Delta R} = 1.7$ and a peak at $\Delta R = 1.3$ for correct jet combinations.

The distributions of these observables are obtained using an angular matching of jets to generator partons. The matching algorithm calculates the ΔR distance of all possible parton-jet combinations and subsequently removes the best matches until each generator parton has one matched jet. Only events in which all generator partons are well matched to reconstructed objects within $\Delta R < 0.5$ are used for the construction of the likelihood. This task has been performed using the $t\bar{t}H$ event sample with a realistic event preselection, i.e. the event has to be triggered and the signal jets are required to pass the simple $p_t > 18$ GeV/c and $|\eta| < 3$ cuts with at least 6 and maximally 7 jets. This corresponds to a fraction of 11% of the events. The requirement of a successful matching within $\Delta R < 0.5$ reduces the fraction to about 8%. No b-tagging cuts are applied at this level, since this would reduce the selection efficiency too much.

In order to construct a likelihood ratio as described below, it is beneficial to also include the information about wrong jet assignments. Wrong combinations are obtained by exchanging one or more of the correct jets by another jet passing the selection cuts in such a way that all possible wrong permutations occur once.

The resulting distributions of these observables for correct as well as wrong jet pairings are displayed in Figures 5.29 to 5.31. These distributions have been obtained in the $t\bar{t}H$ signal sample with $m_H = 120$ GeV/c^2. The dependencies of these variables on the Higgs boson mass is negligible in the considered mass range, hence, no differentiation has been made for the various Higgs mass hypotheses.

It is visible that all of these variables have some discriminating power. In analogy to the method in case of the lepton identification in Section 5.3.2, a global event likelihood is

constructed by combining all these distributions into one likelihood ratio L:

$$L = \Pi_i \frac{P_i^{sig}(x_i)}{P_i^{sig}(x_i) + P_i^{bkg}(x_i)}. \tag{5.5}$$

The resulting distributions of this jet pairing likelihood for correct and wrong combinations are shown in Figure 5.32. For these Figures the events have been preselected according to Section 5.4.1, including a cut on the b-tagging likelihood of $L_{bTag} > 0.225$. It is visible, that the distribution of the best, i.e. highest, likelihood ratio has a significantly higher mean value. This is due to the fact that even with a small number of available jets, i.e. 6 or 7, a very high number of different jet combinations, i.e. 180 or 630, is possible. Especially in events that have, for example, a top mass value far away from the mean value, the correct likelihood ratio might be very small, while the chance to find a higher likelihood value using another (wrong) combination might be quite large. This is illustrated by the right side of Figure 5.32, which displays the "rank" of the likelihood value of the correct jet pairing, compared with all possible likelihood values of wrong jet pairings. Only in roughly 13% of the cases, the highest value is also the correct one. This number might look rather small, but it is expected. In principle, this fraction of 13% represents the so called "pairing efficiency" for the complete reconstruction of the $\mathrm{t\bar{t}H}$ system. Usually, the numbers quoted for a $\mathrm{t\bar{t}H}$ pairing efficiency refer to the efficiency of finding the correct jets for the Higgs boson only. Even in the case where the $\mathrm{t\bar{t}}$ system is not well reconstructed, it is still possible to find the correct jets for the Higgs boson. This is illustrated by Figure 5.33, which shows the likelihood rank for a correct pairing of the jets of the Higgs boson only. In this special setup, 26% of the events are cumulated at the rank one which means that the pairing efficiency is 26%. It is worth noting that also in this case the entries at rank 0 correspond to events in which no correct jet pairing could be found, i.e. that the available jets passing the cuts could not be assigned within $\Delta R < 0.5$ to the generator partons. The term "correct pairing" of a jet means that the

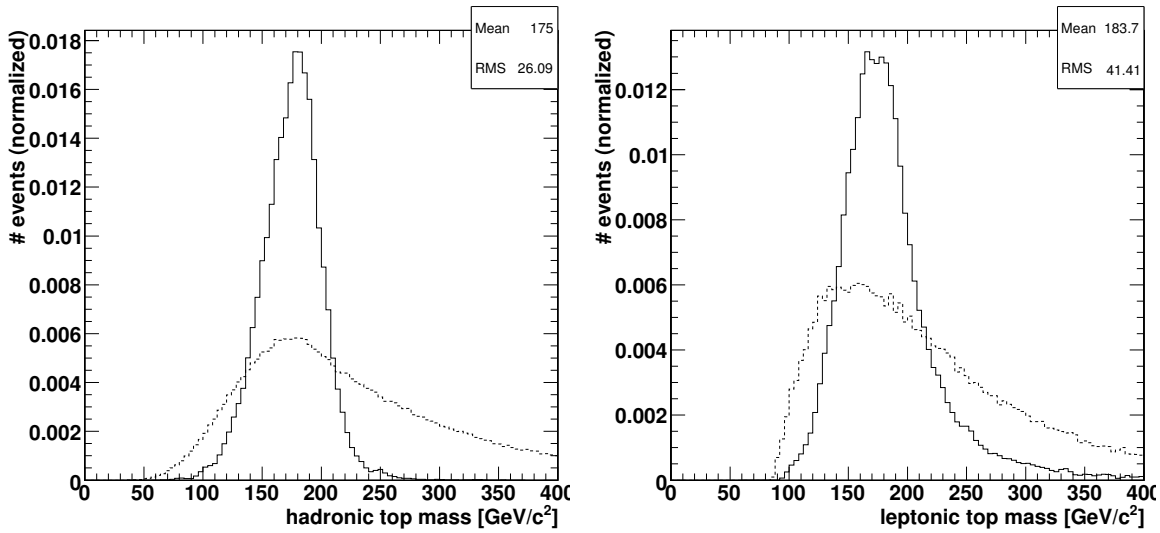

Figure 5.29: Hadronic (left) and leptonic (right) top masses for correct jet assignments (solid line) and wrong jet assignments (dashed line). The mean and RMS values in the histogram are the values for the correct jet assignments.

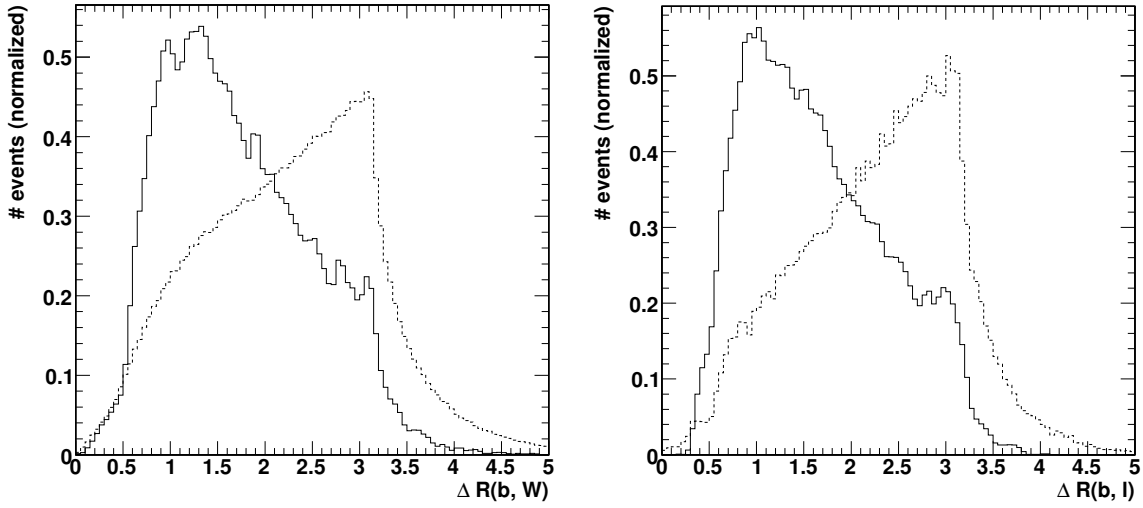

Figure 5.30: On the left: Angular distance ΔR between the b-jet of the hadronically decaying top quark and the W boson. On the right: ΔR between the b-jet of the leptonically decaying top quark and the lepton. As before, the solid line refers to correct jet assignments, while the dashed line represents wrong jet assignments.

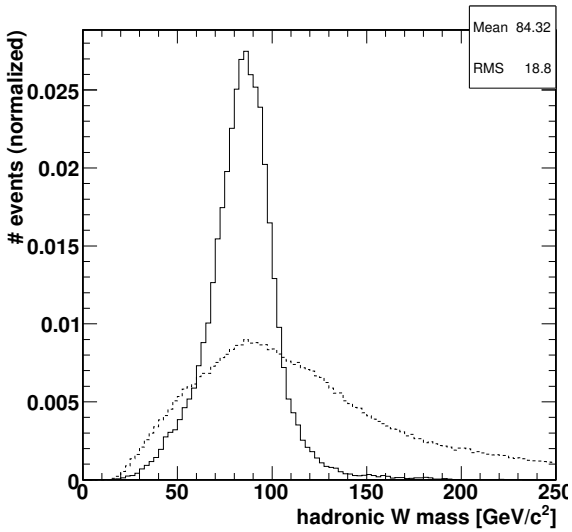

Figure 5.31: Invariant mass of the hadronically decaying W boson for correct jet assignments (solid line) and wrong jet assignments (dashed line).

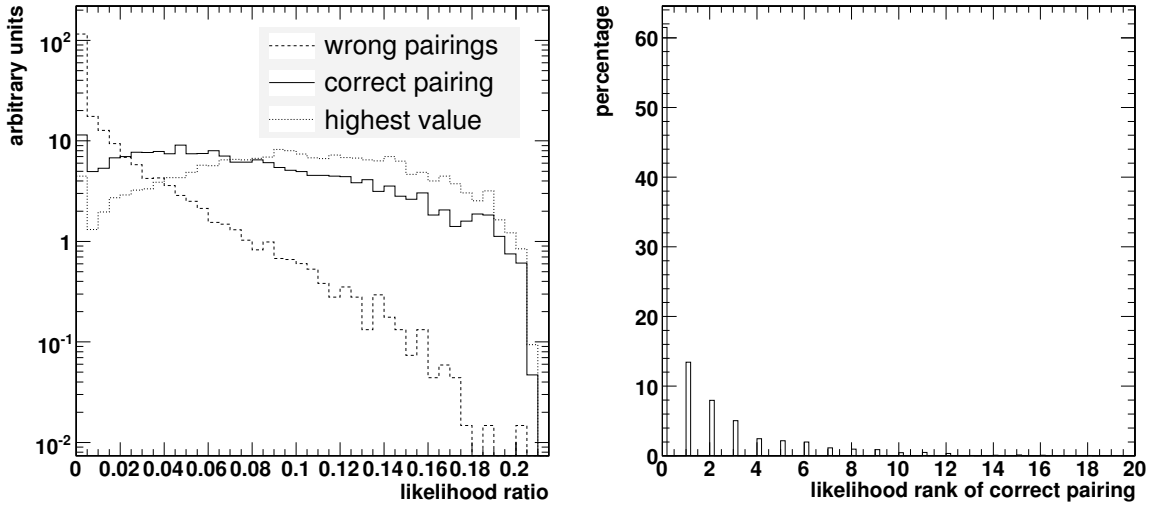

Figure 5.32: On the left: Distributions of the likelihood ratio for correct jet pairings, for all wrong pairings and for the pairing with the best likelihood value (denoted "highest value"). On the right: The "rank" of the correct jet pairing, ordered by the value of the likelihood ratio. The rank 0 means that no correct jet pairing has been found in the respective event, i.e. that it was not possible to match the jets to primary partons within the given ΔR distance.

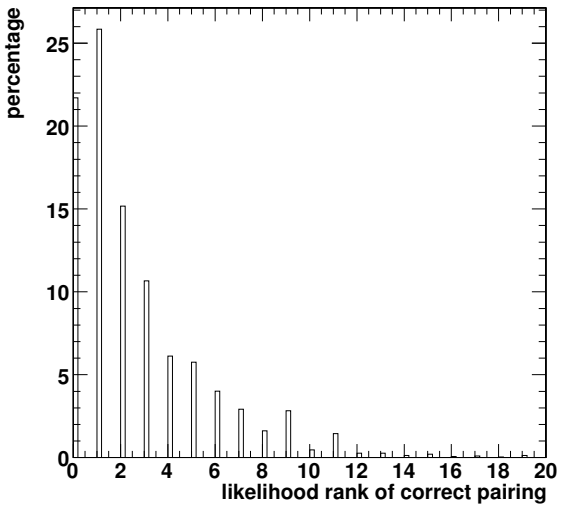

Figure 5.33: Likelihood "rank" of the correct jet pairing of the Higgs boson. Rank 0 means that it was not possible to match the two primary partons of the Higgs boson to jets within the given ΔR distance.

angular distance between jet and its corresponding generator parton is smaller than $\Delta R < 0.5$.

Several attempts to increase the pairing efficiency for the Higgs boson have been tried out. Among these attempts are kinematic fits [87], and complex exploitations of the kinematical characteristics of $t\bar{t}H$ events [2]. Unfortunately, none of these methods was able to increase the pairing efficiency significantly above 30%. Even a large increase of e.g. 10% in the pairing efficiency would not change the final result of the analysis. This means that the invariant Higgs mass distribution displayed in Figure 5.35 would still show a broad signal distribution above a similarly shaped distribution of the background. Therefore, no further attempts to increase this pairing efficiency have been made.

5.5 Discussion of the Results

In this Section, the results for the semileptonic muon- and electron-channel, including all previously discussed optimizations are summarized. The variable with the largest impact on the final result is the cut on the b-tagging likelihood L_{bTag}. Therefore, the expected observability in terms of significance $S/\sqrt{B}$ and purity S/B is shown in dependence on the cut on L_{bTag} in Figure 5.34 for three different Higgs boson mass hypotheses, 115, 120 and 130 GeV/c^2 after an integrated luminosity of 60 fb^{-1}. These plots refer to the full invariant Higgs mass range without applying a mass window. It is visible that the significance reaches its maximum at a cut between 0.125 and 0.225. The invariant Higgs boson mass in the case of $m_H = 115$ GeV/c^2 is shown on the left side of Figure 5.35 in comparison to the combinatorial background . "Combinatorial background" refers to events in which the two b-jets assigned to the Higgs boson are not within $\Delta R < 0.5$ to the generated jets. The right side of this

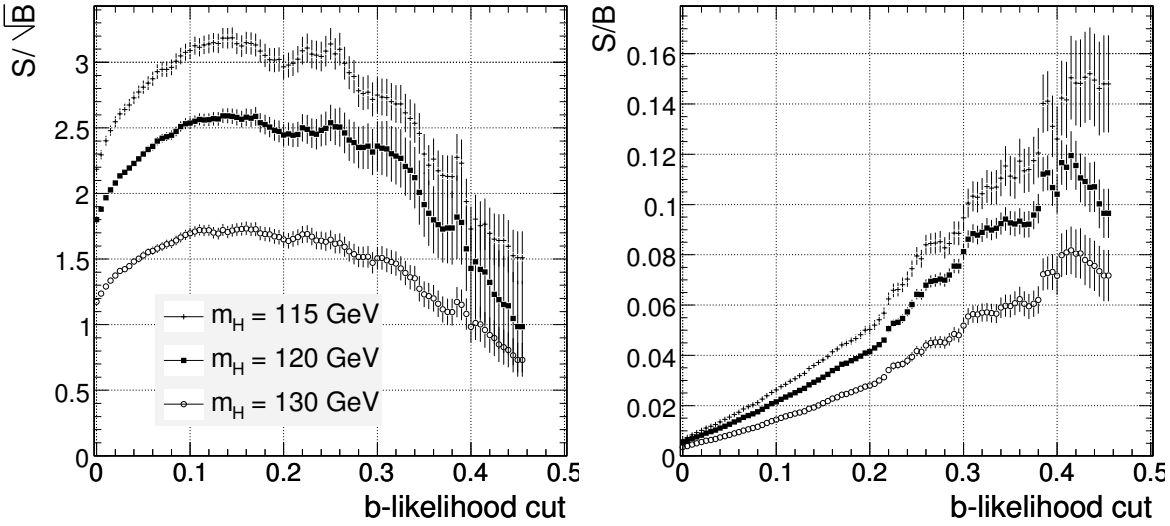

Figure 5.34: Observability in terms of significance $S/\sqrt{B}$ (left plot) and purity S/B (right plot) for three different Higgs boson mass hypotheses (115, 120 and 130 GeV/c^2) in dependence on the cut on the b-tagging likelihood L_{bTag}, after an integrated luminosity of 60 fb^{-1} for the semileptonic (muon and electron) $t\bar{t}H$ decay channel. No mass window has been applied. The error bars indicate the statistical error due to the finite size of datasets. Also here, bin-to-bin correlations occur because of the sliding cut on the b-tagging likelihood.

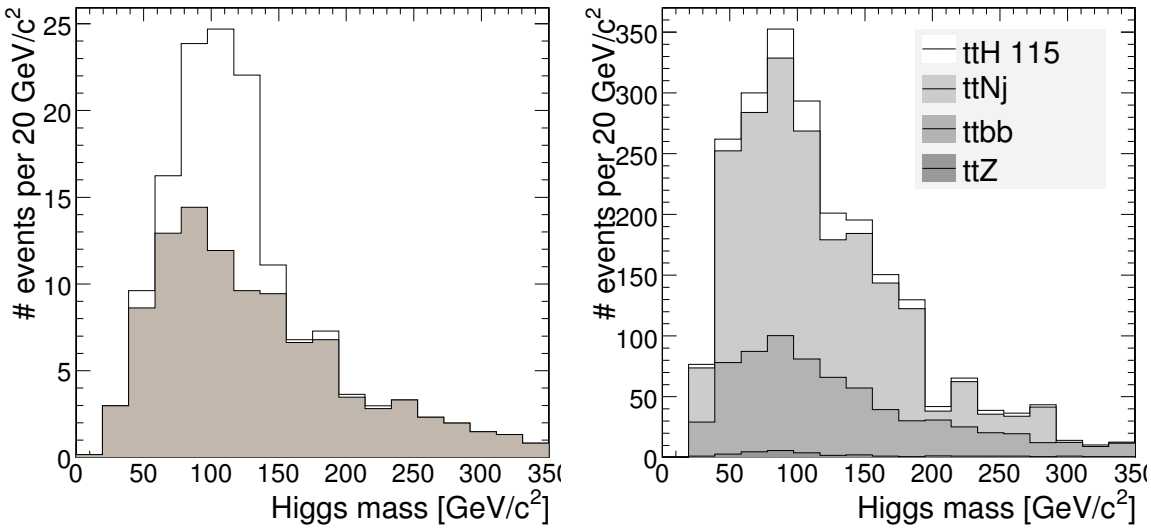

Figure 5.35: Invariant Higgs boson mass spectrum for a L_{bTag} cut of 0.225 and $m_H = 115$ GeV/c^2, after an integrated luminosity of 60 fb^{-1}. On the left: Only signal events; the fraction of combinatorial background is shaded grey. On the right: All relevant physical backgrounds ($\mathrm{t\bar{t}Z}$, $\mathrm{t\bar{t}b\bar{b}}$ and $\mathrm{t\bar{t}}N\mathrm{j}$) and the $\mathrm{t\bar{t}H}$ signal (including the combinatorial background) stacked on top of each other.

Figure shows the physical backgrounds ($\mathrm{t\bar{t}Z}$, $\mathrm{t\bar{t}b\bar{b}}$ and $\mathrm{t\bar{t}}N\mathrm{j}$) and the $\mathrm{t\bar{t}H}$ signal stacked on top of each other. Due to the limited amount of available Monte Carlo statistics for the $\mathrm{t\bar{t}}N\mathrm{j}$ background and the large scale factors that have to be applied to the remaining events, the statistical fluctuations in Figure 5.35 are large. The according distributions for the other two Higgs boson mass hypotheses, $m_H = 120$ GeV/c^2 and $m_H = 130$ GeV/c^2 are shown in Figures C.1 and C.2 in Appendix C.

Compared to the original publication of this analysis [2], Figure 5.34 indicates some improvement which is due to the advancement of the b-tagging algorithms, introduced in Section 4.2.7, and cut optimizations, discussed in Section 5.4.1. The maximum significance of $S/\sqrt{B} = 2.34$ could be increased to 2.6 for $m_H = 120$ GeV/c^2. This corresponds to an improvement of 11%. More impressive is the improved purity at $S/\sqrt{B} = 2.34$ (the result qoted in [2]), which increased from $S/B = 4.2\%$ to 8% which corresponds to a relative increase of more than 90%. Since the purity is the deciding factor as soon as systematic errors are taken into account, one can conclude that the improvements are significant. However, the overall picture concerning the feasibility of this analysis does still not change much.

For the sake of completeness and to be able to compare with the tables in [2], Table 5.9 gives event numbers and selection efficiencies for the semileptonic electron- and muon-channel at two different working points. One "loose" working point, which optimizes $S/\sqrt{B}$, and a "tight" working point, which increases the purity up to the statistical limits, are presented.

In this context it is worth noting that the statistical fluctuations of the number of background events has only moderate impact on the final significance. This is shown in Figure 5.36 which displays the significance expressed as

$$\sigma = \frac{S}{\sqrt{B + \mathrm{d}B}}$$

in dependence on the fractional uncertainty of the number of background events, where dB is the variation of the background prediction. From these figures one can deduce that the hereby stated significances are approximately valid even if the number of background events is wrong by some percent.

The reconstruction of the invariant Higgs boson mass is facilitated by the usage of the described pairing likelihood that selects the two b-jets that can be assigned to the Higgs boson with the highest probability. This procedure also motivates the application of a mass window for the calculation of the final significance. When using the requirement of $m_H < 150$ GeV/c^2, the purity can be increased by about 10%, while the significance does not change much. This minor improvement is due to the shape of the Higgs mass peak shown on the left side of Figure 5.35 which is similar to the shape of the physical background. It should be noted that the reconstruction of a mass peak is important as soon as fitting strategies extracting the significance from the shape of the peak are applied. This will become relevant when background subtraction methods are available.

This leads to the assessment that currently, the analysis of $t\bar{t}H$ with $H \rightarrow b\bar{b}$ has to be carried out in the form of a counting experiment, which heavily relies on the knowledge of event rates and their corresponding systematic errors as evaluated in Section 5.8. Since a precise prediction of the background rates is not possible at present, because NLO calculations for $t\bar{t}$ plus two or more jets are not available yet, the $t\bar{t}H$ analysis will have to rely on the

Table 5.9: Selection efficiency for $L_{bTag} > 0.225$ (ϵ_{loose}) and for $L_{bTag} > 0.350$ (ϵ_{tight}), number of expected events and signal significance in 60 fb^{-1} for the muon and electron $t\bar{t}H$ channels. The signal datasets are labeled by the generated Higgs mass in GeV/c^2 (parentheses). Also quoted are binomial errors arising from the finite sizes of processed datasets. No Higgs mass window has been applied. The last column of $t\bar{t}$4j gives the upper limit corresponding to a confidence level of 68% since no events are remaining after the cuts in this case.

	# Events	ϵ_{loose} (%)	N^{ev}_{loose}	ϵ_{tight} (%)	N^{ev}_{tight}
$t\bar{t}H$ (115)	55395	1.60 ± 0.05	147 ± 5	0.5 ± 0.03	48 ± 3
$t\bar{t}H$ (120)	191133	1.55 ± 0.03	118 ± 2	0.52 ± 0.016	40 ± 1
$t\bar{t}H$ (130)	44595	1.70 ± 0.06	80 ± 3	0.54 ± 0.03	25 ± 2
$t\bar{t}$1j	1297064	0.0045 ± 0.0006	464 ± 60	0.00046 ± 0.0002	47 ± 19
$t\bar{t}$2j	827615	0.0089 ± 0.00103	536 ± 62	0.0011 ± 0.00036	65 ± 22
$t\bar{t}$3j	108778	0.014 ± 0.0035	331 ± 85	0.0028 ± 0.0016	66 ± 38
$t\bar{t}$4j	114054	0.0035 ± 0.0017	128 ± 64	0	< 36
$t\bar{t}b\bar{b}$	384407	0.43 ± 0.01	734 ± 18	0.141 ± 0.006	239 ± 10
$Zt\bar{t}$	94706	0.104 ± 0.011	35 ± 4	0.029 ± 0.005	10 ± 2
Total Backgr.			2230		427
$S/\sqrt{B}$ (115)			3.1		2.3
S/B (115)			6.6%		11%
$S/\sqrt{B}$ (120)			2.5		1.9
S/B (120)			5.3%		9.3%
$S/\sqrt{B}$ (130)			1.7		1.23
S/B (130)			3.6%		5.9%

measurement of the background from data.

5.5.1 Comparison to Previous Results and Expected Suppression for $\mathrm{t\bar{t}}N\mathrm{j}$

Compared to previous studies [88, 89, 74, 75], the $\mathrm{t\bar{t}}$ plus light flavour jets background proved to be dramatically more dominant (by more than a factor of three) than found earlier. This is in fact the sole key to understand the differences to these preceding results. This section shows that, based on the b-tagging performance presented in Section 4.2.7, the $\mathrm{t\bar{t}}N\mathrm{j}$ background rates obtained in this thesis are in agreement with predicting calculations. They are not in contradiction with the preceding results if the conditions of the b-tagging simulations are taken into account.

In order to calculate the predictions for the event rates due to the b-tagging efficiencies, the jet composition of the event samples has to be decomposed. The type of the W boson decays (semileptonic, di-leptonic, or all-hadronic) determines the number of jets in the event samples. Table 5.10 shows the branching ratio of each W boson decay mode. This table has to be compared to Table 5.11 which shows the relative contribution of the W boson decay modes to the background samples after all preselection cuts, except for b-tagging, i.e. $p_t > 18$ GeV/$c, |\eta| < 3$, including HLT and most important, the lepton selection based on the lepton likelihood for the semileptonic channel. The table shows that the preselection enriches the semileptonic fraction in the expected way. The overall preselection efficiencies due to HLT, the cuts on the lepton likelihood, jet p_t, number of jets and $|\eta|$ cuts are quoted in Table 5.12.

For the following calculation, the number of jets and their corresponding flavour compo-

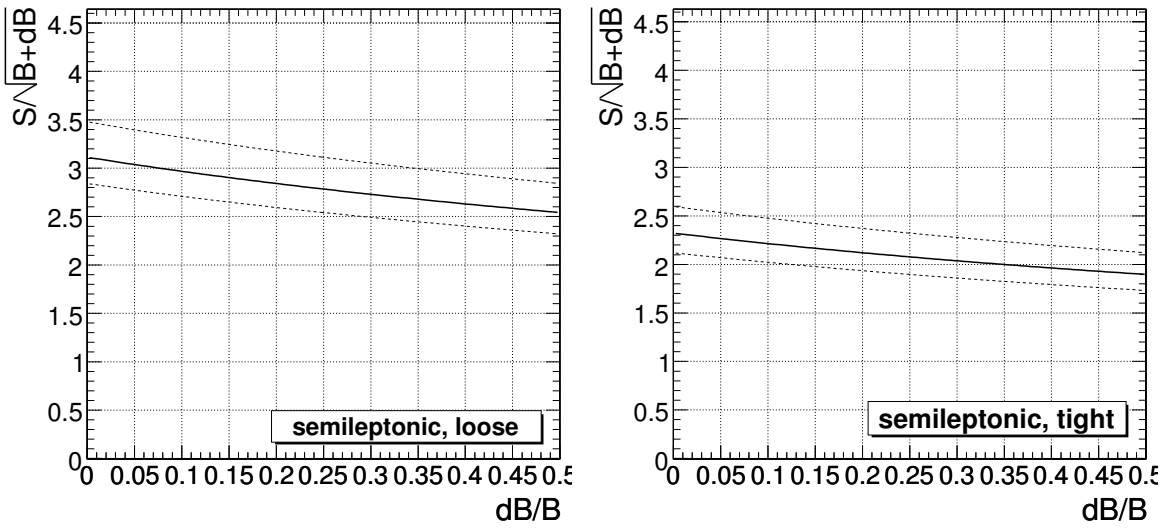

Figure 5.36: Significance $S/\sqrt{B + \mathrm{d}B}$ in dependence on the fractional uncertainty of the number of background events for a hypothetical Higgs boson mass of $m_H = 115$ GeV/c^2 in 60 fb^{-1}. The variation of the background prediction is denoted dB. The dashed lines correspond to a variation of the background cross section of 20% due to the theoretical uncertainty. On the left: "loose" working point, on the right: "tight" working point.

sition has to be known. For this purpose, the reconstructed jets after a preselection identical to the selection in Section 5.4.1 is applied, i.e. the jets are required to have $p_t > 18$ GeV/c, $|\eta| < 3$ with a maximum number of 7 and minimum of 6 jets passing the cuts. Table 5.13 shows the fractional contribution of events with either 6 or 7 jets for the various samples. For the identification of the true jet flavour, the tools described in Section 4.2.6 have been used with the "algorithmic" definition of the true jet flavour. The distribution of the jet flavour for the various samples is given in Table 5.14.

The last ingredient which is necessary are the b-tagging and misidentification efficiencies which can be extracted from Section 4.2.7. The working points used in the analysis are summarized in Table 5.15.

In the following calculations, the information in these tables is combined to evaluate the expected suppression factors. To begin with, the probability of b-tagging exactly i out of n_b b-jets is $C_i^{n_b}\varepsilon_b^i(1-\varepsilon_b)^{n_b-i}$, where $C_i^{n_b}$ is the combinatorial factor for the number of ways i jets out of n_b can be assigned without regards to order.

$$C_i^N \equiv \frac{N!}{i!(N-i)!}. \tag{5.6}$$

After including also the mistagging probabilities, Equation 5.7 gives the probability of b-tagging exactly n jets out of n_b b-jets, n_c charm jets and n_l light flavour jets.

$$\begin{aligned}\varepsilon^n(n_b,n_c,n_l) &= \sum_{i=0}^{n}\sum_{j=0}^{n-i}\left[C_i^{n_b}\varepsilon_b^i\left(1-\varepsilon_b\right)^{n_b-i}\right]\left[C_j^{n_c}\varepsilon_c^j\left(1-\varepsilon_c\right)^{n_c-j}\right] \\ &\quad\times\left[C_{n-i-j}^{n_l}\varepsilon_l^{n-i-j}\left(1-\varepsilon_l\right)^{n_l-(n-i-j)}\right]\end{aligned} \tag{5.7}$$

Table 5.10: Branching Ratios of the three different W boson decay modes. Only muon or electron decays of the W bosons are considered in the leptonic fractions.

decay mode	branching ratio
semileptonic	30%
di-leptonic	5%
all-hadronic	50%
tau modes	15%

Table 5.11: Relative contribution of the W boson decay modes to the background samples after preselection. The $t\bar{t}H$ sample does only contain semileptonic events. The numbers are given in percent.

sample	semileptonic	di-leptonic	all-hadronic
$t\bar{t}H$	100	0	0
$t\bar{t}1j$	95.7	3.9	0.4
$t\bar{t}2j$	94.9	4.7	0.27
$t\bar{t}3j$	92.7	7.1	0.14
$t\bar{t}4j$	87.6	12.2	0.19

Table 5.12: Efficiency of the preselection due to HLT, the cuts on the lepton likelihood, jet p_t, number of jets and $|\eta|$.

sample	preselection efficiency
$\mathrm{t\bar{t}H}$ ($\epsilon^{pre}_{\mathrm{t\bar{t}H}}$)	23.3%
$\mathrm{t\bar{t}1j}$ ($\epsilon^{pre}_{\mathrm{t\bar{t}1j}}$)	4.97%
$\mathrm{t\bar{t}2j}$ ($\epsilon^{pre}_{\mathrm{t\bar{t}2j}}$)	7.79%
$\mathrm{t\bar{t}3j}$ ($\epsilon^{pre}_{\mathrm{t\bar{t}3j}}$)	8.71%
$\mathrm{t\bar{t}4j}$ ($\epsilon^{pre}_{\mathrm{t\bar{t}4j}}$)	4.1%

Table 5.13: Fractional contribution of events with 6 or 7 jets passing the selection cuts for the three different W boson decay modes. The $\mathrm{t\bar{t}H}$ sample does only contain semileptonic events. The numbers are given in percent.

	semilep.		di-lep.		all-had.	
sample	6	7	6	7	6	7
$\mathrm{t\bar{t}H}$	61.4	38.6	-	-	-	-
$\mathrm{t\bar{t}1j}$	66.5	33.5	68.6	31.3	65.8	34.2
$\mathrm{t\bar{t}2j}$	60.6	39.3	64.0	35.9	42.1	57.9
$\mathrm{t\bar{t}3j}$	48.4	51.6	59.8	40.2	36.4	63.6
$\mathrm{t\bar{t}4j}$	33.5	66.4	36.9	63.0	42.8	57.1

Table 5.14: Flavour Composition of the jets of the various data samples. The numbers are the fractions in percent.

sample	light flavour	charm	bottom
semileptonic $\mathrm{t\bar{t}H}$	39.4	6.9	53.6
semileptonic $\mathrm{t\bar{t}1j}$	60.6	8.4	30.9
di-leptonic $\mathrm{t\bar{t}1j}$	62.2	1.5	36.3
all-hadronic $\mathrm{t\bar{t}1j}$	65.6	14	20.4
semileptonic $\mathrm{t\bar{t}2j}$	62.2	8.6	29.2
di-leptonic $\mathrm{t\bar{t}2j}$	63.9	2.1	33.9
all-hadronic $\mathrm{t\bar{t}2j}$	67.7	12.5	19.8
semileptonic $\mathrm{t\bar{t}3j}$	63.9	8.7	27.4
di-leptonic $\mathrm{t\bar{t}3j}$	65.2	3.4	31.3
all-hadronic $\mathrm{t\bar{t}3j}$	72.2	11.1	16.7
semileptonic $\mathrm{t\bar{t}4j}$	67.2	8.12	24.7
di-leptonic $\mathrm{t\bar{t}4j}$	68.6	4.5	26.8
all-hadronic $\mathrm{t\bar{t}4j}$	75	6.8	18.2

Table 5.15: Definitions of variables and values for the working points used in the semileptonic analysis.

	Definition	Value
ε_b	efficiency of b-tagging a b-jet	55%
ε_c	efficiency of b-tagging a c-jet	12%
ε_l	efficiency of b-tagging a light flavour jet	1.2%

In the exact calculation, this number has to be evaluated according to the fractions of jets present in each sample. In order to get a quick approximation of the expected efficiencies, a simplification is done in the following. According to Tables 5.11 and 5.14 it is assumed that only semileptonic events are present and that the jet flavour composition among the samples with different jet multiplicities is the same, like

- 6 jets: $n_b = 2$, $n_c = 1$ and $n_l = 3$
- 7 jets: $n_b = 2$, $n_c = 1$ and $n_l = 4$.

The suppression factor for $\mathrm{t\bar{t}2j}$ is then:

$$\epsilon_{\mathrm{t\bar{t}2j}} = \epsilon^{pre}_{\mathrm{t\bar{t}2j}} \left(0.6 \cdot \epsilon^{i=4}\left(2,1,3\right) + 0.4 \cdot \epsilon^{i=4}\left(2,1,4\right)\right) = 0.012\%, \tag{5.8}$$

which is in good agreement with the observed value at the "loose" working point in Table 5.9. Also the calculated efficiencies for the signal and the other backgrounds, which are given in Table 5.16 are in very good agreement with the observations.

Table 5.16: Selection efficiencies calculated according to Equations 5.7 and 5.8 compared to the observed efficiencies.

sample	calculated selection efficiency (%)	observed selection efficiency (%)
$\mathrm{t\bar{t}H}$	1.47	1.6
$\mathrm{t\bar{t}1j}$	0.0078	0.005
$\mathrm{t\bar{t}2j}$	0.012	0.01
$\mathrm{t\bar{t}3j}$	0.015	0.014
$\mathrm{t\bar{t}4j}$	0.0046	0.004

Furthermore, the excellent compliance of these calculations with the actually observed values is an important consistency check for the analysis as a whole. The remaining question, why this result contradicts the preceding studies is aggravated by the fact that the $\mathrm{t\bar{t}}N\mathrm{j}$ background used in this study was generated with ALPGEN and has a much smaller cross-section than the CompHEP generated samples, which have been used in the earlier analyses. This adds another factor of two to the difference.

The most important difference stems from the fact that a fast simulation of the CMS tracker had to be used in one of the publications mentioned. As shown in Section 4.4.1 the b-tagging performance in the current CMS fast simulation is not in good agreement with full

simulation. Especially the misidentification rate of light flavour jets shows a difference of up to a factor of 5. The $\mathrm{t\bar{t}}$ plus light flavour background is the dominant contribution compared to other backgrounds in the present study. As confirmed by other experiments, e.g. CDF, the light flavour misidentification rates are very hard to be correctly described by simulation programs.

Another difference is the distribution of the jet flavours shown in Table 5.14, which shows a quite large fraction of contamination with charm jets. These jets are coming from gluon splitting as well as from W boson decays. Therefore, a $\mathrm{t\bar{t}2j}$ background cannot be simply understood as consisting of b- and light flavour jets only. If a parameterized b-tagging is used, these jets have to be taken into account properly as done in the calculations in this section.

5.6 Secondary Backgrounds

In Section 5.2.1, the generation of a $\mathrm{t\bar{t}}$ background without light flavour jets has been discussed, but in the further description and presentation of the results, this background has been left out. This is justified by the observation that this background is negligible as soon as soft b-tagging cuts are applied. This is visible in Figure 5.37 which shows the number of remaining events in dependence on the cut on the L_{bTag} variable. It is visible that already at

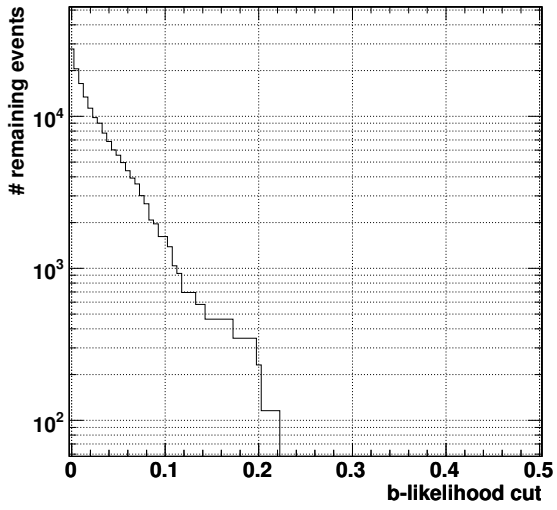

Figure 5.37: Number of remaining events in dependence on the cut on L_{bTag} for the $\mathrm{t\bar{t}0j}$ background.

the "loose" working point introduced in Section 5.5, no $\mathrm{t\bar{t}0j}$ events are passing the cut for an integrated luminosity of $L = 60$ fb^{-1}.

Another possible background source is W plus jets and WW plus jets. These backgrounds have not been considered in the full simulation and reconstruction. Since the predictions for the expected event rates of the $\mathrm{t\bar{t}}N\mathrm{j}$ backgrounds have been discussed in Section 5.5.1 and have been in very good agreement with the observations, a calculation for the event yields of the W plus jets background is presented in the following.

A conservative value for the preselection efficiency $\epsilon^{pre} = 1\%$ is assumed for all jet multiplicities, even though this might overestimate the real efficiencies. This assumption is justified by the fact that only one W boson is present and that the number of jets is very low. While no real b-jets are present in these samples, a contamination stemming from gluon splitting is always existent. The rate of $g \rightarrow b\overline{b}$ is less than 2%, therefore, a fraction of 5% b-jets is assumed.

For instance, the selection efficiency of W + 2jets is then

$$\epsilon_{\mathrm{W2j}} = \epsilon^{pre}_{\mathrm{W2j}} \left(0.05 \cdot \epsilon^{i=4}\left(1,1,4\right) + 0.95 \cdot \epsilon^{i=4}\left(0,0,6\right)\right) = 3.25 \times 10^{-6}\%, \tag{5.9}$$

where the contribution of 5% b-jets is taken into account by dividing the efficiency into one part with a b-jet, weighted with the factor of 0.05 and a part without a b-jet, weighted with 0.95. To simplify this formula, the contribution from charm jets has been taken into account by assuming also 5% as in the case of b-jets. All the W plus jets selection efficiencies including their cross sections and event yields are summarized in Table 5.17. For this table, the b-contamination is increased by 2% for each additional jet. Evidently, these predictions

Table 5.17: Selection efficiencies and yields in 60 fb^{-1} for W + N jets as predicted by Equation 5.9

	Cross-section (pb)	Efficiency (%)	Yield in 60 fb^{-1}
W + 0 jets	90000	0	0
W + 1 jet	24000	1.48×10^{-6}	21
W + 2 jets	7500	3.25×10^{-6}	14
W + 3 jets	2170	4.73×10^{-6}	6
W + 4 jets	522	1.07×10^{-5}	3
W + 5 jets	135	1.32×10^{-5}	1
W + $\geq$6 jets	180	1.57×10^{-5}	2

are small compared to the other backgrounds. The cross sections for the WW plus jets backgrounds are all below 30 pb and are therefore also negligible.

5.7 The All-Hadron and Di-Lepton Channels

The main topic of this thesis is the semileptonic channel that has been presented in the previous sections. In Reference [2] the analyses for the all-hadron and di-lepton channels are described in detail. In comparison with the semileptonic channel, these two channels do not contribute much to the overall discovery potential. In the follwing, a short overview of these two analyses is given. The expected result is that the all-hadron channel suffers too much from QCD background, while the di-lepton channel suffers from its small branching ratio but is probably more promising in a high-luminosity analysis.

5.7.1 The All-Hadron Channel

Since the all-hadron channel has to rely completely on jet reconstruction and has to deal with 8 or more jets, this can be considered as the major bottleneck of this analysis. Therefore, a

dedicated study concerning jet reconstruction and calibration has been performed. The official CMS jet calibration is not optimal for analyses dealing with a large number of low-p_t jets. This calibration only corrects for detector effects, which means that the same jet algorithm is used on generator particles and reconstructed calorimeter towers, and the difference is corrected for. Effects like particles that are outside of the cone radius and particles that are not measured in the calorimeter like muons and neutrinos are ignored. In addition, light flavour jets are not differentiated from b-jets. A dedicated $\mathrm{t\bar{t}H}$ jet calibration has been developed for the all-hadron channel and published in Reference [90]. This advanced calibration corrects the jets using the generated primary partons as reference. The resulting calibration curve is shown in Figure 5.38.

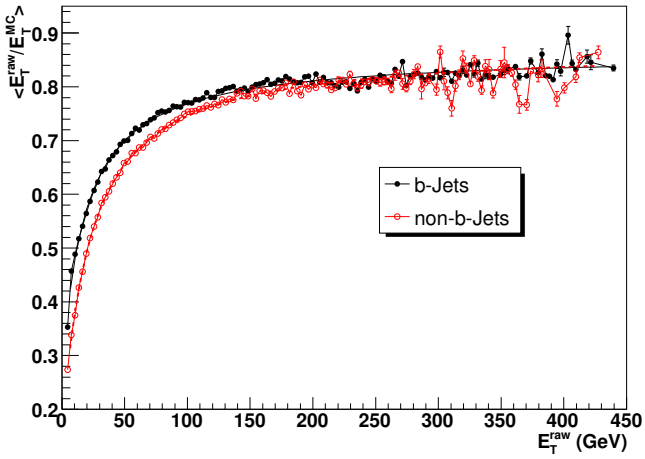

Figure 5.38: Dedicated jet calibration curve for the all-hadron channel. The plot shows the ratio between reconstructed and generated transverse energy E_T^{raw}/E_T^{MC} in dependence on the reconstructed transverse energy E_T^{raw} for the all-hadron $\mathrm{t\bar{t}H}$ signal sample. The primary generator partons are taken as reference. [90]

The determination of the optimal configuration of the jet reconstruction algorithm has been carried out by means of a simple prototype analysis which calculates the purity and significance based on the selection efficiency of the $\mathrm{t\bar{t}H}$ signal and some of the most dominant background events ($\mathrm{t\bar{t}2j}$, $\mathrm{t\bar{t}b\bar{b}}$ and QCD with $\hat{p}_t > 170$ GeV/c). Figure 5.39 shows the result of this study for the Iterative Cone algorithm and suggests a choice of a cone radius of 0.4.

For the task of jet pairing, a χ^2 method, using the invariant masses of top quarks and W bosons as baseline, has been applied in case of the all-hadron channel. The following χ^2 variable is calculated for each possible jet combination:

$$\chi^2_{mass} = \left(\frac{m_{W^+} - m_{jj}}{\sigma(m_W)}\right)^2 + \left(\frac{m_{W^-} - m_{jj}}{\sigma(m_W)}\right)^2 + \left(\frac{m_t - m_{jjj}}{\sigma(m_t)}\right)^2 + \left(\frac{m_{\bar{t}} - m_{jjj}}{\sigma(m_t)}\right)^2 . \quad (5.10)$$

The expected mass values and their σ values are obtained by a parton-jet matching as for the semileptonic channel. The jet combination that yields the minimal χ^2 value is then chosen for the following event selection.

To optimize the signal selection with respect to background rejection, a number of kinematical variables, including the b-tagging discriminator have been studied. These are the following:

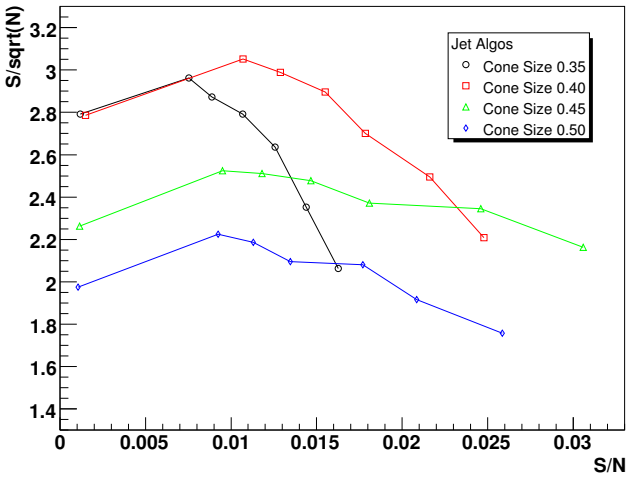

Figure 5.39: Results of the simple prototype analysis. The markers of the same type indicate the same jet finding cone radius as shown in the legend. The cut on the b-tagging discriminator is varied along the lines of same colours. The horizontal axis represents the purity S/N of the event selection, while the vertical axis corresponds to the significance $S/\sqrt{N}$. [2]

- Transverse energies of the jets
- "Combined" b-tagging discriminator
- Event centrality, defined as $\sum_{i=0}^{8} E_T^i/E^i$
- Higgs centrality, defined as above, but only for the two jets assigned to the Higgs boson

The cuts on these variables have been varied simultaneously, thereby mapping out a large phase space of possibilities. As an example, Figures 5.40 to 5.43 show, how the significance $S/\sqrt{B}$ and purity S/B change upon varying one cut while keeping the other cuts fixed.

For the final evaluation of the results in the all-hadron channel, two different sets of cuts have been applied, a "loose" and a "tight" working point, which differ mostly in the choice of the b-tagging discriminator cut, since this has the largest influence on the suppression of light flavour backgrounds. The results and the applied cuts are summarized in Table 5.18.

Even though the "loose" working point gives a better result in terms of significance $S/\sqrt{B}$, the "tight" working point might be a better choice, once systematic errors are included, since the purity S/B has to be optimized in this case. This is discussed in more detail in Section 5.8.

5.7.2 The Di-Lepton Channel

The Di-lepton channel is characterized by its two leptons that are selected according to the lepton selection choices of Section 5.4.1. The Di-lepton channel uses the lepton selection cuts that are used in the semileptonic analyses as vetoes against the double lepton decay. In this way, the sample of events for the di-lepton $t\bar{t}H$ analysis is by construction strictly complementary to those used in the semi-leptonic channels. Furthermore, the Di-lepton channel is accompanied by a significant amount of missing transverse energy because both W bosons decay leptonically and two neutrinos are present. Currently, the di-lepton analysis is

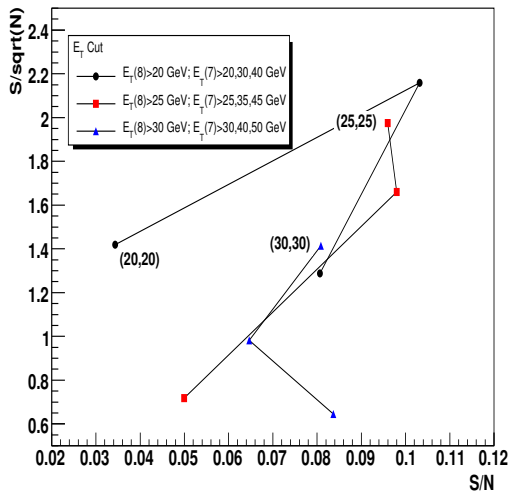

Figure 5.40: E_T cuts on the 7^{th} and 8^{th} jets. The markers of the same type are displaying the variation of the 7^{th} E_T cut while the 8^{th} E_T cut is kept constant as indicated in the legend. [2]

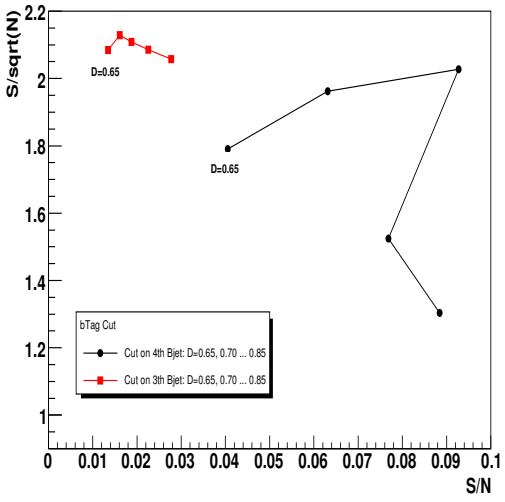

Figure 5.41: Variation of the cut on the "combined" b-tagging discriminator for the 3^{rd} or 4^{th} jets, respectively, ordered by the b-tagging discriminator value. [2]

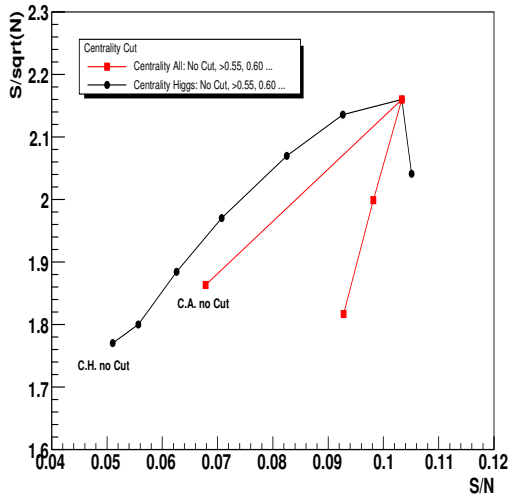

Figure 5.42: Variation of the cuts on Higgs centrality or total event centrality, respectively. [2]

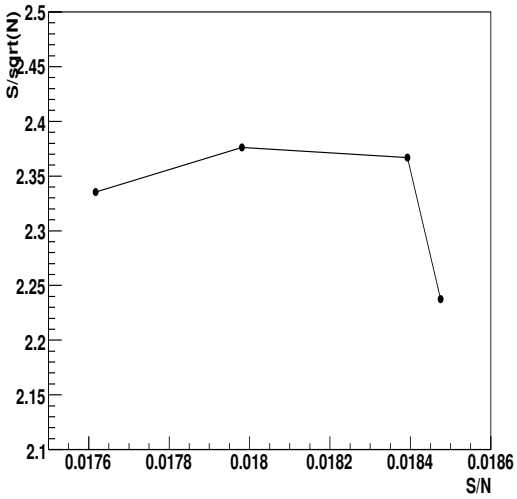

Figure 5.43: Variation of the η cut on all jets, in steps of 0.2, ranging from 2.4 to 3. [2]

a counting experiment, thus no effort has been made to assign the missing transverse energy E_T to the two neutrinos from the hard event.

The details of the di-lepton $t\bar{t}H$ selection are summarized below:

- 2 oppositely charged leptons (e,μ) passing id criteria ($-Log(L_\mu) < 1.4$ for muons, $-Log(L_e) < 1.2$ for electrons)
- corrected $E_T > 40$ GeV
- 4 to 7 jets with calibrated $E_T > 20$ GeV and $|\eta| < 2.5$
- ≥ 3 selected jets b-tagged with discriminator $D > 0.7$

The above is termed the "loose"working point because there are indications that it is possible to increase the purity S/B of the selection by way of more stringent ("tight") criteria in which the maximum of 7 jets is reduced to 6 and the minimum of 3 b-jets is increased to 4. Although the naive significance $S/\sqrt{B}$ decreases, the cleaner selection is plagued less by systematic uncertainties which dominate the more realistic significance $S/\sqrt{B + dB^2}$. However, the numbers quoted for the "tight"working point are currently insufficiently precise because

Table 5.18: Analyzed events, selection efficiency, number of expected events and signal significance in 60 fb^{-1} for the all-hadron $t\bar{t}H$ channel for the two working points ϵ_{loose} and ϵ_{tight}. The signal datasets are labeled by the generated Higgs mass in GeV/c^2 (parentheses). Also quoted are binomial errors arising from the finite sizes of processed datasets. All numbers refer to the full mass range. The applied cuts are $E_T^{8th} > 20$ GeV, $E_T^{7th} > 30$ GeV, χ^2 for W and top within 3σ of their expected values, Higgs centrality > 0.55, $D_3 > 0.8$ for the 3$^{\text{rd}}$ b-tagging discriminator at the "loose" working point. At the "tight" working point, the following additional cuts are employed: $D_3 > 0.85$, $D_4 > 0.7$ and event centrality > 0.8.

	# Events	$\epsilon_{\text{loose}}(\%)$	N^{ev}_{loose} 60fb^{-1}	$\epsilon_{\text{tight}}(\%)$	N^{ev}_{tight} 60fb^{-1}
$t\bar{t}H$ (115)	49636	2.32 ± 0.07	347 ± 10	0.294 ± 0.015	44 ± 4
$t\bar{t}H$ (120)	163494	2.55 ± 0.03	314 ± 5	0.366 ± 0.024	45 ± 2
$t\bar{t}H$ (130)	43254	2.80 ± 0.08	214 ± 6	0.358 ± 0.029	27 ± 2
$t\bar{t}bb$	203135	0.702 ± 0.019	1190 ± 31	0.0645 ± 0.0056	109 ± 9
$t\bar{t}1j$	1031551	0.0084 ± 0.0009	860 ± 92	0.0005 ± 0.0002	49 ± 22
$t\bar{t}2j$	559111	0.0333 ± 0.0024	2000 ± 150	0.0009 ± 0.0004	54 ± 24
$t\bar{t}3j$	68015	0.079 ± 0.011	1910 ± 260	0.0015 ± 0.0015	35 ± 35
$t\bar{t}4j$	97334	0.182 ± 0.014	6660 ± 500	0.0021 ± 0.0015	75 ± 53
$Zt\bar{t}$	80226	0.358 ± 0.021	121 ± 7	0.0312 ± 0.0062	11 ± 2
qcd170	264310	0.0238 ± 0.0030	4810 ± 610	0.0004 ± 0.0004	76 ± 76
qcd120	55128	0.0018 ± 0.0018	83 ± 83	0 ± 0	<95(68%$C.L.$)
Total Backgr.			17600		< 505
$S/\sqrt{B}$ (115)			2.6		2.0
S/B (115)			2.0%		8.7%
$S/\sqrt{B}$ (120)			2.4		2.0
S/B (120)			1.8%		8.9%
$S/\sqrt{B}$ (130)			1.6		1.2
S/B (130)			1.2%		5.4%

of limited dataset sizes at the time of writing. One should therefore not neglect the errors accompanying them.

The selection efficiencies for the two working points, with the corresponding number of expected events and the signal significance, are reported in Tables 5.19. The number of expected events is computed for an integrated luminosity of 60 fb^{-1}.

5.8 Systematic Errors

In the following Section, the systematic uncertainties according to the present knowledge of the expected performance of the CMS detector will be evaluated. The following sources for uncertainties are taken into account:

- Jet energy scale (JES)
- Jet energy resolution
- b-jet and c-jet (mis-)tagging efficiencies
- Light flavour mistagging efficiencies
- Luminosity

Table 5.19: Selection efficiency $\epsilon_{\mathrm{loose}}$ (including branching fraction where applicable) and resulting number of expected events N_{loose} in 60 fb^{-1}, for the di-lepton $\mathrm{t\bar{t}H}$ channel. For a glimpse on possible improvements, the same is provided for a tighter set of cuts ($\epsilon_{\mathrm{tight}}$, N_{tight}). The signal datasets are labeled by the generated Higgs mass in GeV/c^2 (parentheses). Also quoted are binomial errors arising from the finite sizes of processed datasets. All numbers refer to the full mass range.

	# Events	$\epsilon_{\mathrm{loose}}(\%)$	N^{ev}_{loose}	$\epsilon_{\mathrm{tight}}(\%)$	N^{ev}_{tight}
$\mathrm{t\bar{t}H}$ (115)	27900	0.511 ± 0.025	168 ± 8	0.088 ± 0.010	29 ± 3
$\mathrm{t\bar{t}H}$ (120)	26141	0.490 ± 0.025	132 ± 7	0.070 ± 0.009	19 ± 3
$\mathrm{t\bar{t}H}$ (130)	25911	0.490 ± 0.025	82 ± 4	0.072 ± 0.010	12 ± 2
$\mathrm{t\bar{t}b\bar{b}}$	313894	0.637 ± 0.014	1080 ± 24	0.094 ± 0.007	159 ± 12
$\mathrm{t\bar{t}}$1j	280385	0.0125 ± 0.0021	1270 ± 220	0	< 42 (68% C.L.)
$\mathrm{t\bar{t}}$2j	276917	0.0448 ± 0.0040	2690 ± 240	0.00144 ± 0.00072	87 ± 43
$\mathrm{t\bar{t}}$3j	90367	0.0553 ± 0.0078	1330 ± 190	0	< 31 (68% C.L.)
$\mathrm{t\bar{t}}$4j	120042	0.0716 ± 0.0077	2620 ± 280	0.0025 ± 0.0014	92 ± 53
$\mathrm{t\bar{t}Z}$	110156	0.304 ± 0.017	103 ± 6	0.0363 ± 0.0057	12 ± 2
all backgr.			9090		< 422
$S/\sqrt{B}$ (115)			1.8		1.4
S/B (115)			1.8 (%)		6.9 (%)
$S/\sqrt{B}$ (120)			1.4		0.9
S/B (120)			1.5 (%)		4.5 (%)
$S/\sqrt{B}$ (130)			0.9		0.6
S/B (130)			0.9 (%)		2.9 (%)

For the treatment of the jet energy scale and resolution, the procedure follows the commonly agreed CMS prescriptions [85]. The uncertainty due to the JES is implemented by shifting the jet energies systematically up or down by a relative percentage. For jets having a transverse momentum $p_t > 50$ GeV/c, the uncertainty is expected to be 3%, because calibration procedures like the hadronic W boson mass in $t\bar{t}$ events [91] are working well at this energy. In the low p_t region down to 20 GeV/c, where the W boson mass calibration is not available, the energy scale will be set by the GammaJet calibration [92] leading to a linear increase of the uncertainty from 3% to 10%. Below 20 GeV/c, only single particle calibration methods are possible with an accuracy of 10%. This leads to the following functional form of the JES uncertainty:

$$\sigma_E^{jet}/E = \begin{cases} 10\% & p_t < 20 \text{ GeV}/c \\ 10\% - 7\% \cdot (p_t - 20 \text{ GeV}/c)/30 \text{ GeV}/c & 20 \text{ GeV}/c < p_t < 50 \text{ GeV}/c \\ 3\% & p_t > 50 \text{ GeV}/c \end{cases} \quad (5.11)$$

The jet resolution itself is smeared by an overall 10%, which means that the jet fourvector is multiplied by a random number drawn from a gaussian distribution of a mean value of 0 and width 0.1 according to:

$$\vec{p}^{jet} \rightarrow \vec{p}^{jet} \cdot \text{Gauss}(1, 0.1). \quad (5.12)$$

For the b-tagging systematic, the following relative uncertainties in the tagging efficiencies of jets of the various flavours have been assumed:

- 4% for b- and c-jets
- 10% for u, d, s and gluon jets, where "gluon" is defined according to the "algorithmic" definition, in which gluons have the same mistagging rate as u, d and s-jets.

b- and c-jets are treated identically, since they both have real secondary vertices and any systematic effect should be fully correlated between c- and b-jets. Light flavour jets have a higher systematic uncertainty because experience from other hadron collider experiments shows that the tagging rate for these jets is difficult to estimate. Even small deviations in the traversed material budget, and therefore the amount of multiple scattering have large impact on the misidentification rate of light flavour jets.

In [2], the b-tagging uncertainties have been taken into account by simply untagging 4% of the b-jets or –for the variation upwards– tagging a corresponding fraction of untagged b-jets. However, the present analysis applies a more complex likelihood method by tagging four b-jets simultaneously so that there is no simple discriminator cut for each jet that can be passed or not. Thus, for the following study, a different approach is utilized: first, the necessary discriminator cut to obtain the tagging efficiencies used in the analysis (see Table 5.15) is determined according to Figure 5.44. From this Figure, it is visible that an absolute shift in the b-tagging efficiency of 1% corresponds to a shift of the discriminator cut of 0.01. For the c-mistagging rate, an absolute shift of 1% corresponds to a discriminator shift of 0.015, while an absolute shift of 0.1% of light flavour mistagging rate corresponds to a discriminator shift of 0.0103.

Therefore, a relative b-tagging uncertainty of 4% at a working point of 55% corresponds to an absolute shift in b-efficiency of 2.2% or a shift in the discriminator cut of 0.022. As already mentioned, shifting the discriminator cut is not possible, because the analysis does not apply such a cut. Instead, the b-tagging discriminator itself is shifted by the corresponding value.

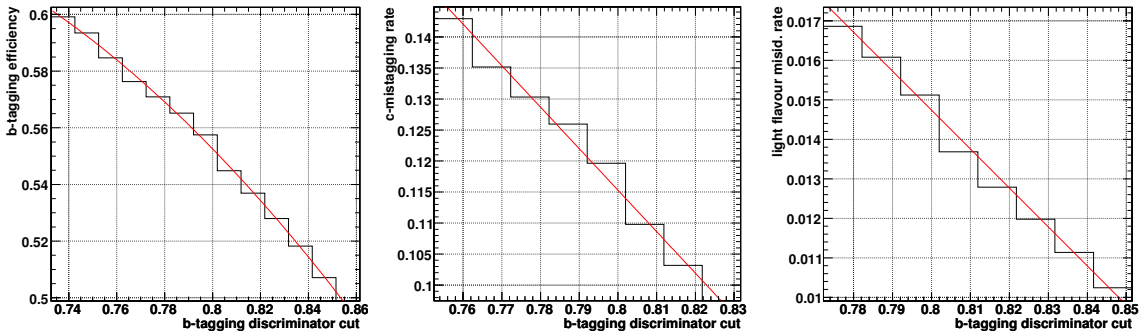

Figure 5.44: Tagging efficiencies and mistagging rates in dependence on the b-tagging discriminator cut. The histograms are zoomed to better identify the behaviour at the working point used in the analysis. From the left to the right: b-, c- and light flavour efficiencies.

This modification can be applied at the very beginning of the analysis and is therefore easy to implement.

The estimation of these uncertainties is accomplished by simultaneously varying the discriminator for b- and c-jets according to the given percentages. The variation for light flavour jets is done independently. In both cases, the variation is performed "upwards" and "downwards", only the direction which gives the larger change in event yields is quoted in Table 5.20.

The systematic uncertainty due to the luminosity affects signal and background in the same way and cancels out completely in the purity S/B. For the significance $S/\sqrt{B}$, this is only a higher order effect which can cause a change of $1.03/\sqrt{1.03} = 1.014 = 1.4\%$ and can therefore be neglected.

Table 5.20 shows the effect of the various systematic uncertainties in terms of a relative change (in %) of the final selection efficiencies. Also the change in final event numbers is given at the two workingpoints ϵ_{loose} and ϵ_{tight}. The relative uncertainties are calculated at the "loose" working points and it is assumed that the same uncertainties apply at the "tight" working point. This is justified by the fact that only the choice of the b-tagging working point is different at the tight working point and that the mistagging efficiencies in dependence on the b-discriminator cut have an approximately linear behaviour. The propagation of the errors to the tight working point is necessary because of the small statistical significance of some of these calculations. For instance, the number of remaining events after all selection cuts at the loose working point in the $t\bar{t}$4j sample is 4 which leads to a relative statistical error of $\sqrt{4}/4 = 50\%$. Obviously, the obtained numbers for this specific sample cannot be considered to be very meaningful. The statistical errors due to the finite sizes of data samples are given in Table 5.9 for all samples in order to be able to judge the reliability of the obtained numbers. Fortunately, reliable numbers are available for all signal samples, for the $t\bar{t}$1j, $t\bar{t}$2j and $t\bar{t}b\bar{b}$ and $t\bar{t}$Z samples. From these samples, conservative estimations for $t\bar{t}$4j are possible as indicated in Table 5.20.

The numbers obtained hereby are compatible with the results in [2] considering the statistical bounds of these calculations. One can conclude that the resulting systematic uncertainty in terms of background event yields is 34% at the loose working point and 31% at the tight working point.

Table 5.20: Systematic uncertainties relative to final selection efficiencies (in percent) for the semi-leptonic $t\bar{t}H$ channels. Σ is the quadrature sum of all changes in the given row. The last two colums show the abolute uncertainty (in number of events) at the two different working points ϵ_{loose} and ϵ_{tight}. The $t\bar{t}4j$ line is given in brackets because this particular background does not give reliable results since the systematical variation is based on a number of only four remaining events. A conservative upper limit of 40% for $t\bar{t}4j$ is estimated from the other backgrounds and used in the following.

	JES (%)	Jet res. (%)	bc-tagging (%)	uds-tagging (%)	Σ (%)	# events ϵ_{loose}	# events ϵ_{tight}
$t\bar{t}H$ (115)	5.4	4.4	23.8	0.2	24.8	36	12
$t\bar{t}H$ (120)	3.4	1.6	21.5	0.07	21.9	26	9
$t\bar{t}H$ (130)	3.3	1.1	23.1	0.3	23.3	19	6
$t\bar{t}1j$	23.7	8.5	25.4	0	35.8	166	17
$t\bar{t}2j$	4	5.4	37.8	2.7	38.5	207	25
$t\bar{t}3j$	26.7	6.7	26.7	0	38	127	25
($t\bar{t}4j$)	(175)	(100)	(50)	(0)	(207)	(266)	(0)
$t\bar{t}4j$					≈ 40	≈ 50	(0)
$t\bar{t}b\bar{b}$	6.4	1.4	25.3	0.12	26.2	192	62
$t\bar{t}Z$	6.1	2	28.3	1.01	29	10	3
total Bg.						753 (34%)	133 (31%)

An interesting observation is the fact that the impact due to the 10% uncertainty of light flavour mistagging rate is mostly below 3%. This confirms the observation, that the largest part of the misidentified $t\bar{t}N$j events consists of events with splitting of gluons into real b-jets or W boson decays into charm jets.

The impact of these systematic uncertainties on the final significance is given in Table 5.21. Under the assumption that these systematic errors follow a Gaussian distribution, the error on the number of background events dB has to be included quadratically and it follows for the significance σ:

$$\sigma = \frac{S}{\sqrt{B + \mathrm{d}B^2}}. \tag{5.13}$$

Another important aspect to be investigated is the question of how precise the background has to be known in order to reach a specific significance. The limit of $S/\sqrt{B} = 3.1$ for a Higgs boson mass of $m_H = 120$ GeV/c^2 which is given by the semileptonic analysis can be increased to 3.9 if the two other channels, the all-hadron channel and the di-lepton channel, are combined together with the semileptonic channel. This is possible by simply adding the event yields for signal and background, because the channels are disjoint due to the lepton selection. Figure 5.45 shows the behaviour of the combined significance in dependence on the background uncertainty for the semileptonic channel and for the combination of all channels.

It is visible that the combined significance reaches higher values at 0, but due to the low purity of the all-hadron and di-lepton channels, the significance drops quickly as soon as the uncertainty increases. Figure 5.45 also shows the uncertainty due to the theoretical knowledge of the background cross section which is varied by 20% up or down in the plot.

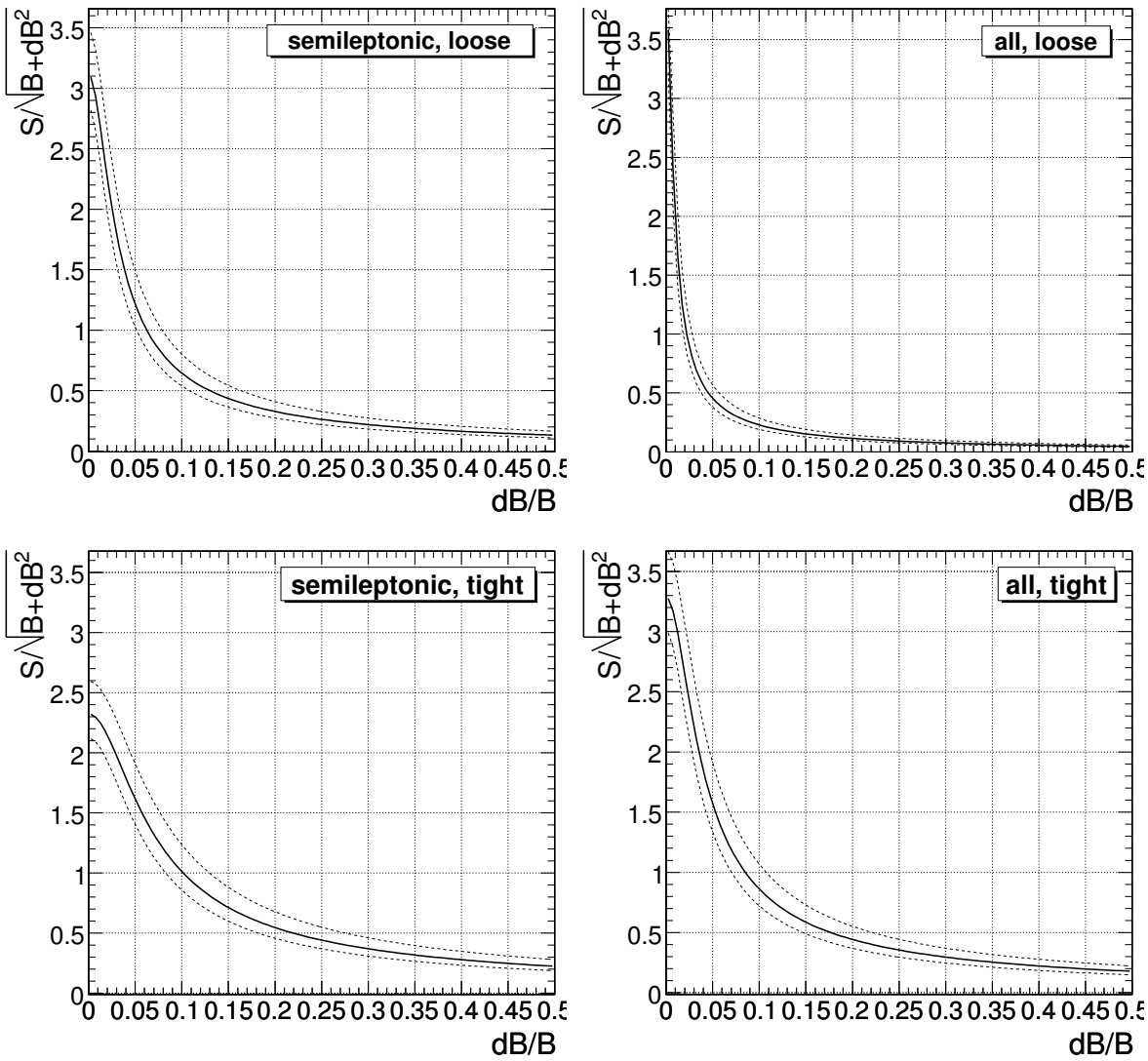

Figure 5.45: Significance $S/\sqrt{B + \mathrm{d}B^2}$ in dependence on the fractional uncertainty $\mathrm{d}B/B$ of the background at the "loose" and "tight" working points in the Gaussian error model for a Higgs boson mass of $m_H = 115$ GeV/c^2 and an integrated luminosity of 60 fb^{-1}. The dashed line corresponds to a variation of the background cross section of 20% due to the theoretical uncertainty. Two plots on the left: Semileptonic channel only. Plots on the right: All channels combined. The upper row shows the loose working points, while the two plots on the bottom show the tight working point.

The tight working point shows better results compared to the loose working point as soon as the background uncertainty reaches realistic values above 5%.

This kind of error model has its justification in the sense that it reflects the uncertainty on the measurements from the current point of view. As soon as CMS starts to take data, control samples will be available to help reducing the systematic uncertainties. These uncertainties are certainly smaller than the uncertainty on the current performance estimations. As an alternative model, the errors stated here are considered to reflect the upper limit to the expected uncertainties. They can be taken into account into the final significance by assuming a "rectangular" model for the distribution of the error. This is done analytically by convoluting the assumed gaussian distribution of the statistical error with the rectangular distribution of the systematic error [93]. The distribution of the number of events within this error model is shown in comparison to the gaussian error model in Figure 5.46. Figure 5.47

Table 5.21: Significance of the semileptonic channels before and after taking into account the uncertainty dB in the total number of background events due to systematics. The result is shown for the two working points ϵ_{loose} and ϵ_{tight}, assuming the same systematic uncertainties for both.

	S/B	$S/\sqrt{B}$	$S/\sqrt{B+\mathrm{d}B^2}$
$\epsilon_{\mathrm{loose}}$			
$t\bar{t}H$ (m_H=115 GeV/c^2)	0.07	3.1	0.20
$t\bar{t}H$ (m_H=120 GeV/c^2)	0.053	2.5	0.16
$t\bar{t}H$ (m_H=130 GeV/c^2)	0.036	1.7	0.11
$\epsilon_{\mathrm{tight}}$			
$t\bar{t}H$ (m_H=115 GeV/c^2)	0.11	2.3	0.35
$t\bar{t}H$ (m_H=120 GeV/c^2)	0.09	1.9	0.29
$t\bar{t}H$ (m_H=130 GeV/c^2)	0.06	1.2	0.19

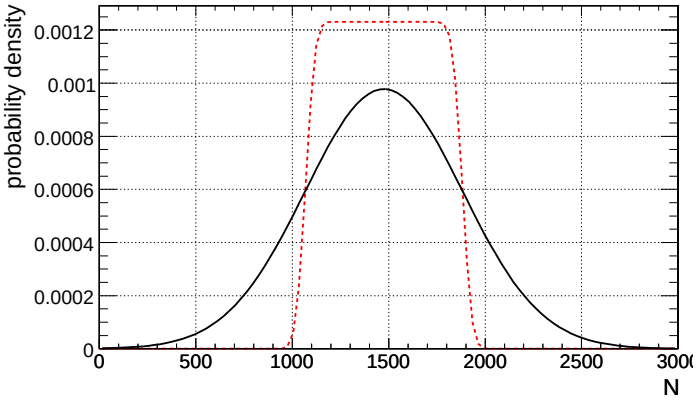

Figure 5.46: Probability densities of the expected number of events at the "tight" working point including all channels. The solid line shows the gaussian error model while the dashed line shows the rectangular error model. In both cases, the statistical (gaussian) error is convoluted with the systematical error.

shows the resulting significance in this error model in comparison with the gaussian error model. As expected, the significance decreases more quickly in the gaussian model than in the rectangular model for low uncertainty values. Obviously, not only the size of the error, but also the assumed model of the error distribution determines the final significance.

The conclusion from these studies is that the uncertainty of the knowledge of the background level has to be much less than 10% before a measurement is possible in this channel. This is an enormous challenge, but it will be possible by employing methods to measure the background directly from data. This can be done with a high statistical precision because of the abundance of $\mathrm{t\bar{t}}$ events at the LHC.

5.8.1 Prospects for Improvements

A number of possibilities to improve the results remain to be implemented and tested. Most of them have to rely on further developments of the performance of the CMS detector and the analysis tools as a whole. For example, the poor jet reconstruction performance needs to be improved urgently. For this purpose, a promising "energy flow" project has been launched within CMS. The aim is to integrate all detector components to the jet finding, not only the calorimeters as done at present. Muons and tracks carry important information that can be used to improve the performance as has been done in previous particle physics experiments.

In a mature experiment, more complex triggers should be implemented. As discussed in Section 5.3.1, a dedicated $\mathrm{t\bar{t}H}$ trigger has not been available for this study, but is expected to improve the signal selection efficiency by some percent. Even the single lepton trigger performances, which are around 60% as given in Table 5.8, are not statisfying. The inefficiencies are mostly related to the p_t thresholds in the lepton triggers. The efficiency therefore depends on the p_t spectrum. Hence, the thresholds should be lowered or the lepton triggeres should be combined with alternative trigger paths. Experiences from other experiments motivate efficiencies around 90% or better.

Furthermore, the exploitation of differences between signal and background in terms of

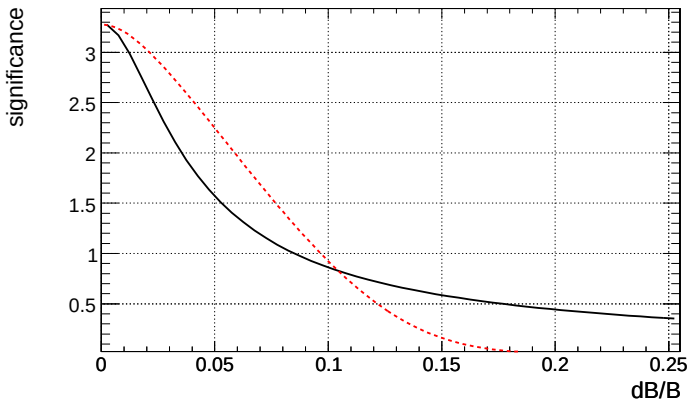

Figure 5.47: Significance in dependence on the fractional uncertainty $\mathrm{d}B/B$ of the background at the "tight" working point for all channels and a Higgs boson mass of $m_H = 115$ GeV/c^2 in 60 fb^{-1}. The solid line shows the gaussian error model, while the dashed line shows the rectangular error model. In both cases, the statistical (gaussian) error is convoluted with the systematical error.

kinematical variables can be used to extract a clearer signal. Some discriminating variables have been identified by applying a neural network [94]. The suppression power has been found to be of the order of 20% at a signal efficiency of 90%.

Probably the most promising approach to get rid of the huge -especially theoretical- systematical uncertainties is the determination of background rates from real data. For example, the light flavour jet mistagging rate can be obtained from a high purity semileptonic $t\bar{t}$ sample that has been obtained without applying b-tagging, e.g. with a top-mass window. The jets belonging to the W boson provide a well defined sample of light flavour and charm jets that can be used to measure the tagging rates at the corresponding energies.

Chapter 6

Summary and Conclusions

In this thesis, the potential of the CMS experiment to discover the Standard Model Higgs boson in the decay channel $H \rightarrow b\bar{b}$ has been evaluated. This channel has the highest branching ratio in the mass range just above the exclusion limit from the LEP experiments at 114.4 GeV/c^2. Because of the large abundance of other processes with two b-quarks in the final state at the Large Hadron Collider, this search has to be carried out in the mode of associated production with top quarks, which deliver a clearer signature and less backgrounds, thus holding promise for a discovery.

This study of the $t\bar{t}H$ discovery potential has been performed as realistically as possible by applying a full Monte Carlo simulation of the CMS detector and by using trigger and reconstruction algorithms that will be applied also on real data. The full simulation and reconstruction has been done for the first time in this channel, hence pointing out previously unknown limitations.

Since this specific analysis imposes the highest demands on detector performance and physics reconstruction tools, it has been selected as a benchmark for the Physics Technical Design Report (PTDR) [1]. Therefore, a significant effort has been invested in developing, evaluating and optimizing the reconstruction and analysis tools. In particular, the b-flavour tagging algorithms, which are the most powerful instruments applied in the $t\bar{t}H$ analysis because of the presence of four b-jets, have been studied and improved in full as well as in the fast detector simulation of CMS.

For the fast simulation, an interface for the b-tagging algorithms has been implemented, released and maintained in the production versions of the CMS software. The performance has been compared to the full simulation and several approaches to improve the agreement have been adopted. The main result is that the observables at the b-hadron decay vertices have been found to be very well described in the fast simulation. The remaining discrepancies are due to the number of charged tracks at the primary event vertex. This investigation can be considered as an important step towards a satisfying agreement between full and fast detector simulation in CMS.

In case of the b-flavour tagging in the full detector simulation and reconstruction, some major improvements have been introduced to the algorithms. By combining a secondary vertex based b-tagging algorithm with a soft lepton tagging algorithm, an improvement of more than 15% has been reached in terms of light flavour jet rejections. Together with an application of an improved vertex finder that makes use of tracks from tertiary decay vertices, the improvement of the b-tagging performance arrived at more than 25% compared to the

standard algorithms.

Furthermore, the impact of systematic errors due to uncertainties concerning the performance of the detector and reconstruction tools has been studied for the very first time in conjunction with the $t\bar{t}H$ analysis. Effects like the uncertainty of the energy scale of jets as well as b-tagging and mistagging rates have been taken into account. Also the theoretical uncertainty of the knowledge of cross sections of $t\bar{t}$ plus N jets processes due to the unavailability of next-to-leading order calculations has been investigated. The influence on the final event yields and the discovery potential has been estimated from different viewpoints. The results have been found to depend on the type and model of the assumed uncertainties.

A separate topic, but still related to the $t\bar{t}H$ analysis, is the technical realization of the study by applying grid technologies in order to cope with the large amount of required data storage and computing resources associated with this task. A first proof of concept has been accomplished during this thesis in order to analyze the CMS data of $t\bar{t}H$ signal and background event samples using distributed grid resources. The prototyping, deployment, configuration and maintenance of event catalogues and databases at the German Tier 1 center GridKa have been a part of the successful realization of this analysis.

Several conclusions can be drawn from the studies presented in this thesis. First, the results published in [2] and [1] have been confirmed and cross-checked since the analysis has been completely reimplemented and optimized, followed by a thorough verification of the code. In addition, some significant improvements in the optimization of kinematical cuts, b-tagging, event selection and the statistical reliability due to the size of datasets, have been achieved. Compared to the results in [2], the significance without systematic errors could be improved by about 10% in the semileptonic channel, while the purity improved by more than 90%. The purity is the deciding factor as soon as systematic errors are taken into account. Therefore also the result including systematic errors improved, but still stays below the limit of observability if the conservative model of Gaussian errors is assumed. In a different, more optimistic error model, the situation looks better and the significance can reach values of up to three for $m_H = 115$ GeV/c^2 after a period of three years of datataking at a luminosity of $2 \cdot 10^{33}\text{cm}^{-2}\text{s}^{-1}$ corresponding to an integrated luminosity of 60 fb^{-1}. However, the knowledge of the systematic uncertainties has to be improved by methods of measuring the background directly from data.

An important result of the full simulation and reconstruction is the determination of the impact of $t\bar{t}$ plus N light flavour jets processes which act as backgrounds to the $t\bar{t}H$ analysis. It has been found that these backgrounds have been underestimated in previous studies. These specific backgrounds can be suppressed with b-flavour tagging methods but the majority of the misidentified events consist of gluons splitting into b- or c-quarks which cannot be rejected efficiently. Moreover, W bosons decaying into charm quarks contribute a large fraction to the misidentification of this background. Therefore, a calculation of the background rejection efficiency has to take these effects into account.

Obviously, the measurement of the $H \rightarrow b\bar{b}$ decay will be a big challenge. Even though it cannot be considered to be a discovery channel, a measurement will be possible after the determination of the Higgs boson's mass in other decay channels in case of its existence. This way, important consistency checks within the standard model will be facilitated. Especially the measurement of the combined top-Higgs, Higgs-bottom Yukawa coupling, which is only possible in the $t\bar{t}H$ channel, can be performed by determining the cross section of the signal process.

Appendix A

Comparison of b-Tagging Observables for ORCA and FAMOS

In this section, the observables of Figure 4.10 are shown in comparison between ORCA and FAMOS. These observables are all related to the secondary vertex, which means that the successful reconstruction of a secondary vertex is mandatory, i.e. that the jet has to be categorized in the first one ("RecoVertex") of the three vertex categories, which are introduced in Section 4.2.6.

Figure A.1 shows the comparison of the invariant mass of charged particle tracks associated to the secondary vertex for three different jet flavours. Figures A.2 to A.6 show the comparisons for the remaining observables which are:

- Multiplicity of charged particle tracks associated to the secondary vertex (Figure A.2).
- Distance between primary and secondary vertex in the transverse plane, devided by its error, called flight distance significance (Figure A.3).
- Energy of charged particle tracks divided by the total energy of charged particles associated to the jet (Figure A.4).
- Rapidities of charged particle tracks associated to the secondary vertex with respect to the jet direction $y = \frac{1}{2} \ln \left(\frac{E+p_\parallel}{E-p_\parallel} \right)$. This enters for each track associated to the secondary vertex (Figure A.5).
- The track impact parameter significance of the first track exceeding the charm mass threshold in the transverse plane (Figure A.6).

All of these variables show a fair agreement between FAMOS and ORCA except for the energy fraction of charged particles at the secondary vertex. This variable is calulated by simply dividing the energy sum of the charged particle tracks at the secondary vertex by the energy sum of all the charged particle tracks associated to the jet. Apparently, the energy is cumulated at the secondary vertex in the FAMOS case. As Figure A.2 indicates, the number of tracks associated to the secondary vertex is in good agreement and is not responsible for this behaviour. However, the total number of tracks in the jet, which is shown

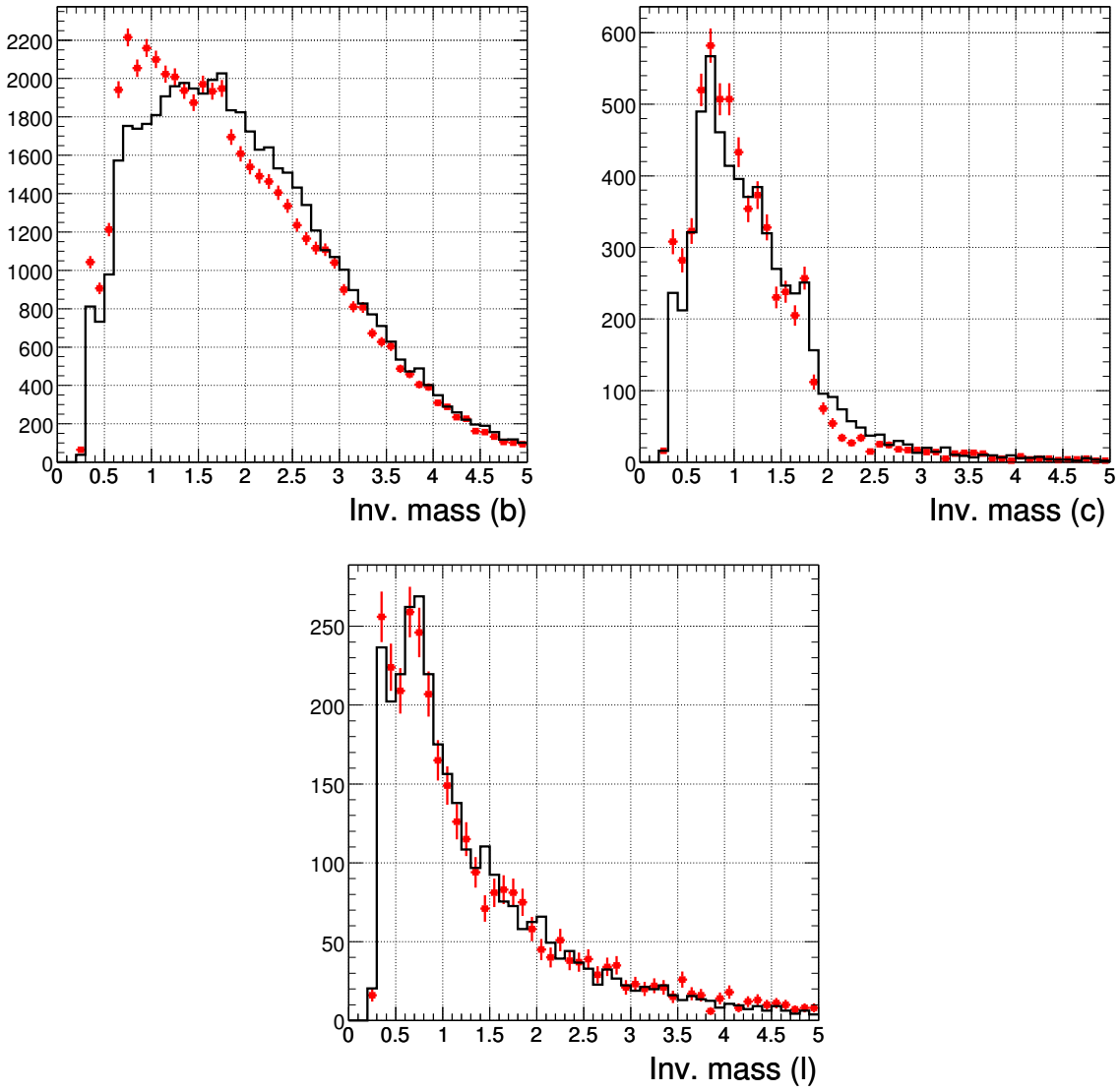

Figure A.1: Invariant mass of charged particles associated to the secondary vertex for ORCA (black solid line) and FAMOS (grey markers). The top left plot shows b-jets only, the top right c-jets and the plot on the bottom displays uds- and gluon-jets.

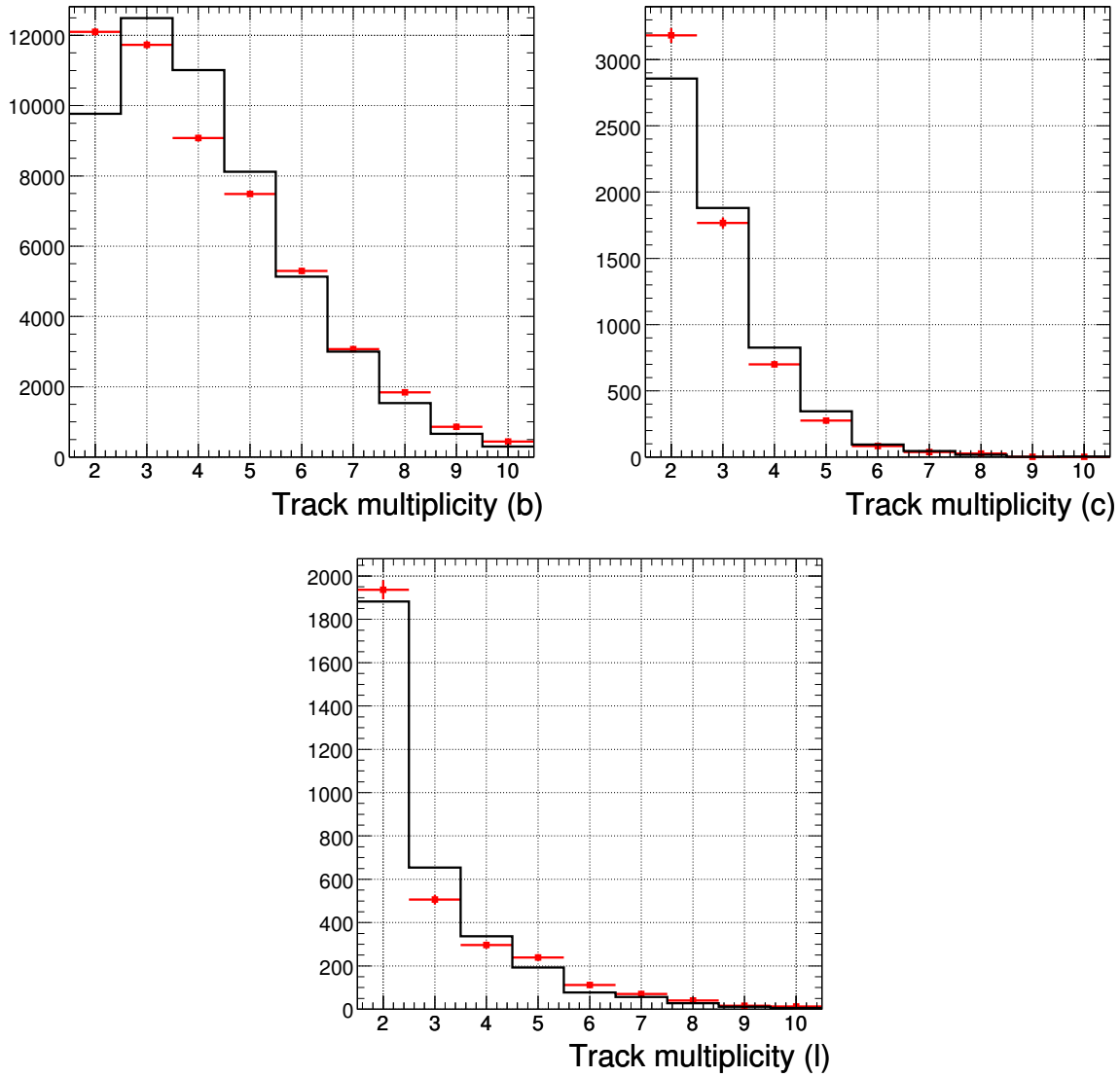

Figure A.2: Multiplicity of charged particle tracks associated to the secondary vertex for ORCA (black solid line) and FAMOS (grey markers). The top left plot shows b-jets only, the top right c-jets and the plot on the bottom displays uds- and gluon-jets.

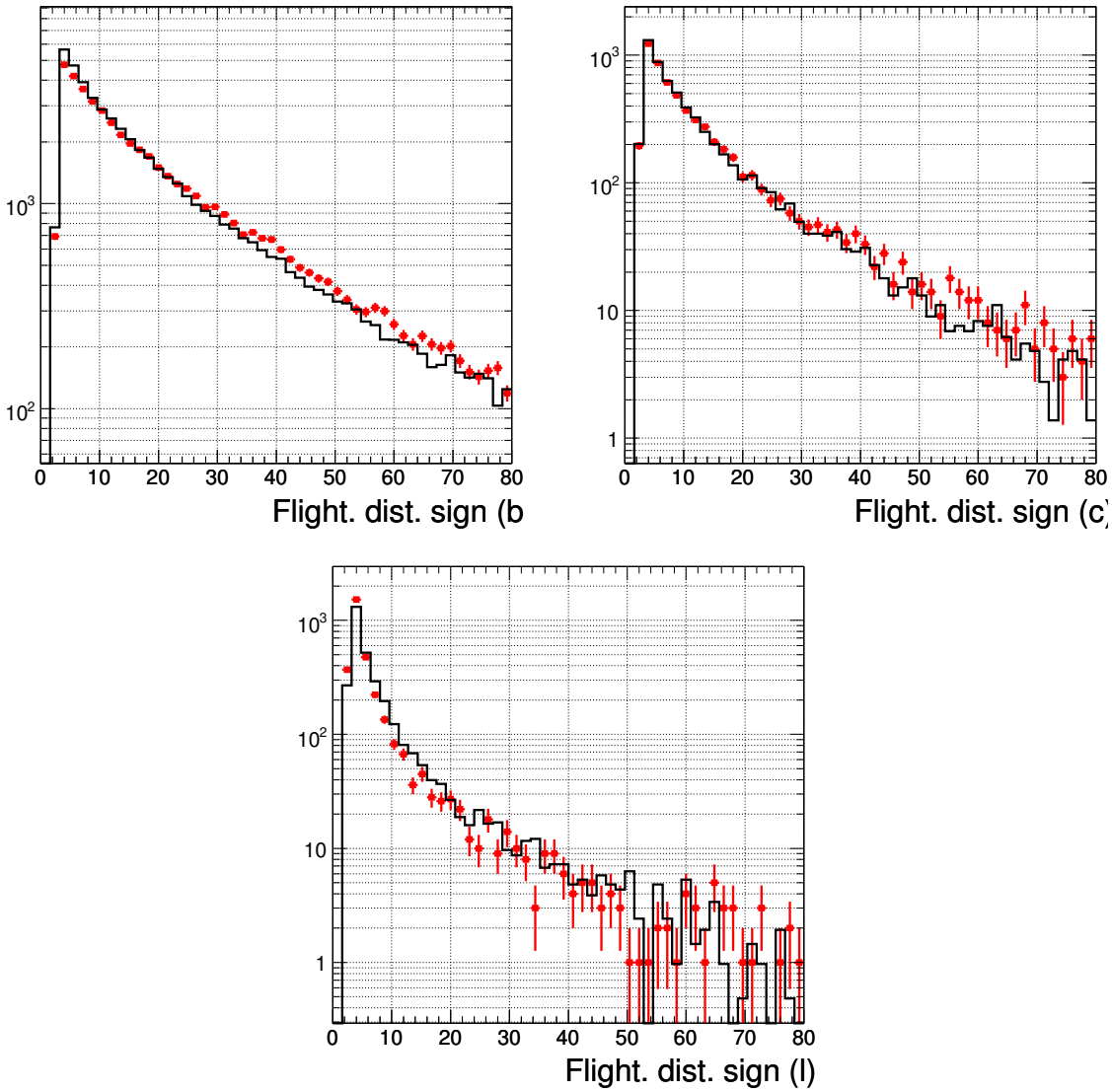

Figure A.3: Flight distance significance at the secondary vertex for ORCA (black solid line) and FAMOS (grey markers). The top left plot shows b-jets only, the top right c-jets and the plot on the bottom displays uds- and gluon-jets.

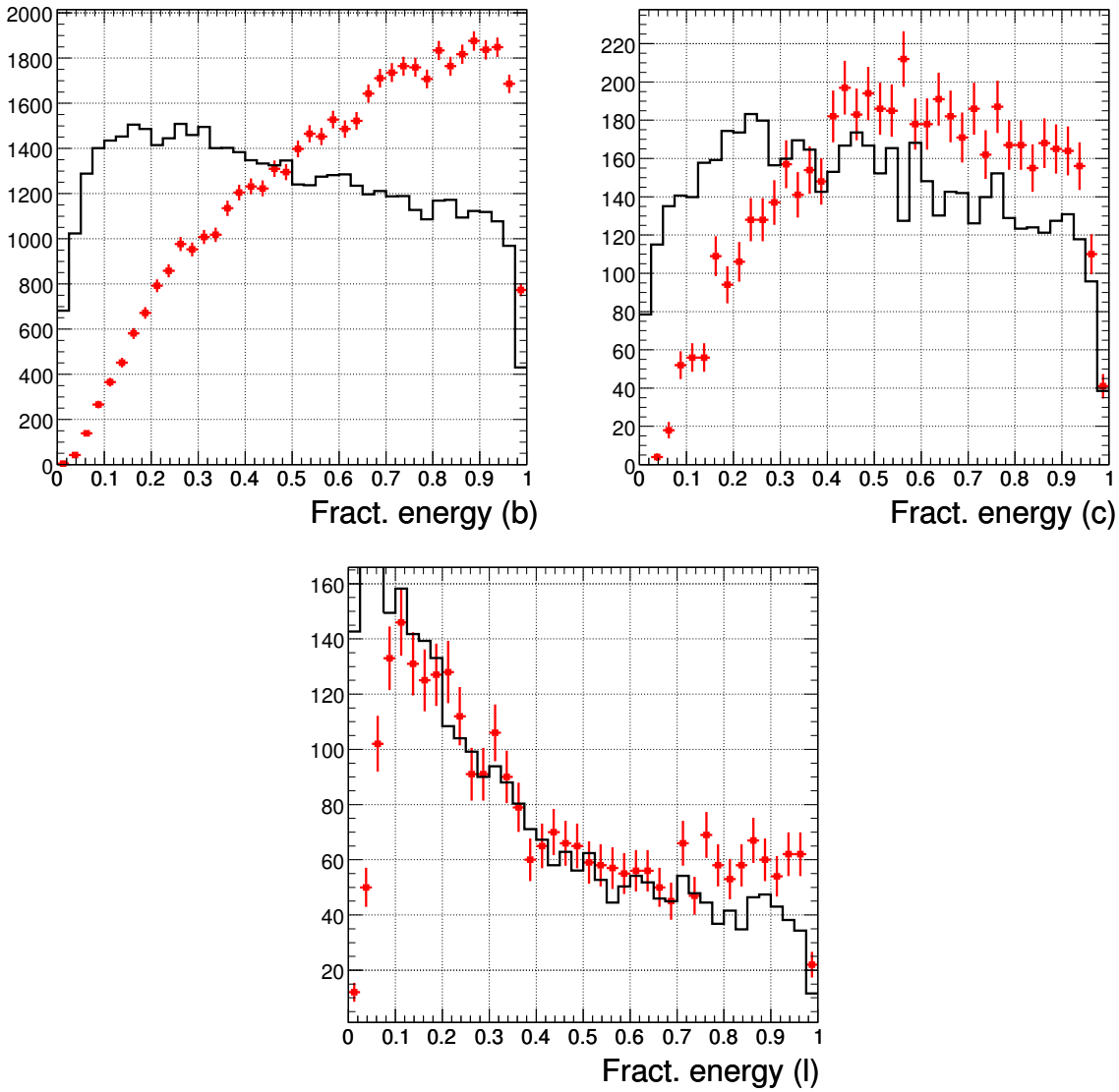

Figure A.4: Energy of charged particle tracks divided by the total energy of charged particles associated to the jet for ORCA (black solid line) and FAMOS (grey markers). The top left plot shows b-jets only, the top right c-jets and the plot on the bottom displays uds- and gluon-jets.

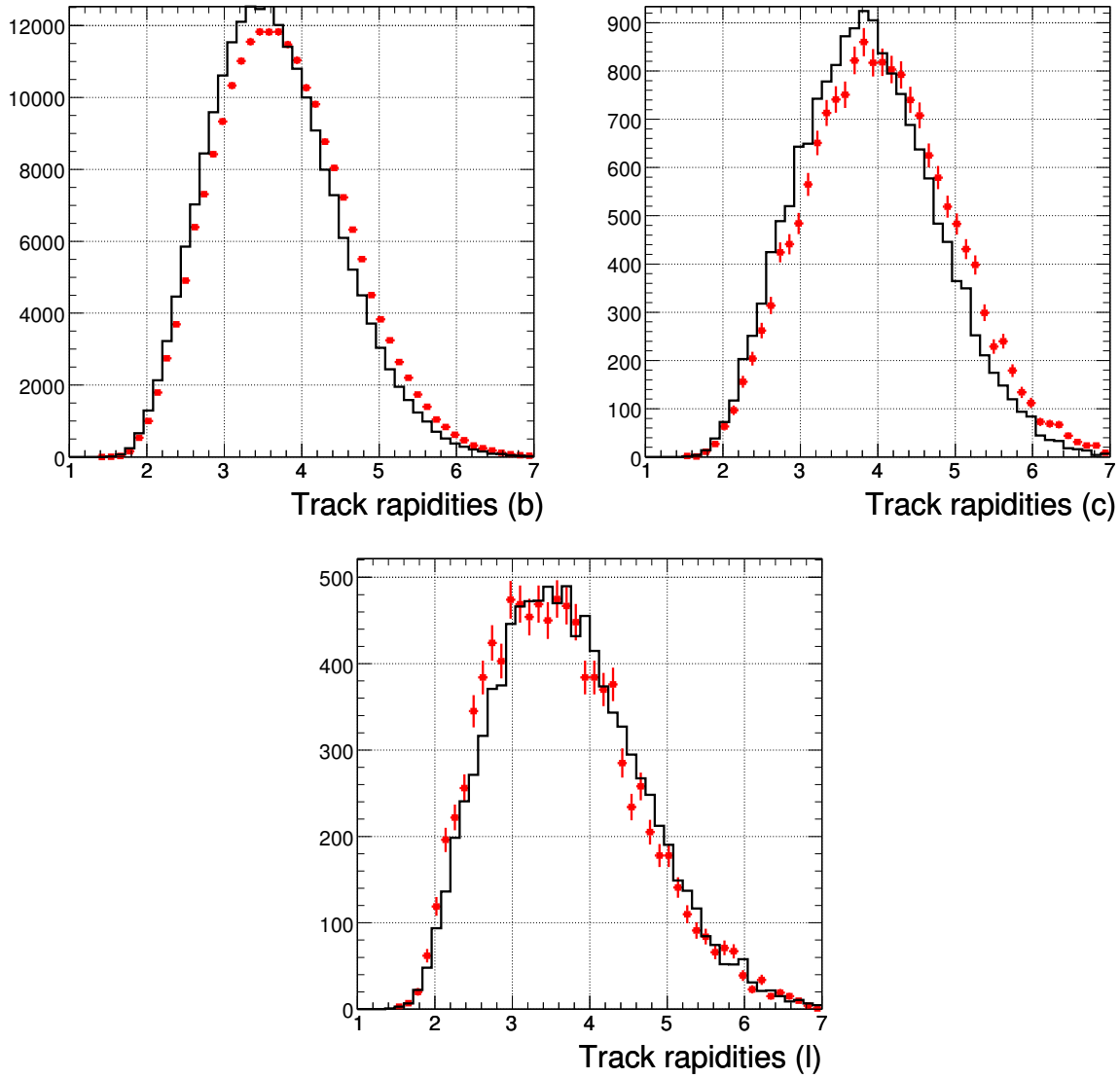

Figure A.5: Rapidities of charged particle tracks associated to the secondary vertex with respect to the jet direction for ORCA (black solid line) and FAMOS (grey markers). The top left plot shows b-jets only, the top right c-jets and the plot on the bottom displays uds- and gluon-jets.

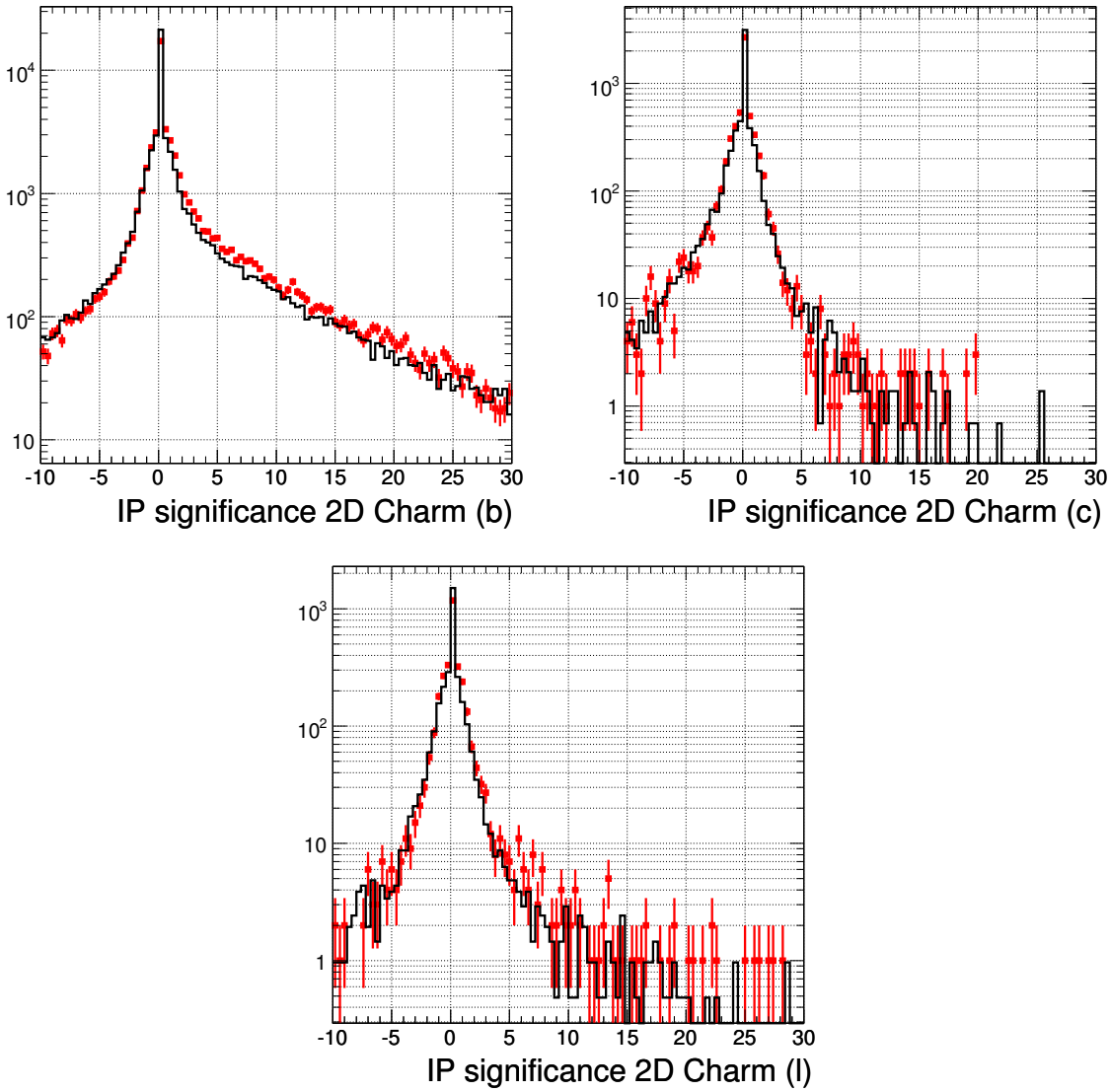

Figure A.6: Signed transverse impact parameter significance of the first track exceeding the charm mass threshold for ORCA (black solid line) and FAMOS (grey markers). The top left plot shows b-jets only, the top right c-jets and the plot on the bottom displays uds- and gluon-jets.

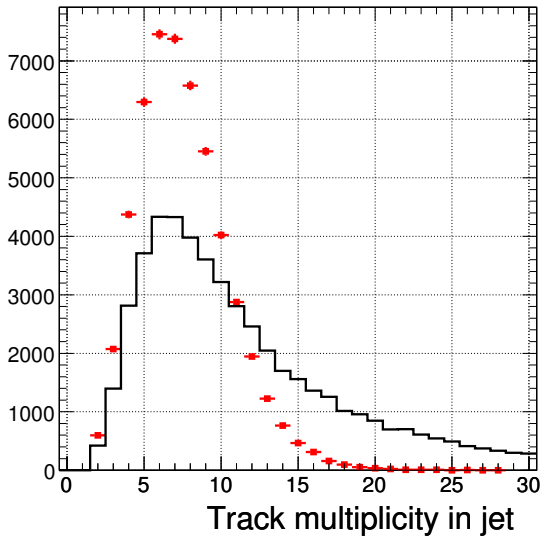

Figure A.7: Total number of charged particle tracks associated to the jet for ORCA (black solid line) and FAMOS (grey markers) in the case of b-jets.

in Figure A.7, gives a different picture. Obviously, the number of tracks at the primary vertex is in very bad agreement between fast and full simulation. This leads to the conclusion, that also the amount of energy associated to the primary vertex is too small which leads to the observed misbehaviour of the energy fraction. The reason for this discrepancy needs further investigation, it might be related to the amount of fake tracks. This is surely also related to the observation that the secondary vertex reconstruction is more efficient in FAMOS and that the distribution of the impact parameter significance for light flavour jets is narrower in FAMOS.

Appendix B

Interpretation of the Generator Output

A sample listing of the generator output in case of a $t\bar{t}2j$ event is given in the following:

---#	ID	st	Mo1	Mo2	Da1	Da2	px	py	pz	E
1	p+	3	0	0	0	0	0.00	0.00	7000.00	7000.00
2	p+	3	0	0	0	0	0.00	0.00	-7000.00	7000.00
3	u	3	1	0	0	0	-0.32	-0.00	1897.69	1897.69
4	s	3	2	0	0	0	1.60	0.19	-664.75	664.75
5	g	3	3	0	0	0	-1.80	-4.81	401.50	401.54
6	s	3	4	0	0	0	0.87	0.22	-566.38	566.38
7	t	3	5	6	0	0	40.38	-65.99	82.05	208.19
8	t~	3	5	6	0	0	-67.11	131.51	181.79	292.35
9	s	3	5	6	0	0	8.99	-23.75	-415.58	416.35
10	g	3	5	6	0	0	16.81	-46.35	-13.14	51.03
11	W+	3	7	0	0	0	-38.77	-58.53	69.96	127.64
12	b	3	7	0	0	0	79.15	-7.46	12.09	80.56
13	W-	3	8	0	0	0	-78.79	19.85	112.24	160.21
14	b~	3	8	0	0	0	11.68	111.65	69.55	132.14
15	c	3	11	0	0	0	-27.98	12.24	39.21	49.72
16	s~	3	11	0	0	0	-10.79	-70.77	30.75	77.91
17	d	3	13	0	0	0	-62.50	16.70	124.01	139.87
18	u~	3	13	0	0	0	-16.29	3.15	-11.77	20.34
19	gamma	1	10	0	0	0	0.74	-4.25	-1.06	4.44
20	W+	2	11	0	72	77	-35.48	-56.44	67.89	124.58
21	W-	2	13	0	22	90	-76.43	19.72	109.16	156.89
22	gamma	1	18	0	0	0	-0.87	0.18	0.21	0.91
23	K0	2	2	0	190	190	-0.90	0.33	-3023.90	3023.90
24	u	2	3	0	191	191	0.60	0.48	364.59	364.59
25	g	2	0	0	191	191	-14.45	-6.98	126.37	127.38
26	g	2	0	0	191	191	-2.47	-1.06	14.83	15.07
27	g	2	14	0	191	191	0.53	35.20	34.57	49.34
28	b~	2	14	0	191	191	9.80	74.62	35.09	83.17
29	s	2	9	0	213	213	1.83	-0.70	-10.65	10.83
30	s~	2	9	0	213	213	2.47	-9.54	-131.42	131.79
31	u	2	10	0	220	220	-0.32	-1.99	-0.06	2.04
32	g	2	10	0	220	220	-0.58	-1.16	0.31	1.34
33	g	2	10	0	220	220	8.62	-11.09	-6.31	15.40

```
  34 g       2    10     0    220    220     1.47    -4.15      -2.20     4.92
  35 g       2    10     0    220    220     3.50    -9.37      -3.19    10.50
  36 g       2    10     0    220    220     3.08   -12.59      -0.97    13.00
  37 g       2     0     0    220    220     2.44    -1.04      -1.08     2.86
  38 g       2     0     0    220    220     2.56    -0.63       2.28     3.49
  39 g       2     0     0    220    220    14.45     6.98      28.50    32.71
  40 g       2     0     0    220    220     0.86     0.36       2.53     2.70
  41 g       2     0     0    220    220     0.83     0.32       5.82     5.88
  42 g       2     0     0    220    220     1.76     1.14      10.59    10.79
  43 g       2     0     0    220    220    -2.44     1.04     113.23   113.26
  44 g       2     0     0    220    220    -1.06    -1.55     138.93   138.95
  45 g       2     0     0    220    220    -0.08    -0.46      26.62    26.63
  46 g       2     3     0    220    220     0.36     0.28       2.78     2.81
  47 g       2     3     0    220    220    -0.77    -0.10       4.54     4.61
  48 g       2     3     0    220    220    -0.63     0.70       4.58     4.67
  49 g       2     3     0    220    220    -0.54     1.82       4.35     4.75
  50 g       2     3     0    220    220     1.04    -0.05      11.70    11.74
  51 g       2     3     0    220    220     1.55     0.26     100.50   100.51
  52 g       2     3     0    220    220     0.99     0.98     136.73   136.73
  53 g       2     3     0    220    220    -1.13     0.43     865.88   865.89
  54 ud_0    2     1     0    220    220     0.32     0.00    2714.26  2714.26
  55 u~      2    10     0    265    265     0.04    -1.07       0.12     1.13
  56 g       2     0     0    265    265    -1.59    -3.31    -102.71   102.77
  57 g       2     9     0    265    265    -0.16     0.41      -0.85     0.96
  58 g       2     9     0    265    265     0.36    -1.24     -31.92    31.94
  59 g       2     9     0    265    265     0.96    -2.64     -58.82    58.89
  60 s       2     9     0    265    265     3.39    -9.68    -175.76   176.06
  61 b       2    12     0    279    279    74.74    -2.75      12.63    76.00
  62 g       2    12     0    279    279     6.33    -3.51       0.48     7.26
  63 g       2     7     0    279    279    -0.86     0.13      -0.52     1.01
  64 g       2     7     0    279    279    -5.00    -2.89      -0.97     5.86
  65 g       2     7     0    279    279     0.03     0.41      -0.43     0.59
  66 g       2     4     0    279    279    -0.43    -0.36      -0.73     0.92
  67 g       2     4     0    279    279     0.67     1.48      -9.04     9.19
  68 g       2     4     0    279    279     0.11    -0.26      -2.20     2.21
  69 g       2     4     0    279    279     0.30    -0.56      -8.14     8.16
  70 g       2     4     0    279    279     0.09    -0.34     -77.73    77.73
  71 uu_1    2     2     0    279    279    -0.70    -0.53   -1010.20  1010.20
  72 s~      2    16     0    303    303    -7.72   -61.82      28.09    68.34
  73 g       2    16     0    303    303    -0.51    -1.52      -0.69     1.75
  74 g       2    15     0    303    303     0.29    -4.11       7.86     8.87
  75 g       2    15     0    303    303    -0.44    -0.89       3.74     3.87
  76 g       2    15     0    303    303   -14.33     3.48      13.66    20.10
  77 c       2    15     0    303    303   -12.77     8.43      15.24    21.65
  78 u~      2    18     0    323    323   -10.34    14.86       1.47    18.16
  79 g       2    18     0    323    323    -0.93    -0.22       0.03     0.95
  80 g       2    18     0    323    323    -1.50    -0.65       0.67     1.77
  81 g       2    18     0    323    323    -1.62    -2.43      -0.05     2.92
  82 g       2    18     0    323    323    -5.64    -7.04       1.79     9.20
  83 g       2    17     0    323    323    -5.69    -0.85      32.93    33.43
  84 g       2    17     0    323    323    -3.07     0.80       3.43     4.68
  85 g       2    17     0    323    323    -2.04     0.80       3.80     4.39
(...)
```

In this listing, the `ID-st` column indicates the status of the particle. Staus 1 means that the line represents a stable final state particle, status 2 means that the particle is unstable and will decay into status 1 particles. Lines with status 3 have to be interpreted as a sort of "documentation line" which reflects an intermediate state in the treatment of initial and final state radiation, which are calculated by PYTHIA after the event has been produced with ALPGEN or CompHEP.

The aim is to get the kinematics of the primary partons before any radiation has taken place. In the present case of a $t\bar{t}$2j event, for example, the two extra jets correspond to lines 9 and 10. The kinematic information of these lines can't be taken directly, as can be seen in the case of the W+ boson in lines 11 and 20 which has different kinematic properties in the case of the status 2 particle.

The applied algorithm to obtain the kinematical states of the primary partons takes all status 2 particles that have the same status 3 particle as mother (e.g. particles 31 to 36 and 55) and calculates the sum of these fourvectors. This way, exactly two extra jets are obtained in the case of $t\bar{t}$ plus two jets, three extra jets for $t\bar{t}$ plus three jets and so on.

It should be noted that this algorithm is not able to give the exact solution for the primary quark kinematics, if such a solution exists at all, since higher order effects are not taken into account. Another possible solution would be a sort of jet clustering algorithm which combines the final state particles that have the same direction. But also this approach is not optimal, since it depends on the kind of clustering algorithm and its parameters (e.g. cone radius).

Appendix C

Invariant Higgs Boson Mass Distributions

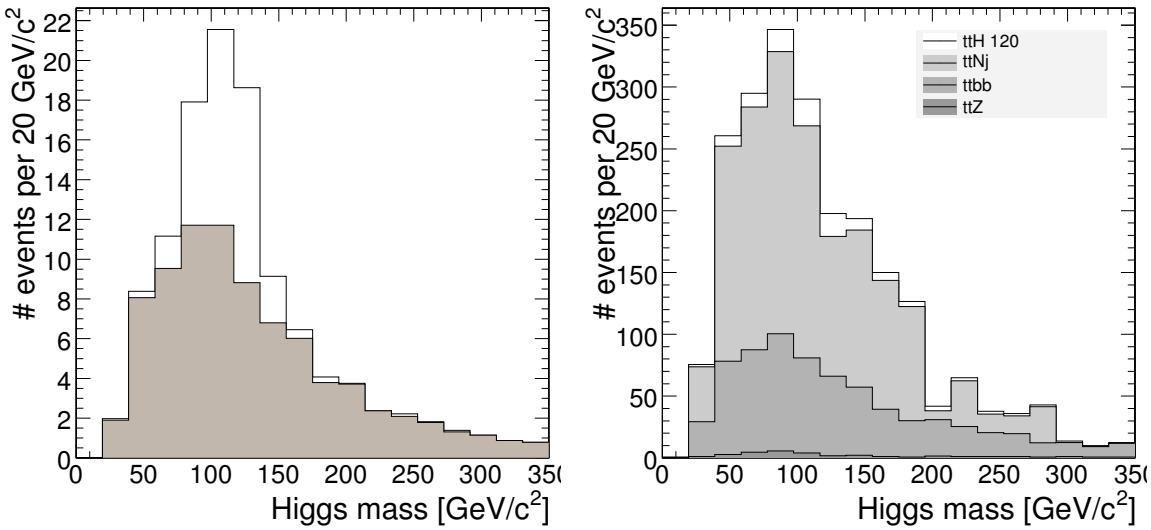

Figure C.1: Invariant Higgs boson mass spectrum for a L_{bTag} cut of 0.225 and m_H = 120 GeV/c^2, after an integrated luminosity of 60 fb^{-1}. On the left: Only signal events; the fraction of combinatorial background is shaded grey. On the right: All relevant physical backgrounds ($t\bar{t}Z$, $t\bar{t}b\bar{b}$ and $t\bar{t}N$j) and the $t\bar{t}H$ signal stacked on top of each other.

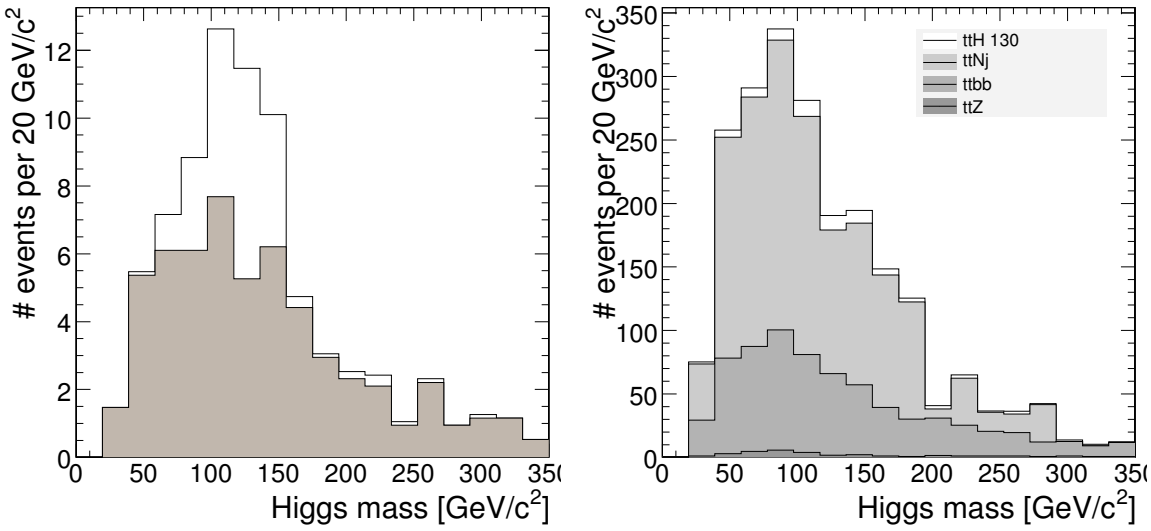

Figure C.2: Invariant Higgs boson mass spectrum for a L_{bTag} cut of 0.225 and m_H = 130 GeV/c^2, after an integrated luminosity of 60 fb^{-1}. On the left: Only signal events; the fraction of combinatorial background is shaded grey. On the right: All relevant physical backgrounds ($t\bar{t}Z$, $t\bar{t}b\bar{b}$ and $t\bar{t}Nj$) and the $t\bar{t}H$ signal stacked on top of each other.

List of Tables

List of Figures

Bibliography

[1] CMS Collaboration, "The CMS Physics Technical Design Report, Volume 2," *CERN/LHCC* **2006-021** (2006). CMS TDR 8.2.

[2] D. Benedetti, S. Cucciarelli, C. Hill, J. Incandela, S. Koay, C. Riccardi, A. Santocchia, A. Schmidt, P. Torre, and C. Weiser, "Search for $H \to bb$ in association with a tt pair at CMS," *CMS Note* **2006/119** (2006).

[3] The LEP Collaborations: ALEPH Collaboration, DELPHI Collaboration, L3 Collaboration, OPAL Collaboration, the LEP Electroweak Working Group, "A Combination of Preliminary Electroweak Measurements and Constraints on the Standard Model," `arXiv:hep-ex/0511027`. CERN-PH-EP/2005-051.

[4] The ALEPH Collaboration, the DELPHI Collaboration, the L3 Collaboration, the OPAL Collaboration, the SLD Collaboration, the LEP Electroweak Working Group, the SLD electroweak, heavy flavour groups, "Precision Electroweak Measurements on the Z Resonance," *Phys. Rept.* **427** (2006) 257, `arXiv:hep-ex/0509008`.

[5] CMS Collaboration, "The Magnet Project Technical Design Report," *CERN/LHCC* **97-010** (1997). CMS TDR 1.

[6] CMS Collaboration, "The Muon Project Technical Design Report," *CERN/LHCC* **97-32** (1997). CMS TDR 3.

[7] S. Cucciarelli, D. Kotlinski, and T. Todorov, "Position Determination of Pixel Hits," *CMS Note* **2002-049** (2002).

[8] CMS Collaboration, "The Tracker Project Technical Design Report," *CERN/LHCC* **98-006** (1998). CMS TDR 5, Addendum CERN/LHCC 2000-016.

[9] CMS Collaboration, "The Electromagnetic Calorimeter Technical Design Report," *CERN/LHCC* **97-033** (1997). CMS TDR 4, Addendum CERN/LHCC 2002-027.

[10] CMS Collaboration, "The Hadron Calorimeter Technical Design Report," *CERN/LHCC* **97-031** (1997). CMS TDR 2.

[11] CMS Collaboration, "The TriDAS Project Technical Design Report, Volume 1: The Trigger Systems," *CERN/LHCC* **2000-38** (2000). CMS TDR 6.1.

[12] F. Halzen and A. D. Martin, "Quarks and Leptons: An Introductory Course in Modern Particle Physics". John Wiley and Sons, 1984.

[13] D. Griffiths, "Introduction to Elementary Particles". John Wiley and Sons, 1987.

[14] A. Djouadi, "The anatomy of electro-weak symmetry breaking. I: The Higgs boson in the standard model," arXiv:hep-ph/0503172.

[15] A. Djouadi, J. Kalinowski, and M. Spira, "HDECAY: A Program for Higgs Boson Decays in the Standard Model and its Supersymmetric Extension," *Comput. Phys. Commun.* **108** (1998) 56–74, arXiv:hep-ph/9704448.

[16] H. Spiesberger, M. Spira, and P. M. Zerwas, "The Standard Model: Physical Basis and Scattering Experiments," arXiv:hep-ph/0011255.

[17] E. Braaten and J. P. Leveille, "Higgs Boson Decay and the Running Mass," *Phys. Rev.* **D22** (1980) 715. doi:10.1103/PhysRevD.22.715.

[18] N. Sakai, "Perturbative QCD Corrections to the Hadronic Decay Width of the Higgs Boson," *Phys. Rev.* **D22** (1980) 2220. doi:10.1103/PhysRevD.22.2220.

[19] T. Inami and T. Kubota, "Renormalization Group Estimate of the Hadronic Decay Width of the Higgs Boson," *Nucl. Phys.* **B179** (1981) 171. doi:10.1016/0550-3213(81)90253-4.

[20] S. G. Gorishnii, A. L. Kataev, and S. A. Larin, "The Width of Higgs Boson Decay into Hadrons: Three Loop Corrections of Strong Interactions," *Sov. J. Nucl. Phys.* **40** (1984) 329–334.

[21] M. Drees and K.-i. Hikasa, "Heavy Quark Thresholds in Higgs Physics," *Phys. Rev.* **D41** (1990) 1547. doi:10.1103/PhysRevD.41.1547.

[22] M. Drees and K.-i. Hikasa, "Note on QCD Corrections to Hadronic Higgs Decay," *Phys. Lett.* **B240** (1990) 455. doi:10.1016/0370-2693(90)91130-4.

[23] K. G. Chetyrkin, "Correlator of the quark scalar currents and Gamma(tot)(H → hadrons) at O(alpha(s)**3) in pQCD," *Phys. Lett.* **B390** (1997) 309–317, arXiv:hep-ph/9608318.

[24] J. Fleischer and F. Jegerlehner, "Radiative Corrections to Higgs Decays in the Extended Weinberg-Salam Model," *Phys. Rev.* **D23** (1981) 2001–2026. doi:10.1103/PhysRevD.23.2001.

[25] D. Y. Bardin, B. M. Vilensky, and P. K. Khristova, "Calculation of the Higgs boson decay width into fermion pairs," *Sov. J. Nucl. Phys.* **53** (1991) 152–158.

[26] A. Dabelstein and W. Hollik, "Electroweak corrections to the fermionic decay width of the standard Higgs boson," *Z. Phys.* **C53** (1992) 507–516.

[27] B. A. Kniehl, "Radiative corrections for $H \to f\bar{f}(\gamma)$ in the standard model," *Nucl. Phys.* **B376** (1992) 3–28. doi:10.1016/0550-3213(92)90065-J.

[28] H. Georgi, S. Glashow, M. Machacek, and D. Nanopoulos, "Higgs Bosons from Two Gluon Annihilation in Proton Proton Collisions," *Phys. Rev. Lett.* **40** (1978) 692. doi:10.1103/PhysRevLett.40.692.

[29] R. N. Cahn and S. Dawson, "Production of Very Massive Higgs Bosons," *Phys. Lett.* **B136** (1984) 196. doi:10.1016/0370-2693(84)91180-8.

[30] K.-i. Hikasa, "Heavy Higgs Production in e+ e- and e- e- Collisions," *Phys. Lett.* **B164** (1985) 385. doi:10.1016/0370-2693(85)90346-6.

[31] G. Altarelli, B. Mele, and F. Pitolli, "Heavy Higgs Production at Future Colliders," *Nucl. Phys.* **B287** (1987) 205–224. doi:10.1016/0550-3213(87)90103-9.

[32] S. Glashow, D. Nanopoulos, and A. Yildiz, "Associated Production of Higgs Bosons and Z Particles," *Phys. Rev.* **D18** (1978) 1724–1727. doi:10.1103/PhysRevD.18.1724.

[33] Z. Kunszt, Z. Trocsanyi, and W. J. Stirling, "Clear signal of intermediate mass Higgs boson production at LHC and SSC," *Phys. Lett.* **B271** (1991) 247–255. doi:10.1016/0370-2693(91)91308-I.

[34] W. Beenakker, S. Dittmaier, M. Kraemer, B. Pluemper, M. Spira, and P. Zerwas, "Higgs radiation off top quarks at the Tevatron and the LHC," *Phys. Rev. Lett.* **87** (2001) 201805, arXiv:hep-ph/0107081.

[35] W. Beenakker, S. Dittmaier, M. Kraemer, B. Pluemper, M. Spira, and P. Zerwas, "NLO QCD corrections to $t\bar{t}H$ production in hadron collisions," *Nucl. Phys.* **B653** (2003) 151–203, arXiv:hep-ph/0211352.

[36] S. Dawson, L. H. Orr, L. Reina, and D. Wackeroth, "Associated top quark Higgs boson production at the LHC," *Phys. Rev.* **D67** (2003) 071503, arXiv:hep-ph/0211438.

[37] ALEPH Collaboration, DELPHI Collaboration, L3 Collaboration, OPAL Collaboration and The LEP Working Group for Higgs Boson Searches, "Search for the Standard Model Higgs boson at LEP," *Phys. Lett. B* **565** (2003) 61–75. doi:10.1016/S0370-2693(03)00614-2.

[38] CDF Collaboration, D0 Collaboration, the TEVNPH Working Group, "Combined D0 and CDF Upper Limits on Standard-Model Higgs-Boson Production," *CDF Note* **8384** (2006).

[39] The LEP Electroweak Working Group, "Status of July 2006." http://lepewwg.web.cern.ch.

[40] S. Abdullin et al., "Summary of the CMS potential for the Higgs boson discovery," *Eur. Phys. J.* **C39S2** (2005) 41–61. doi:10.1140/epjcd/s2004-02-003-9.

[41] CMS Collaboration, "The CMS Physics Technical Design Report, Volume 1," *CERN/LHCC* **2006-001** (2006). CMS TDR 8.1.

[42] V. Innocente, L. Silvestris, and D. Stickland, "CMS software architecture: Software framework, services and persistency in high level trigger, reconstruction and analysis," *Computer Physics Communications* **140** (2001) 31–44. doi:10.1016/S0010-4655(01)00253-3.

[43] **GEANT4** Collaboration, S. Agostinelli et al., "GEANT4: A simulation toolkit," *Nucl. Instrum. Meth.* **A506** (2003) 250–303. doi:10.1016/S0168-9002(03)01368-8.

[44] T. Sjostrand, L. Lonnblad, and S. Mrenna, "PYTHIA 6.2: Physics and manual," arXiv:hep-ph/0108264.

[45] CMS Collaboration, "The TriDAS Project Technical Design Report, Volume 2: Data Acquisition and High-Level Trigger," *CERN/LHCC* **2002-26** (2002). CMS TDR 6.2.

[46] R. Frühwirth, "Application of Kalman Filtering to Track and Vertex Fitting," *Nucl. Instrum. and Methods* **A262** (1987) 444. doi:10.1016/0168-9002(87)90887-4.

[47] V. Innocente, M. Maire, and E. Nagy, "GEANE: Average Tracking and Error Propagation Package," *CERN Program Library* (1991). IT-ASD W5013-E.

[48] C. Weiser, "A Combined Secondary Vertex Based B-Tagging Algorithm in CMS," *CMS Note* **2006/014** (2006).

[49] T. Speer, K. Prokofiev, R. Frühwirth, W. Waltenberger, and P. Vanlaer, "Vertex Fitting in the CMS Tracker," *CMS Note* **2006/032** (2006).

[50] T. Müller, C. Piasecki, G. Quast, and C. Weiser, "Inclusive Secondary Vertex Reconstruction in Jets," *CMS Note* **2006/027** (2006).

[51] C. Piasecki, "Development of the CMS Tracker and Reconstruction of Secondary Vertices of b- and c-Hadrons". PhD thesis, Institut für Experimentelle Kernphysik, Universität Karlsruhe (TH), 2006.

[52] A. Bocci, P. Demin, R. Ranieri, and S. de Visscher, "Tagging b jets with electrons and muons at CMS," *CMS Note* **2006/043** (2006).

[53] D. Benedetti, S. Cucciarelli, J. D'Hondt, A. Giammanco, J. Heyninck, A. Schmidt, and C. Weiser, "Study of jet clustering algorithms at the LHC," in *Les Houches physics at TeV colliders 2005, standard model, QCD, EW, and Higgs working group: Summary report*, pp. 48–55. 2006. arXiv:hep-ph/0604120.

[54] M. Erdmann, D. Hirschbühl, C. Jung, S. Kappler, Y. Kemp, M. Kirsch, D. Miksat, C. Piasecki, G. Quast, K. Rabbertz, P. Schemitz, A. Schmidt, T. Walter, and C. Weiser, "Physics Analysis Expert PAX: First Applications," in *Computing in High Energy and Nuclear Physics (CHEP03)*. La Jolla (USA), 2003. arXiv:physics/0306085.

[55] M. Erdmann, U. Felzmann, D. Hirschbühl, C. Jung, S. Kappler, M. Kirsch, G. Quast, K. Rabbertz, J. Rehn, S. Schalla, P. Schemitz, A. Schmidt, T. Walter, and C. Weiser, "New Applications of PAX in Physics Analyses at Hadron Colliders," in *Computing in High Energy and Nuclear Physics (CHEP04)*. Interlaken (Switzerland), 2004.

[56] M. Erdmann, U. Felzmann, A. Flossdorf, S. Kappler, M. Kirsch, G. Mueller, G. Quast, C. Saout, A. Schmidt, and J. Weng, "Concepts, Developments and Advanced Applications of the PAX Toolkit," in *Computing in High Energy and Nuclear Physics (CHEP06)*. Bombay (India), 2006. arXiv:physics/0605063.

[57] M. Erdmann, U. Felzmann, D. Hirschbühl, S. Kappler, M. Kirsch, G. Quast, A. Schmidt, and J. Weng, "The PAX Toolkit and Its Applications at Tevatron and LHC," *IEEE Transactions on Nuclear Science* **53** (April, 2006) 506–512. doi:10.1109/TNS.2006.870179.

[58] H1 Collaboration, "Internal software manual for H1PHAN".

[59] H. Albrecht, E. Blucher, and J. Boucrot, "ALPHA User's Guide," *ALEPH Internal Note* **99-087** (2000). SOFTWR 99-001.

[60] R. Brun and F. Rademakers, "ROOT - An Object Oriented Data Analysis Framework," *Nucl. Inst. and Meth. in Phys. Res. A* **389** (1997), no. 1-2, 81–86. See also `http://root.cern.ch/`. `doi:10.1016/S0168-9002(97)00048-X`.

[61] L. Lonnblad, "CLHEP: A project for designing a C++ class library for high-energy physics," *Comput. Phys. Commun.* **84** (1994) 307–316.

[62] E. Gamma, "Design Patterns. Elements of Reusable Object-Oriented Software.". Addison-Wesley Professional, 1997.

[63] LCG Collaboration, "LHC Computing Grid : Technical Design Report," *CERN/LHCC* **2005-024** (2005). LCG TDR 001.

[64] CMS Collaboration, "The Computing Project Technical Design Report," *CERN/LHCC* **2005-23** (2005). CMS TDR 7.

[65] The EGEE (Enabling Grids for E-sciencE) Project. http://www.eu-egee.org/.

[66] A. D. Peris, P. M. Lorenzo, F. Donno, A. Sciabà, S. Campana, and R. Santinelli, "LCG-2 User Guide", 2.3 edition, 2005. CERN-LCG-GDEIS-454439.

[67] DataGrid, "Job Description Language HowTo", 2001. DataGrid-01-TEN-0102-0_2.

[68] DataGrid, "JDL Attributes", 2003. DataGrid-01-TEN-0142-0_2.

[69] V. Lefébure and J. Andreeva, "RefDB: The Reference Database for CMS Monte Carlo Production," in *Computing in High Energy and Nuclear Physics (CHEP03)*. La Jolla (USA), 2003.

[70] D. Duellmann, "The LCG POOL Project, General Overview and Project Structure," in *Computing in High Energy and Nuclear Physics (CHEP03)*. La Jolla (USA), 2003.

[71] "PHP: Hypertext Preprocessor." http://www.php.net.

[72] "MySQL." http://www.mysql.com.

[73] "CRAB, CMS Remote Analysis Builder." http://cmsdoc.cern.ch/cms/ccs/wm/www/Crab/.

[74] V. Drollinger, D. Denegri, R. Salerno, and Y. Sirois, "Searching for Higgs Bosons in Association with Top Quark Pairs in the H → bb Decay Mode," *CMS Note* **2001/054** (2001).

[75] V. Drollinger, "Reconstruction and Analysis Methods for Searches of Higgs Bosons in the Decay Mode $H^0 \to b\bar{b}$ at Hadron Colliders". PhD thesis, Institut für Experimentelle Kernphysik, Universität Karlsruhe (TH), 2001.

[76] D. L. Rainwater, M. Spira, and D. Zeppenfeld, "Higgs Boson Production at Hadron Colliders: Signal and Background Processes," `arXiv:hep-ph/0203187`.

[77] W. Beenakker, S. Dittmaier, M. Kraemer, B. Plümper, M. Spira, and P. Zerwas, "Higgs Radiation off Top Quarks at the Tevatron and the LHC," *Phys. Rev. Lett.* **87** (2001) 201805, arXiv:hep-ph/0107081.

[78] M. Mangano. Private communication and rough guesses (CERN), 2006.

[79] M. Dubinin et al., "CompHEP - a package for evaluation of Feynman diagrams and integration over multi-particle phase space," *INP-MSU* **41** (1998) 542.

[80] M. L. Mangano, M. Moretti, F. Piccinini, R. Pittau, and A. D. Polosa, "ALPGEN, a generator for hard multiparton processes in hadronic collisions," *JHEP* **07** (2003) 001, arXiv:hep-ph/0206293. doi:10.1088/1126-6708/2003/07/001.

[81] H. Lai, J. Huston, S. Kuhlmann, F. Olness, J. Owens, D. Soper, W. Tung, and H. Weerts, "Improved Parton Distributions from Global Analysis of Recent Deep Inelastic Scattering and Inclusive Jet Data," *Phys. Rev.* **D55** (1997) 1280–1296, arXiv:hep-ph/9606399.

[82] H. L. Lai, J. Huston, S. Kuhlmann, J. Morfin, F. Olness, J. F. Owens, J. Pumplin, and W. K. Tung, "Global QCD Analysis of Parton Structure of the Nucleon: CTEQ5 Parton Distributions," *Eur. Phys. J.* **C12** (2000) 375–392, arXiv:hep-ph/9903282.

[83] E. James, Y. Maravin, M. Mulders, and N. Neumeister, "Muon Identification in CMS," *CMS Note* **2006/010** (2006).

[84] S. Baffioni, C. Charlot, F. Ferri, D. Futyan, P. Meridiani, I. Puljak, C. Rovelli, R. Salerno, and Y. Sirois, "Electron Reconstruction in CMS," *CMS Note* **2006/040** (2006).

[85] J. Rohlf and C. Tully, "Recommendations for Jet and Missing Transverse Energy Reconstruction Settings and Systematics Treatment," *CMS Internal Note* **2006/025** (2006).

[86] H. Pi, P. Avery, D. Green, J. Rohlf, and C. Tully, "Measurement of Missing Transverse Energy With the CMS Detector at the LHC," *CMS Note* **2006/035** (2006).

[87] J. D'Hondt, S. Lowette, O. Buchmuller, S. Cucciarelli, F. Schilling, M. Spiropulu, S. Mehdiabadi, D. Benedetti, and L. Pape, "Fitting of Event Topologies with External Kinematic Constraints in CMS," *CMS Note* **2006/023** (2006).

[88] S. Kappler, T. Müller, G. Quast, and C. Weiser, "Progress Report on Studies of the Channel $t\bar{t}H$ with $H \to b\bar{b}$ and $t\bar{t} \to WWb\bar{b} \to qq'\mu\bar{\nu}_\mu b\bar{b}$ for CMS in Full Simulation," *CMS Internal Note* **2004/048** (2004).

[89] S. Kappler, "Higgs Search Studies in the Channel $t\bar{t}H$ with the CMS Detector at the LHC". PhD thesis, Institut für Experimentelle Kernphysik, Universität Karlsruhe (TH), 2004.

[90] A. Santocchia, "Optimization of Jet Reconstruction Settings and Parton-Level Correction for the ttH Channel," *CMS Note* **2006/059** (2006).

[91] J. D'Hondt, S. Lowette, J. Heyninck, and S. Kasselmann, "Light quark jet energy scale calibration using the W mass constraint in single-leptonic $t\bar{t}$ events," *CMS Note* **2006/025** (2006).

[92] V. Konoplianikov, O. Kodolova, and A. Ulyanov, "Jet Calibration using γ+jet Events in the CMS Detector," *CMS Note* **2006/042** (2006).

[93] S. Brandt, "Datenanalyse". Spektrum, 1999.

[94] D. Schieferdecker, "Analysis of the $t\bar{t}H$ Channel at the CMS Detector of LHC with Neural Networks." Diploma Thesis, Institut für Experimentelle Kernphysik, Universität Karlsruhe (TH), 2006.

Acknowledgements - Danksagung

Mein Dank geht an Prof. Quast für die ausgezeichnete Betreuung und sein unübertroffenes Engagement gegenüber all seinen Schützlingen. Herrn Prof. Müller danke ich für die Übernahme des Korreferats und seine Unterstützung bei nicht nur wissenschaftlichen Dingen. Beide ermöglichten mir einen 18 monatigen Aufenthalt am CERN in Genf, wo ich eine fantastische Zeit erleben durfte, was ich sehr zu schätzen weiß.

Am CERN teilte ich ein Büro mit Christian Weiser, der mich die Geheimnisse der Teilchenphysik lehrte und in dem ich einen sehr guten Freund fand. Er hat diese Arbeit maßgeblich mitgestaltet, indem er jederzeit bereit war, Fragen aller Art zu beantworten. Ich habe Genf daher nur ungern verlassen.

Ich danke allen, die ich am CERN kennengelernt habe für die tolle Zeit.

Den Mitgliedern des Instituts für Experimentelle Kernphysik danke ich für viele Jahre guter Zusammenarbeit und hervorragender Atmosphäre. Den Administratoren der Computersysteme die ihre Zeit zum Wohle des Institutes zur Verfügung stellen, ebenfalls ein großes Dankeschön. Ich hoffe, daß die hier geschlossenen Freundschaften von langer Dauer sind.

Ich bedanke mich bei den kritischen Korrekturlesern, die geholfen haben, ein fehlerreduziertes Werk zu erhalten: Yves Kemp, Cornelia Krebs et. al., Christian Piasecki, Klaus Rabbertz, Christian Sander, Christian Weiser und Joanna Weng.

I want to thank the members of the CMS $t\bar{t}H$ working group for a very cooperative and constructive period. Especially during the completion of our CMS Note, just before the deadline, they literally worked around the clock: Daniele Benedetti, Susanna Cucciarelli, Chris Hill, Joe Incandela, Sue Ann Koay, Cristina Riccardi, Attilio Santocchia, Paola Torre and Christian Weiser.

Dank gebührt meinen Eltern und meiner Großmutter für die uneingeschränkte Unterstützung in jeder Hinsicht.

Besonders danke ich Cornelia für die Inspiration, die ich durch sie erfahren habe, und für alles andere, das hier niemals Platz finden würde.