Identifikation von Elektronen im Vorwärtsbereich des CDF-Experiments zur Suche nach elektroschwacher Top-Quark-Produktion

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

genehmigte

DISSERTATION

von

Dipl.-Phys. Yves Kemp
aus Innsbruck (A)

Tag der mündlichen Prüfung: 10.02.2006

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

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

Deutsche Zusammenfassung

Das Tevatron in der Nähe von Chicago ist der Kollider mit der zur Zeit höchsten Schwerpunktsenergie. Protonen und Antiprotonen kollidieren mit einer Energie von 1.96 Teraelektronvolt. Diese Kollisionen werden mittels zweier Experimente gemessen: CDF und DØ.

Das Tevatron ist bis zur Inbetriebnahme des *Large Hadron Collider* am CERN der einzige Beschleuniger, mit dem Topquarks erzeugt werden können. Überwiegend werden diese schwersten Quarks als Top-Antitop-Paare durch die starke Wechselwirkung erzeugt. Dieser Prozess ist nachgewiesen worden: Eine Kombination von Messungen mit dem *Collider Detector at Fermilab*, CDF ergibt einen Wirkungsquerschnitt $\sigma_{t\bar{t}} = (7.1 \pm 0.6_{\text{stat.}} \pm 0.7_{\text{syst.}} \pm 0.4_{\text{lumi.}})$ pb. Das Standardmodell der Teilchenphysik sagt weiterhin die Erzeugung von Topquarks über Prozesse der elektroschwachen Wechselwirkung voraus, was aber experimentell noch nicht bestätigt werden konnte. Zwei Prozesse sind am Tevatron von Bedeutung: Der t-Kanal mit einem vorhergesagten Wirkungsquerschnitt von $\sigma_{t\text{-Kanal}} = 1.98^{+0.28}_{-0.22}$ pb und der s-Kanal mit einem vorhergesagten Wirkungsquerschnitt $\sigma_{s\text{-Kanal}} = (0.88 \pm 0.11)$ pb. Diese Prozesse sind unter anderem deswegen interessant, weil der Wirkungsquerschnitt proportional zum Quadrat des CKM Matrixelements $|V_{tb}|$ ist. Die Messung der Wirkungsquerschnitte erlaubt so im Prinzip eine direkte Messung von $|V_{tb}|$ ohne weitere Annahmen bezüglich der Zahl der Quarkfamilien.
Das Topquark zerfällt nahezu ausschließlich in ein W-Boson und ein b-Quark. Das b-Quark wird als Quarkjet nachgewiesen. Die experimentell am leichtesten zugänglichen Zerfallskanäle des W-Boson sind $W \to e\nu_e$ und $W \to \mu\nu_\mu$. Da Neutrinos nur extrem selten mit Materie wechselwirken, können sie mit dem CDF-Experiment nicht direkt nachgewiesen werden. Vielmehr verraten sie sich durch fehlende Transversalenergie.
Die Signatur eines Ereignisses ist also ein isoliertes Elektron oder Muon, fehlende Transversalenergie und zwei oder drei zusätzliche Quarkjets, von denen mindestens einer als b-Quark identifiziert werden muß.

2005 wurde von CDF eine Messung der Wirkungsquerschnitte der elektroschwachen Topquarkerzeugung veröffentlicht, in der Daten von März 2002 bis September 2003 ausgewertet wurden. Dies entspricht einer integrierten Luminosität von 162 pb^{-1}. Diese Messung konnte allerdings nur obere Grenzen für die Wirkungsquerschnitte angeben. Momentan ist eine neue Analyse in Vorbereitung, die einen erweiterten Datensatz ausnützen wird: Daten bis September 2004 finden Verwendung,

was einer integrierten Luminosität von 320 pb^{-1} entspricht. Neben dem erweiterten Datensatz sind noch andere Verbesserungen geplant. Unter anderem wird die geometrische Akzeptanz von Elektronen erhöht, indem Ereignisse analysiert werden, in denen das Elektron im Vorwärtskalorimeter des CDF Experimentes nachgewiesen wurde. Das jetzige Vorwärtskalorimeter wurde erst 2001 nach einem Umbau in Betrieb genommen und wurde bisher noch nicht in einer Analyse verwendet, in der nach Ereignissen gesucht wurde, die als Signatur ein isoliertes Elektron oder Muon und Quarkjets aufweisen.

Simulationsstudien mit Monte Carlo Generatoren zum t-Kanal-Prozess haben gezeigt, daß durch Einbeziehen von Elektronen (beziehungsweise Positronen) im Bereich $1.2 < |\eta| < 2.0$[1] die Akzeptanz um 26.7% erhöht werden kann. Allerdings ist die Identifikation von Elektronen in diesem Bereich schwieriger als im Bereich $|\eta| < 1.0$: Bedingt durch Schwierigkeiten bei der Spurrekonstruktion stehen weniger Informationen zur Verfügung, anhand derer man richtige Elektronen von Untergrund aus QCD-Prozessen unterscheiden kann.

Ein erstes Ziel meiner Arbeit ist, diese Elektronen mittels eines neuronalen Netzes optimal zu identifizieren. Dies ist eine wichtige Voraussetzung für Analysen zur Physik des W-Boson und des Topquarks. Herauszufinden, inwieweit diese neuen Methoden den QCD-Untergrund in W+jets Ereignissen besser unterdrücken, ist ein zweites Ziel meiner Arbeit. Weiterhin zeige ich mögliche Verbesserungen in Analysen mit W-Bosonen auf und mache eine Abschätzung, welchen Zugewinn die Analyse von elektroschwacher Topquarkerzeugung durch Vorwärtselektronen erwarten kann.

Neuronale Netze sind seit langem eine anerkannte Methode, um verschiedene korrelierte Variablen optimal zu kombinieren. In einem ersten Schritt werden sie trainiert, das heißt, sie lernen anhand von historischen oder simulierten Daten die Unterschiede zwischen Signalereignissen und Untergrundereignissen. Im Falle des neuronalen Netzes zur Identifikation von Vorwärtselektronen werden sowohl Signal als auch Untergrund aus Daten gewonnen. Der Signaldatensatz wurde aus $Z \rightarrow e^+e^-$ Ereignissen gewonnen. Unter anderem über die Masse des Z-Bosons hat man eine Kontrolle über die Reinheit des Signaldatensatzes. Der Untergrunddatensatz wurde aus QCD Ereignissen bestimmt, in denen genau zwei Quarkjets auftreten. Anhand kinematischer Betrachtungen kann man auch hier die Reinheit des Untergrunddatensatzes steuern.
In einem zweiten Schritt wird das trainierte neuronale Netz benutzt, um neue Daten in Signal und Untergrund zu klassifizieren. Hier zeigt sich, daß diese neue Methode den Untergrund um 13% besser unterdrückt, als dies die Standardmethode erlaubt.

Die neue Methode der Identifikation mittels neuronaler Netze kommt dann zum Einsatz in Ereignissen, in denen genau ein Elektronkandidat, fehlende Transversalenergie und zwischen null und drei Quarkjets gemessen wurden. In diesen Ereignissen interessiert besonders die Unterdrückung von QCD-Ereignissen, die eine W+jets Signatur vorspiegeln. Da die auf Seitenbändern in den Daten beruhende Standard-

methode der Untergrundabschätzung nicht funktioniert, wenn man Elektronen mit dem neuronalen Netz identifiziert, habe ich zwei neue Methoden eingeführt. Diese zwei neuen Methoden funktionieren für beliebige Schnitte. Beide beruhen auf der Verwendung charakteristischer Verteilungen. Die Anteile von Signal und Untergrund werden durch Anpassung von Signal- und Untergrundvorlagen an die Daten gewonnen. Die erste Methode benutzt als Variable die fehlende Transversalenergie, die zweite Methode die Ausgabe des neuronalen Netzes. Die Ergebnisse der beiden Methoden sind in etwa vergleichbar. Allerdings hat sich gezeigt, daß die Standardmethode den Untergrund systematisch unterschätzt. Die Abschätzung mittels dieser Methode ist unterschiedlich zu meinen beiden Methoden, wenn die Standardschnitte auf die Daten angewendet werden.
Wenn man Schnitte auf die Ausgabe des neuronalen Netzes macht, sind die Resultate im Vergleich zu den Standardschnitten je nach gewähltem Schnitt unterschiedlich. Man kann die Reinheit und Effizienz zum Beispiel so einstellen, daß 7% mehr Signalereignisse und 25% weniger QCD-Untergrundereignisse als mit den Standardschnitten selektiert werden. Allein diese Verbesserung zeigt das Potential dieser Methode der Elektronselektion. Wichtig ist hierbei, daß die Effizienz beziehungsweise die Reinheit des Datensatzes frei wählbar ist, was einen weitereren Vorteil gegenüber der Standardmethode bedeutet.

Die Identifikationsmethode mittels neuronaler Netze steht bereit für den Einsatz zur Suche nach elektroschwacher Top-Quark-Produktion. Eine erste Studie hat gezeigt, daß, je nach Schnittszenario, eine um ca. 26% höhere Akzeptanz im Elektron plus Jets-Kanal durch das Einbeziehen von Vorwärtselektronen erreicht werden kann.

Identification of Electrons in the Forward Region of the CDF Experiment for the Search for Electroweak Top Quark Production

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

genehmigte

DISSERTATION

von

Dipl.-Phys. Yves Kemp
aus Innsbruck (A)

Tag der mündlichen Prüfung: 10.02.2006
Referent: Prof. Dr. Th. Müller, Institut für Experimentelle Kernphysik
Korreferent: Prof. Dr. G. Quast, Institut für Experimentelle Kernphysik

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Until the start of the Large Hadron Collider, the Tevatron Accelerator is the facility that can look deepest into the heart of matter. In Run II of the Tevatron, protons and antiprotons collide at a center of mass energy of 1.96 TeV. Two experiment have been constructed to track known and new phenomena: CDF and DØ. At the end of 2005, more than 1 fb^{-1} of data has been recorded by each experiment.

The Institut für Experimentelle Kernphysik in Karlsruhe is deeply involved in the search for electroweak top quark production at the CDF experiment. Indeed, the heaviest particle found up to now is the top quark. It was discovered in Run I of the Tevatron in 1995, in production mechanisms involving the strong interaction. The electroweak production mode is predicted to have about 40% the cross section of the strong interaction production mode, but the experimental signature makes it more difficult to separate the signal from the larger background. Looking for electroweak top quark production is, however, a very interesting and important field; it is a very good field for testing the Standard Model in many ways.
It is possible to directly measure the CKM matrix element V_{tb}, which could indicate a possible fourth generation of quarks if it deviates significantly from 1. The b quark distribution function of the proton can be measured. Electroweak top quark production is an important background for the search for a light Standard Model Higgs boson at the Tevatron. It is therefore mandatory to understand and measure this production mode well.

In experimental high energy physics, one single person can no longer perform an analysis like the search for electroweak top quark production by himself. Therefore, a group of about 20 people is actively looking for improvements of this CDF analysis. In February 2005, the first CDF search was published with 162 pb^{-1} of Run II data. This search could only set upper limits on the cross section. The next analysis, which is planned to be published in 2006 with more data, will be improved in many ways. My thesis investigates the possibility of enlarging the dataset by taking into account not only electrons detected with the central part of the CDF detector, but also with the forward part. This has not been done yet for the electroweak top quark production, neither for other analyses based on a lepton plus jets signature.

Monte Carlo simulations predict that in the electron decay channel the acceptance can be raised by 26.7% by including electrons in the region $1.2 < |\eta| < 2.0$. η is the pseudorapidity and is defined as $\eta = -\ln\tan\theta/2$. In this formula, θ is the polar angle with respect to the beam axis. The detector is described in more detail in chapter 2.

Chapter 3 gives an overview of the parts of the CDF detector relevant to this analysis. One of my tasks was to maintain and expand the computing cluster EKP-plus which provides the necessary computing power to perform an analysis like this. The lessons from the daily work are briefly summarized in chapter 4. In chapter 5, analysis techniques used by some or all members of the CDF collaboration are detailed. These are not specific for this thesis, but without them, this thesis would, however, not be complete. In chapter 6, I present a new method of identifying forward electrons which uses a neural network to discriminate between real electrons and background from QCD events faking an electron. The variables which the neural network combines in one single variable are the same as those used for the standard cut-based selection method of CDF. The network is trained on data, which is rather unusual. The reasons for this are the bad description of electrons in simulations with Monte Carlo techniques and the quasi-absence of any QCD fake Monte Carlo simulation.

The electron identification method is applied to W+jets events in chapter 7. First, I perform the standard CDF cut-based selection method. I evaluate the QCD background content using the standard 4-sector method separately for all jet multiplicities. The standard 4-sector method can only be applied when using the standard CDF cuts. To estimate the background content in a sample selected using my neural network, I have developed two novel methods which work for any cut scenario. The first method is based on a fit to the missing transverse energy $\not\!\!E_T$. A background template and a signal template are fitted to the data distribution in a range in which signal and background are both present in the data. After a cut on $\not\!\!E_T > 20$ GeV as applied in a typical analysis, the background fraction is computed. The second method is called the *In-Situ* fit method, as the final data sample is taken to evaluate the signal and background content. The signal and background templates derived from the training of the neural network are fitted to the neural network output of the data sample. The fit result is the fraction of signal and background events in data. The self-consistency of all the methods is checked by testing them with an adequate data sample. Different quantities are shown which will help to decide which cut scenario is optimal for a given analysis.

To show that my methods work, I present the transverse mass distribution of W bosons with electrons detected in the forward part of the detector using my algorithm. Also presented is an estimation of the acceptance increase of the data sample used for the electroweak top quark search when using forward electrons.

In the conclusion, I summarize and discuss the results obtained in this thesis.

Chapter 2

The Top Quark in the Standard Model

The Standard Model of Elementary Particle Physics describes the fundamental particles of matter and their interactions except gravity. The Standard Model has been very successful in predicting a vast variety of properties of particles and interactions. The Institut für Experimentelle Kernphysik (EKP) of the Universität Karlsruhe (TH) is deeply involved in measuring the properties of the heaviest particle discovered up to now: the top quark.
This chapter will give a short overview of the properties of the top quark and the different production modes that are relevant at the Tevatron accelerator ring. This chapter will also place my work into the larger picture of the physics program of the CDF experiment. For an in-depth introduction to the Standard Model, I would like to point the reader to textbooks like references [1, 2, 3]. Reference [4] is an overview article which focuses on top quark physics in hadron colliders.

2.1 Properties of the Top Quark

The top quark has been discovered at the Tevatron by the CDF and DØ experiments in 1995 [5, 6]. Up to the turn on of the LHC, the Tevatron remains the only facility which can produce top quarks. Currently the best measurement of the top quark mass is $M_t = (172.7 \pm 2.9)$ GeV/c^2, as one can see in figure 2.1 [7]. The results of different $t\bar{t}$ cross section measurements at CDF are shown in figure 2.2. The combination of these measurements is also shown in this figure. The theoretical $t\bar{t}$ cross section at the Tevatron center of mass system energy of 1.96 TeV depends on the top mass which is shown in figure 2.3. The measured cross section is in good agreement with the theoretical calculations. The helicity of the W boson in top quark decays is also accounted for as a property of the top quark. The latest most precise measurement states that the fraction of longitudinally polarized W bosons is $F_0 = (81.3^{+11.4}_{-12.4}(\text{stat.})^{+4.7}_{-3.7}(\text{sys.}))\%$ [8] while the Standard Model predicts $F_0 = 70\%$, which is in good agreement.

The width of the top quark is predicted to be $\Gamma_t = 1.57$ GeV/c^2 at an assumed top quark mass of 180 GeV/c^2 [9]. The top quark lifetime is $\tau_t \approx 4 \cdot 10^{-25}$ s. This is

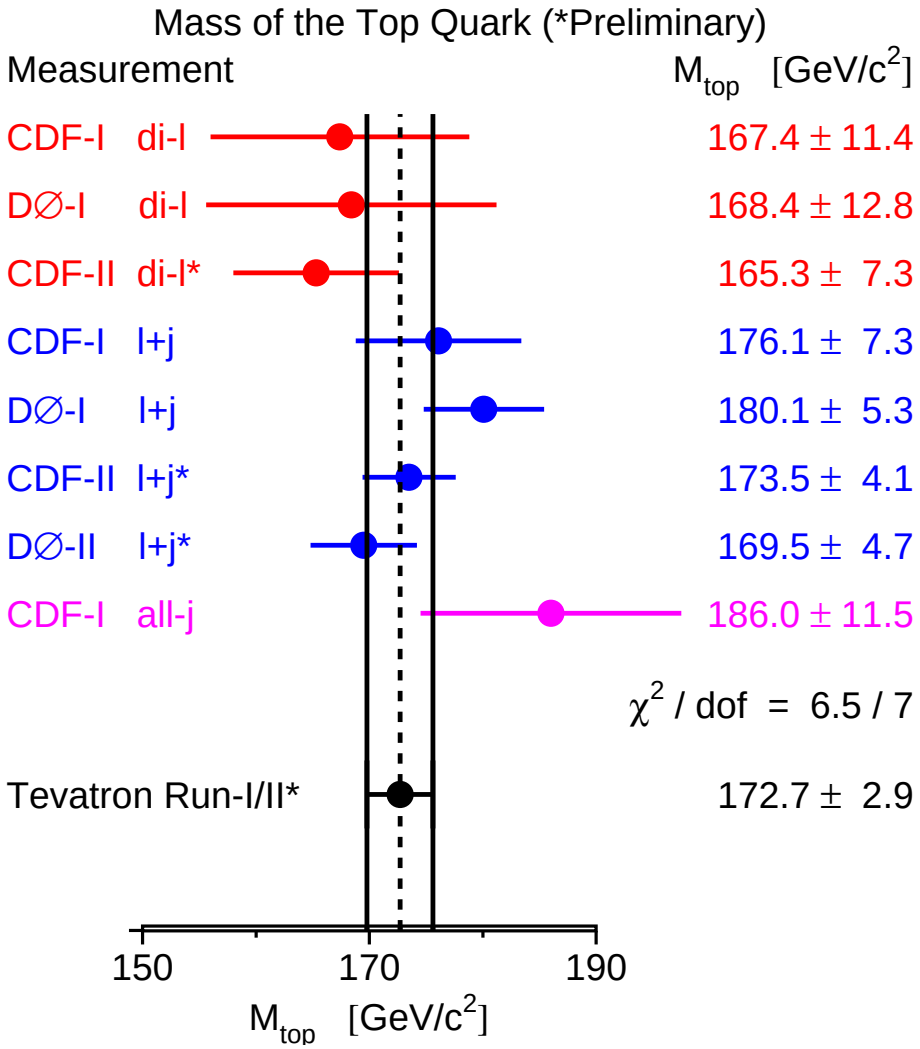

Figure 2.1: CDF and DØ measurements of the top mass, combined by the Tevatron Electroweak Working Group [7].

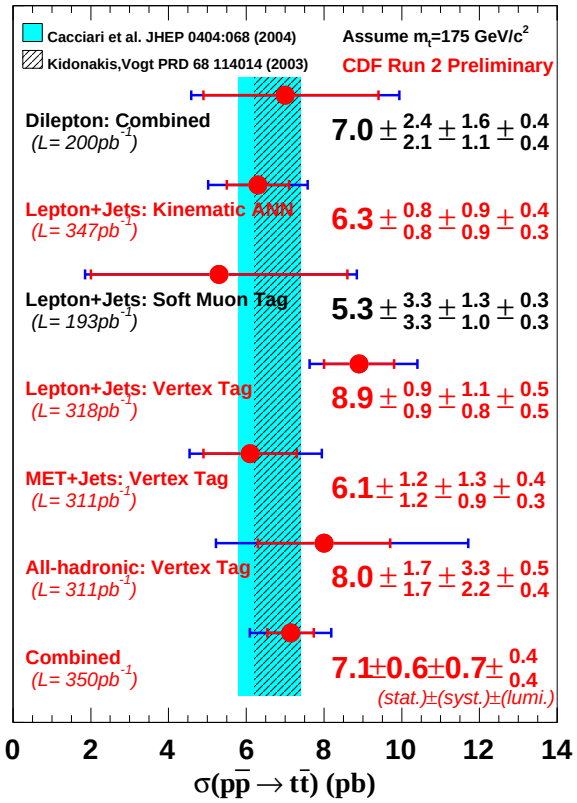

Figure 2.2: CDF measurements of the $t\bar{t}$ cross section in pb and their combination [12].

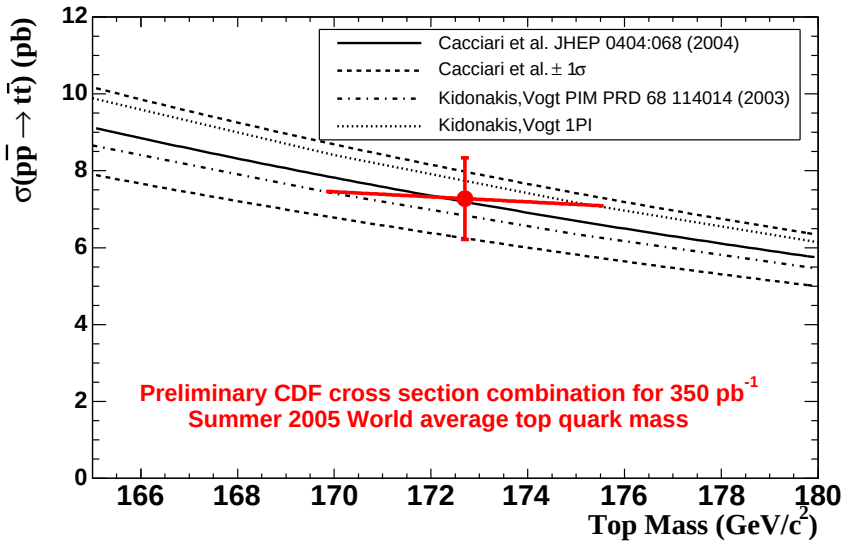

Figure 2.3: $t\bar{t}$ cross section versus top mass, theoretical calculations and the combined measurement. The measurement is in good agreement with the theory [10, 11]. Different CDF cross section measurements have been combined in reference [12].

shorter than the QCD timescale $\tau_{QCD} \approx 3 \cdot 10^{-24}$ s, therefore the top quark decays before forming a hadron. In this sense, one can say that the top quark is a quasi free particle, which allows for interesting measurements. The decay of the top quark is governed by the Cabibbo-Kobayashi-Maskawa mixing matrix. This matrix defines the transformation from the mass eigenstates to the eigenstates of the electroweak interaction. By convention, it is expressed by a 3×3 unitary matrix V operating on the charge $-e/3$ quarks.

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix}$$

The 90% confidence limits on the magnitude of the elements of the complete matrix are [13]:

$$\begin{pmatrix} 0.9739 \text{ to } 0.9751 & 0.221 \text{ to } 0.227 & 0.0029 \text{ to } 0.0045 \\ 0.221 \text{ to } 0.227 & 0.9730 \text{ to } 0.9744 & 0.039 \text{ to } 0.044 \\ 0.0048 \text{ to } 0.014 & 0.037 \text{ to } 0.043 & 0.9990 \text{ to } 0.9992 \end{pmatrix}$$

Not all of these numbers have been measured experimentally, some have been deduced from the assumed unitarity of this 3×3 matrix. This assumption relies on the existence of three and only three families (or generations) of particles. As V_{tb} is much larger than V_{td} and V_{ts} and very close to 1, the top quark decays nearly exclusively into a W boson and a b quark.
The different decay modes of the W boson and their branching fractions are given in table 2.1

Decay mode	Branching fraction (in %)
$W \to e\nu_e$	$(10.75 \pm 0.13)\%$
$W \to \mu\nu_\mu$	$(10.57 \pm 0.15)\%$
$W \to \tau\nu_\tau$	$(11.25 \pm 0.20)\%$
$W \to$ hadrons	$(67.60 \pm 0.27)\%$

Table 2.1: Decay modes of the W boson and their branching fractions [13].

2.2 Production Modes of the Top Quark at the Tevatron

At the Tevatron, the most important production mode of top quarks is the production of top quark pairs via the strong interaction. The Standard Model predicts the production of single-top quarks via electroweak interactions. The latter production mode is predicted to have a cross section of about 40% of the first production mode. However, as the background for the electroweak production mode is much higher than for the strong interaction mode, the top quark was discovered at the Tevatron

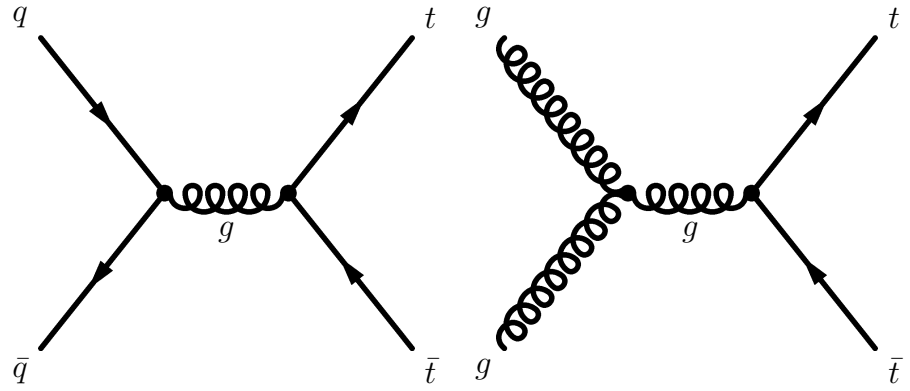

Figure 2.4: Main tree level Feynman diagrams for the $t\bar{t}$ production. The left diagram accounts for 85% of the cross section, the right diagram and other diagrams with two gluons in the initial state for 15% at the Tevatron.

in 1995 in the top quark pair production mode. The single-top production mode still remains to be discovered.

2.2.1 Top Quark Pair Production

Two groups have independently computed the theoretical cross section. With an assumed top quark mass of 175 GeV, they obtain $\sigma_{t\bar{t}} = 6.70^{+0.71}_{-0.88}$ pb [10] and $\sigma_{t\bar{t}} = (6.77 \pm 0,42)$ pb [11] at the Tevatron. One can see that these two different theoretical computations are in very good agreement. Different measurements of $\sigma_{t\bar{t}}$ at CDF are shown in figure 2.3. They have been combined to $\sigma_{t\bar{t}} = (7.1 \pm 0.6_{\text{stat.}} \pm 0.7_{\text{syst.}} \pm 0.4_{\text{lumi.}})$ pb [12]. This result is in good agreement with the theoretical computations. Figure 2.4 shows some of the tree level Feynman diagrams for the $t\bar{t}$ production. At the Tevatron, diagrams with quarks in the initial state account for 85% of the cross section whereas diagrams with gluons in the initial state account for the remaining 15%. The two top quarks decay mostly into a W boson and a b quark. Top-antitop quark events are experimentally classified according to the decay mode of the W bosons. There are four main categories:

- The so-called leptonic or di-lepton channel. In this channel, both of the two W bosons decay either into an electron or muon and their respective neutrino. This channel has the clearest experimental signature and the lowest backgrounds, but occurs only in 5% of the $t\bar{t}$ events.

- The lepton plus jets channel. In this channel, one W boson decays into an electron or muon and its respective neutrino while the other W boson decays into two quarks, forming jets. Events with this decay topology occur in 30% of all $t\bar{t}$ events. This channel is not as clean as the di-lepton channel, but still experimentally well accessible. One could say that this channel is the golden channel. The signature of such an event is one isolated electron or muon, two light flavor jets, two b flavor jets and missing transverse energy ($\not\!\!E_T$) originating

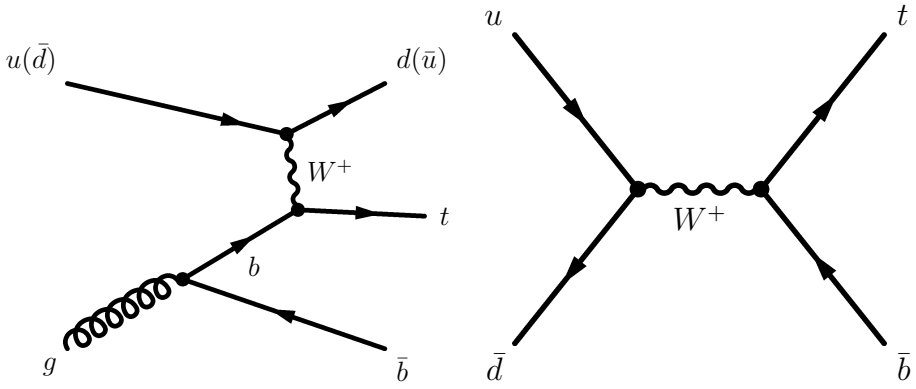

Figure 2.5: Example Feynman diagram for the single-top production modes. The left diagram represents the t-channel or gluon-fusion process, the right diagram represents the s-channel process.

from the neutrino that is not detected. The main background sources are W+-jets events and QCD multijet events that contain a misidentified electron or muon and mismeasured $\not\!\!E_T$.

- The third channel is the all-hadronic mode. In this channel, both W bosons decay in a quark-antiquark pair. This channel occurs in 44% of all $t\bar{t}$ decays. However, the signature is rather challenging as a total of six jets are formed. The background from QCD is rather important. Additionally, when reconstructing the W bosons and the top quarks for a mass measurement or to form a discriminating variable, the number of combinations is very large, and can only partially be resolved by constraints on the W mass or by applying b tags.

- The remaining 20% of the decays involve decays of W bosons into τ leptons. The identification of τ leptons is difficult as it has many different decay modes. Therefore, no top quark mass or $t\bar{t}$ cross section measurement has been published with CDF data involving this channel up to now.

The $t\bar{t}$ production mode is best suited to measure the top quark mass, as this mode has the highest statistics. The W helicity in top quark decays was also measured in this channel.

2.2.2 Electroweak Top Quark Production

The two most important diagrams for electroweak top quark production are shown in figure 2.5. One can see that only one top or antitop quark is produced, hence the name single-top quark production. The cross sections expected at the Tevatron are $1.98^{+0.28}_{-0.22}$ pb and (0.88 ± 0.11) pb for the t-channel and the s-channel respectively [14, 15]. In total, single-top production amounts to about 40% of the $t\bar{t}$ cross section. However, the background rates with the same experimental signature are much larger. The cleanest channel to look at is the electron or muon decay channel of the W boson. However, single-top events only come with two, rarely three additional

jets. Therefore, W+jet production and QCD multijet events have a greater impact on the background. Another important background is $t\bar{t}$ lepton plus jets production.

The electroweak production channels are best suited for measuring the CKM matrix element V_{tb}, since the quark production cross section is proportional to $|V_{tb}|^2$. These channels are therefore sensitive to a potential fourth generation or other effects diminishing V_{tb}. Also with V_{tb} close to 1, a deviation from the predicted cross section could indicate new physics, like Flavor Changing Neutral Currents.
Additionally, the t-channel is a probe for the b quark PDF of the proton.
Last, it is important to understand well the single-top quark production processes. They have, especially the s-channel, the same signature as the associated production of a light Higgs boson with a W boson.
The single-top production processes have not been discovered yet, i.e. their cross section has not been determined experimentally. However, in Run I and Run II, upper limits have been set on the cross sections [16, 17, 18, 19]. The latest prospects predict that, given the Standard Model cross section, a first evidence can be achieved with 1.5 fb^{-1} of integrated luminosity [20]. In the previous CDF analysis [18], not all available data was used: for instance, W bosons decaying into electron and neutrino were only considered if the electron was in the acceptance region of the central part of the detector. This central part of the detector is equivalent to a pseudorapidity range $|\eta| < 1.0$. However, the Standard Model predicts that one can increase the acceptance for electrons by 26% when extending the detection region to $1.2 < |\eta| < 2.0$ for electrons with $P_T > 20$ GeV/c. Figure 2.6 shows the η distribution of electrons in t-channel Monte Carlo events generated by MadEvent. One distribution shows the η distribution for all electrons, the other shows the η distribution for electrons which have at least $P_T > 20$ GeV/c, to simulate the analysis cuts. One can see that this requirement cuts slightly more electrons in the forward region (B and C in the plot) as in the central region (A). From the theory, the total momentum of the electron should be independent of its direction. Forward electrons, however, have more longitudinal momentum as central electrons, therefore they will have a lower transverse momentum. The electrons in the very forward region $|\eta| > 2.0$ are experimentally hard to identify since there is no tracking information available in this region. Compared to the number of central electrons, only 5.2% of the electrons are in region C. The main background to electrons are jets from hard QCD jet production. The fall off of the η distribution for electrons from W bosons is shown in figure 2.6. The fall off of the η distribution for jets from hard QCD dijet production is measured in reference [22]. This measurement, although it is done with data from Run I, clearly shows that jets from hard QCD dijet production gain importance compared to electrons from W boson decays at higher values of η. The background expectation is therefore higher in region B and region C than in region A.

This thesis will deal with the identification of the electrons in the pseudorapidity region $1.2 < |\eta| < 2.0$, and the prospects for the search for electroweak top quark production.

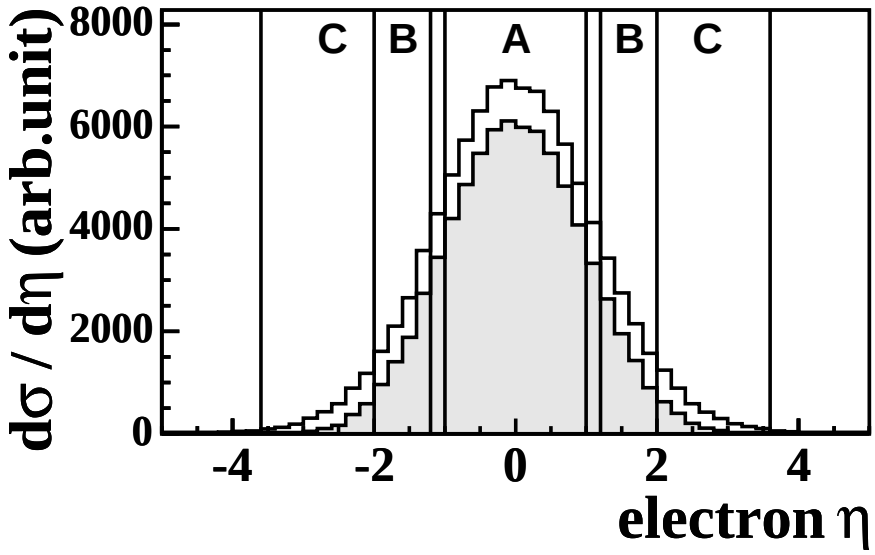

Figure 2.6: η distributions for electrons without a P_T requirement (solid line) and with $P_T >$ 20 GeV/c (gray area). Three regions are indicated, region A ($|\eta| < 1.0$) represents the central calorimeter, region B ($1.2 < |\eta| < 2.0$) represents the part of the forward calorimeter where tracking is possible and region C represents the remaining part of the forward calorimeter ($2.0 < |\eta| < 3.6$) which is not covered by any tracking detector. The sample used for this plot is a MadEvent Monte Carlo simulation of the t-channel mode for single-top quark production as described in reference [21]. 52905 (61077) events are in region A, 13902 (19860) are in region B and 2741 (7697) are in region C fulfilling the P_T requirement (in brackets numbers without P_T requirement).

Chapter 3

The Experiment

The CDF experiment is located at the Tevatron collider at the Fermi National Laboratory (Fermilab or FNAL), in Batavia/Illinois (USA), where protons and antiprotons circulate in opposite directions in a ring with a diameter of 2 km. The Tevatron is the accelerator with the highest center of mass energy currently in operation. The first collisions at a center of mass energy of 1.8 TeV were initiated in 1985. The data collected until 1996 in the so-called Run I phase amount to 106 pb^{-1} (as used in reference [23]) and allowed, among other interesting results, the first experimental evidence of the top quark, followed by the precise determination of its mass. Starting in 1996, the accelerator complex was upgraded to increase the instantaneous luminosity and the center of mass energy to 1.96 TeV. CDF and DØ, the second Tevatron experiment, were upgraded as well. The Run II phase started at the end of 2001 and is scheduled until 2009. During this time 4.4 to 8.5 fb^{-1} of data are expected to be delivered[1]. The physics programs of CDF and DØ include Higgs searches, top quark physics, rare processes and the measurement of the frequency of the $B_s^0 - \bar{B}_s^0$ oscillation. In this chapter the experimental setup of the CDF detector is presented with special focus on the facilities relevant to this analysis.

3.1 The Accelerators

The Tevatron is one of the large facilities at FNAL. Protons and antiprotons are brought to collision at a center of mass energy of 1.96 TeV. To reach this center of mass energy, a system of different accelerators is needed.
The first step in the accelerator chain is the Cockroft-Walton pre-accelerator. Hydrogen gas is ionized to create negative ions that are accelerated by a positive voltage to an energy of 750 keV. The negative ions then enter a linear accelerator, called LINAC, about 130 m long, which accelerates the ions to 400 MeV by means of an oscillating electric field. The ions then pass a carbon foil, where the electrons are stripped off. The next step is the booster, a circular accelerator that uses magnets

[1] The delivered luminosity is the luminosity produced by the accelerator. The recorded luminosity differs as downtimes of the detector or the data acquisition are not included. Again a subset of the recorded luminosity is used for a specific analysis, which requires certain parts of the detector to be in a good state (good run requirement).

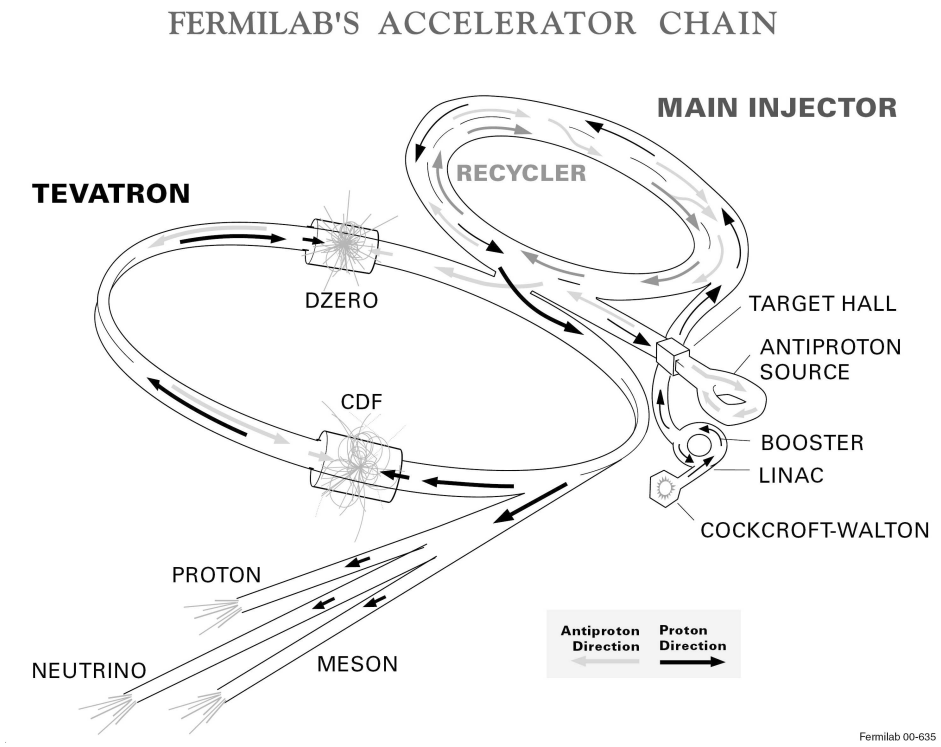

Figure 3.1: Schematic view of the Tevatron accelerator chain.

to bend the beam of protons into a circular path. After some 20.000 revolutions, the protons leave the booster with an energy of 8 GeV.
Protons are then transferred to the Main Injector which has four functions:

- It accelerates protons from 8 GeV to 150 GeV
- It produces 120 GeV protons, which are used for antiproton production
- It receives antiprotons from the Antiproton Source and increases their energy to 150 GeV
- Finally, it injects protons and antiprotons into the Tevatron.

Additionally, the Main Injector tunnel hosts the so-called Recycler. The initial plan was to reuse antiprotons from former collider stores. This plan has, however, been abandoned. Now, the only but important purpose of the Recycler is the cooling and stacking of fresh antiprotons. Electron cooling of the antiprotons has been done in the Recycler since July 2005, resulting in a higher luminosity.
To produce the antiprotons, the Main Injector sends 120 GeV protons to the Antiproton Source, where the protons collide with a nickel target. The collisions produce a wide range of secondary particles including many antiprotons. The antiprotons are collected, focused and then stored in the Accumulator ring. When a sufficient number of antiprotons has been produced, they are sent to the Main Injector.
The last step in the long chain of accelerators to reach the 1.96 TeV energy is

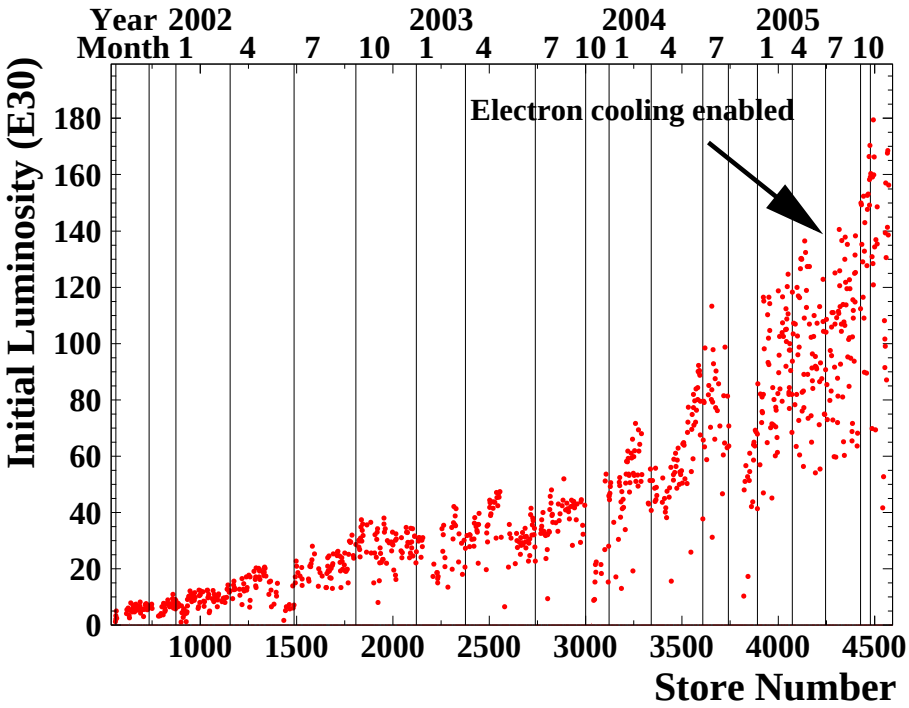

Figure 3.2: Initial luminosity per store [$cm^{-2}s^{-1}$].

the Tevatron accelerator and storage ring, a collider with a circumference of about six kilometers. Protons and antiprotons are circulating in opposite directions at 0.98 TeV. They are brought to collision at two interaction points: D0 and B0 (where the CDF experiment is located).
Figure 3.1 shows a schematic view of the Tevatron accelerator chain.

The energies of the protons and antiprotons determine the cross section σ of the physical processes one wants to observe. The number n of produced events in a time period is given by $n = \sigma \int Ldt$. The quantity L is called instantaneous luminosity, the quantity $\int Ldt$ is the integrated luminosity over time. In the particle physicists jargon, the $\int Ldt$ is often only referred to as luminosity and distinguished from the instantaneous luminosity by looking at the units.

In the beginning of Run II in June 2001, the instantaneous luminosity did not meet the design goals, partially because the new main injector and recycler were not well understood and under control. As knowledge about the accelerators grew, the instantaneous luminosity increased, as can bee seen in figure 3.2 [24]. A lot of improvements have been made by the Beams Division of Fermilab after 2001 to achieve a better performance. To give just one example, electron cooling of the antiprotons in the recycler storage ring has been installed.

The amount of data delivered by the Tevatron and written to tape is presented as a function of time in figure 3.3.

It is planned to continue the Tevatron operation until the end of the fiscal year 2009. The baseline goal is to achieve an integrated luminosity of 4.4 fb^{-1}, the design luminosity goal is 8.4 fb^{-1}. Now it seems that the design luminosity will be reached,

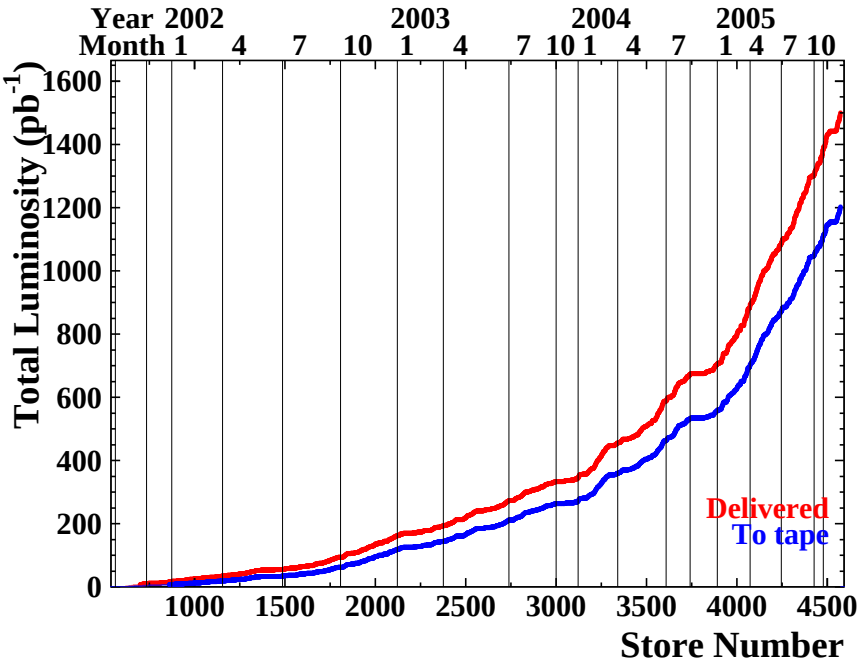

Figure 3.3: Delivered (upper curve) and recorded (lower curve) integrated luminosity since the start of Run II. This analysis uses data taken from March 2002 to September 2004.

for the end of the fiscal year 2005 it was 1.2 fb^{-1}, a goal that was attained. More details about the Tevatron goals can be found in reference [25], up-to-date and historical information about the Tevatron performance is found on the web page given in reference [26].

3.2 The Collider Detector at Fermilab

Two multipurpose detectors are measuring the collisions of protons and antiprotons at the Tevatron: CDF and DØ. The general layout is similar for both of these detectors, they cover most of the 4π solid angle around the beam spot and feature azimuthal and forward-backward symmetry. Both can measure the tracks made by charged particles in the core of the detector, the energy deposit in calorimeters and identify muons. Magnetic fields help identifying charged particles. As this analysis was performed with the CDF experiment, the key features of this experiment needed for this analysis will be described. An in-depth description can be found in the technical design report [27].

The CDF was built and is maintained by a collaboration of more than 50 institutions in eleven countries. The Institut für Experimentelle Kernphysik in Karlsruhe is the only German institute in this collaboration and member since 1996.

Figure 3.4 shows an elevation view of one half of the CDF II detector. In the following sections, angles and directions are often referred to the CDF coordinate system. The polar angle θ is measured with respect to the proton beam axis (z-axis),

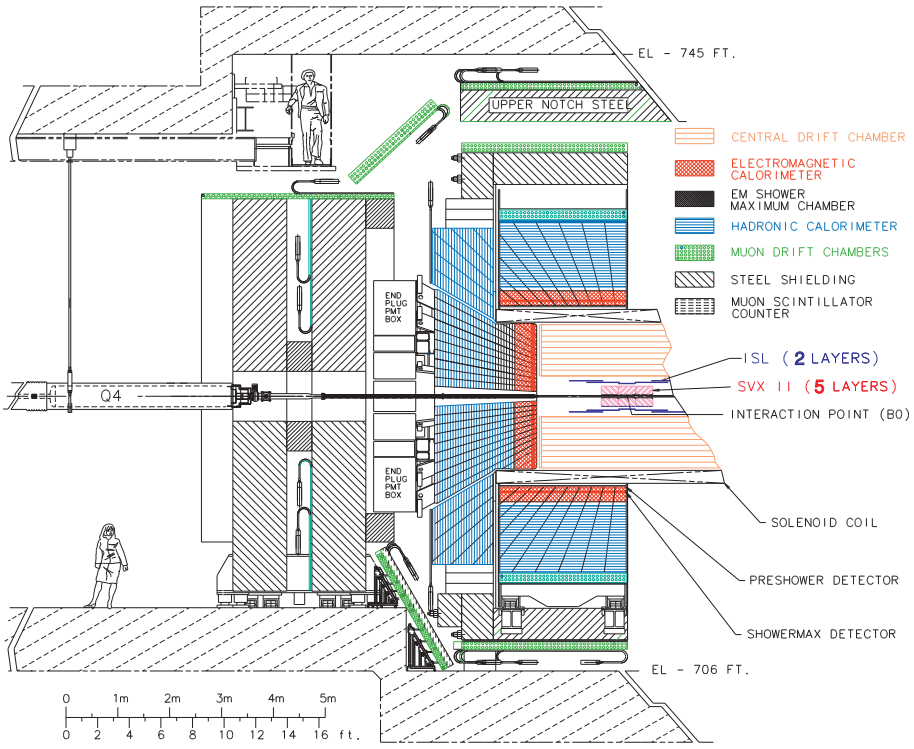

Figure 3.4: Elevation view of one half of the CDF detector in Run II.

pointing in east direction. The azimuthal angle φ is measured from the plane of the Tevatron. Transverse and longitudinal are meant with respect to the proton beam, i.e. parallel and perpendicular to the proton beam respectively. An often-used quantity is the pseudorapidity defined by $\eta = -\ln\left(\tan\frac{\theta}{2}\right)$.

3.2.1 The Tracking System

The Tracking System in Run II consists of 4 parts: Layer 00, the SVX II, a silicon vertex detector, the ISL, Intermediate Silicon Layers, and the COT, the Central Outer Tracker, an open drift chamber.

A schematic overview of the Layer 00, SVX II and ISL detectors is shown in figure 3.5. Glued to the beam pipe, Layer 00 is closest to the beam, with its modules placed at radii $r = 1.35$ cm and $r = 1.62$ cm of the beam pipe. It is a single-sided radiation hard silicon microstrip detector and provides a coverage of $|\eta| < 4.0$. Layer 00 was added later to the design of the vertex detector to enhance its resolution and longevity [28].

Layer 00 is enclosed by the SVX II. The SVX II detector design is driven by high luminosity, the Tevatron short bunch spacing of 396 ns, and by the physics requirement of b decay vertex identification within collimated high-P_T jets [29]. SVX II is comprised of three cylindrical barrels which cover $\approx 2.5\sigma$ of the interaction region providing track information to pseudorapidity $|\eta| < 2$. Five layers of double-sided silicon sensors at radii from 2.4 to 10.7 cm supply $r - \varphi$ as well as 3 $r - z$ and

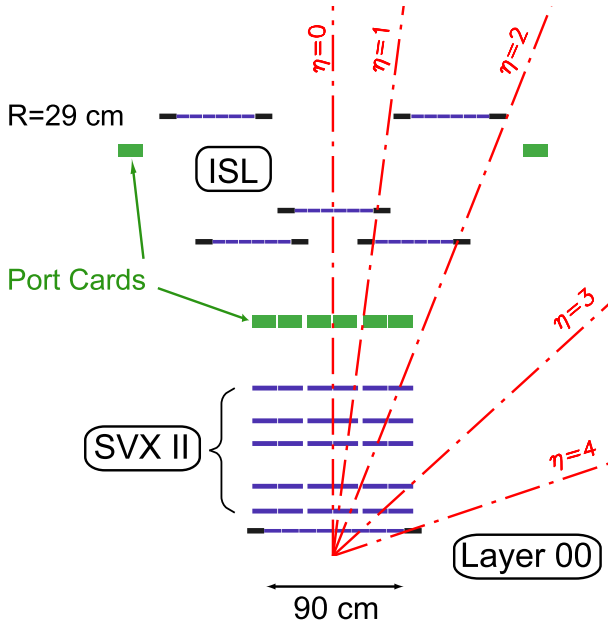

Figure 3.5: Schematic view of the Layer 00, SVX II and ISL silicon tracking detectors.

2 small angle stereo measurements. The results provide good pattern recognition and 3-d vertex reconstruction with an impact parameter resolution $\sigma_\varphi < 30$ μm and $\sigma_{z_0} < 70$ μm for central high momentum tracks. The impact parameter is the distance of closest approach of the track helix to the beam axis measured in the plane perpendicular to the beam. The SVX II provides coverage up to $|\eta| \approx 2$. In the region $|\eta| < 1$ the combination of the SVX II and the COT can provide full 3D tracking, but the reconstruction is mainly anchored on COT tracks. For $|\eta| > 1$, SVX II can only allow for 2D tracking. To increase the tracking volume, the three layers of the silicon detector ISL are placed between the SVX II and the COT [30]. The outer part of the tracking system is the Central Outer Tracker, a 3.1 m long cylindrical open drift chamber, to provide tracking at large radii in the region $|\eta| < 1.0$. The COT covers the radial range from 40 to 137 cm and provides 96 measurement layers organized into alternating axial and stereo superlayers. The hit position resolution is approximately 14μm and the momentum resolution $\sigma(p_T)/p_T^2 = 0.0015(GeV/c)^{-1}$. Due to the high luminosity and the short bunch spacing, the COT is designed to operate with a maximum drift time of 100 nsec by reducing the maximum drift distance and by using a gas mixture with a fast drift velocity [27].

3.2.2 The Calorimetry System

The solenoid and tracking volume is surrounded by the calorimeters, designed to measure the energy of particles and jets by fully absorbing all particles except muons and neutrinos. There are, altogether, five calorimeter systems: the central electromagnetic calorimeter (CEM), the central hadron calorimeter (CHA), the end-wall hadron calorimeter (WHA), the end-plug electromagnetic (PEM) and hadron

calorimeter (PHA), covering 2π in azimuth and pseudorapidity $|\eta| < 3.6$. Each calorimeter module is divided into projective towers, pointing to the nominal interaction point. The calorimeters are sampling calorimeters. The active medium is a scintillator, the absorber is lead in the electromagnetic calorimeter and iron in the hadronic calorimeter. The different energy resolutions and segmentation in η and φ for the several calorimeters are given in table 3.1. The central calorimeters are described in more detail in references [32, 33, 34].

System	η range	$\Delta\varphi$	$\Delta\eta$	Energy resolution
CEM	$\lvert\eta\rvert < 1.1$	$15°$	≈ 0.1	$14\%\sqrt{E_T}$
PEM	$1.1 < \lvert\eta\rvert < 1.8$	$7.5°$	≈ 0.1	$16\%\sqrt{E_T}$
	$1.8 < \lvert\eta\rvert < 2.1$	$7.5°$	≈ 0.16	
	$2.1 < \lvert\eta\rvert < 3.64$	$15°$	$0.2 - 0.6$	
CHA	$\lvert\eta\rvert < 0.9$	$15°$	≈ 0.1	$75\%\sqrt{E}$
WHA	$0.7 < \lvert\eta\rvert < 1.3$	$15°$	≈ 0.1	$80\%\sqrt{E}$
PHA	$1.2 < \lvert\eta\rvert < 1.8$	$7.5°$	≈ 0.1	$5\% + 80\%\sqrt{E}$
	$1.8 < \lvert\eta\rvert < 2.1$	$7.5°$	≈ 0.16	
	$2.1 < \lvert\eta\rvert < 3.64$	$15°$	$0.2 - 0.6$	

Table 3.1: Summary of the CDF calorimeter properties in Run II. CEM and CHA are the central electromagnetic and the central hadronic calorimeters respectively. PEM and PHA are their counterparts in the plug region. WHA is the end wall hadronic calorimeter. The transverse energy E_T and the energy E are given in GeV. The resolution was measured in test beam data using electrons for the electromagnetic calorimeters and using single pions for hadronic calorimeters. $\Delta\varphi$ and $\Delta\eta$ are the segmentation in azimuth and pseudorapidity respectively.

3.2.3 The Plug Upgrade Calorimeter

In contrast to the central calorimetry, the forward calorimetry was upgraded for Run II of the Tevatron [27, 35]. These systems are called "Plug" or "CDF Plug Upgrade" calorimeter in the CDF jargon. The plug calorimeter is a shower-sampling device consisting of plastic scintillating plates with optical fiber readout. It replaces the previous gas sampling calorimeters employed in Run I. This new technology can cope better with the higher bunch crossing rate. It represents the first application of the tile-fiber technique on a large scale. Besides faster response of the calorimeter, another design criterion was a performance comparable to that of the central calorimeter system.

The calorimeter part closest to the interaction point is the plug electromagnetic calorimeter (PEM), followed by the hadron calorimeter, as shown in figure 3.6. The electromagnetic calorimeter is divided into 22 sets of alternating layers of absorber and polystyrene scintillator. The absorber plates are 4.5 mm thick lead sheets, glued together at the edges with a stainless steel supporting structure. The thickness of the scintillator is 4 mm. There are 24 sections or wedges of 15^0 each, the separation into the 20 towers can be seen in figure 3.7. The coverage in the radial angle φ is

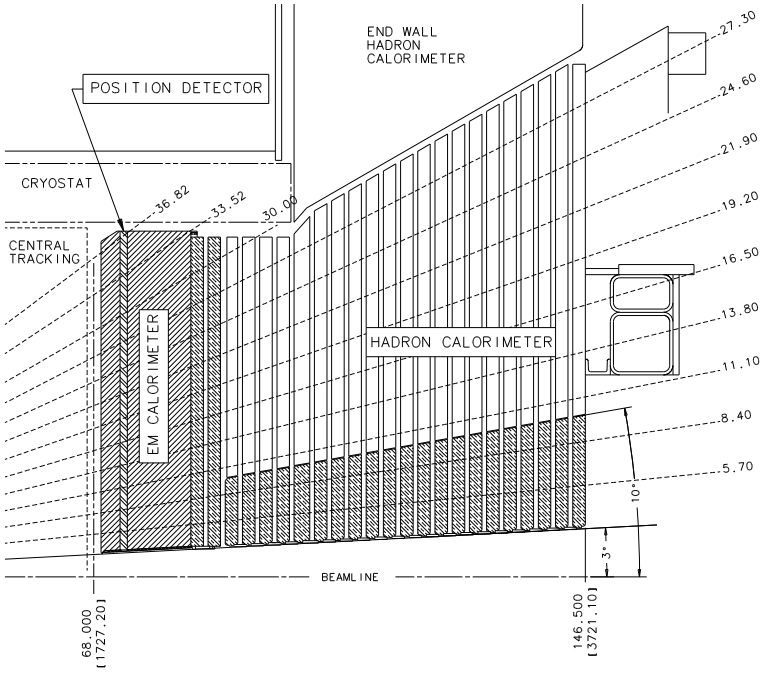

Figure 3.6: Sketch of the plug calorimeter system.

2π, the coverage in pseudorapidity is $1.1 < |\eta| < 3.6$. Thus, the tower segmentation is roughly $7.5^0 \times 0.12\eta$. In front of the first lead layer is another scintillator layer, which is read out separately from the rest of the calorimeter, to act as a preshower detector (PPR). Its structure is the same, except that the thickness is of 10 mm.

The plug hadron calorimeter (PHA) covers $1.3 < |\eta| < 3.64$ and has a structure similar to the PEM, except that the thickness of the scintillator is 6 mm and the absorber is of 5.08 cm thick iron. A total of 23 sets of scintillator/absorber layers form the PHA.
Most jets shower inside the hadron calorimeter. However, some shower early, inside the EM section. It is difficult to distinguish electrons from these jets using just the energy deposited inside the calorimeter. Inside the EM section is a shower maximum detector. The precise location of the energy deposition in this subdetector helps distinguishing jets from electrons. This detector also allows the separation of photons from π^0. The shower maximum detector (PES) is a coarse tracking chamber, located where the average EM object (electron, photon or π^0) deposits the largest fraction of energy. This is equivalent to six radiation length of material. Therefore, the PES is just after the fourth lead plate. The shower maximum detector is segmented in azimuth into eight sectors of 45^0 each, every segment consists of two layers, u and v, oriented at $\pm 22.5^0$ with respect to the radial dimension shown in figure 3.8. The layers are constructed of 200 5 mm wide by 6 mm deep strips of scintillator material. The strips are installed such that the detector is segmented into low ($1.13 < |\eta| < 2.60$) and high ($2.60 < |\eta| < 3.50$) η regions.

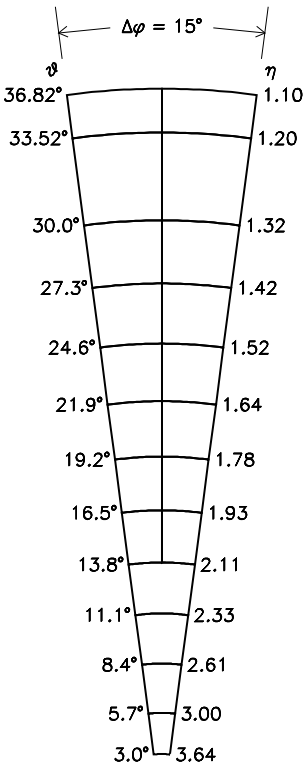

Figure 3.7: Segmentation of the towers in the PEM.

3.2.4 The Muon Chambers

Four systems of scintillators and drift tubes are used to detect muons with the CDF [36]. The central calorimeters act as hadron absorbers for the Central Muon Detection System (CMU). The CMU consists of four layers of drift chambers located outside the central hadronic calorimeter. Its range is $|\eta| < 0.6$ and can be reached by muons with transverse momenta greater than 1.4 GeV/c. Four additional layers of drift chambers are located behind a 0.6 m thick absorber layer of steel. This system is called Central Muon Upgrade (CMP). The CMP covers the same η range. In addition, the pseudorapidity range of $0.6 < |\eta| < 1.0$ is covered by the Central Muon Extension (CMX). These systems have been already used in Run I, however, new chambers have been added to the CMP and CMX in order to close gaps in the azimuthal coverage. The Run I forward muon system has been replaced by the Intermediate Muon System (IMU) covering a range of $1.0 < |\eta| < 1.5$. Table 3.2 gives an overview of the different muon systems in Run II.

3.2.5 The CDF Trigger System

The trigger plays an important role to efficiently extract the most interesting physics events from the large number of minimum bias events, because the collision rate is equal to the mean crossing rate of 1.7 MHz while the tape writing speed is about 75 Hz at present. The CDF trigger is a three level system with each level providing a sufficient rate reduction for the processing of the next level, shown in figure 3.9 [37].

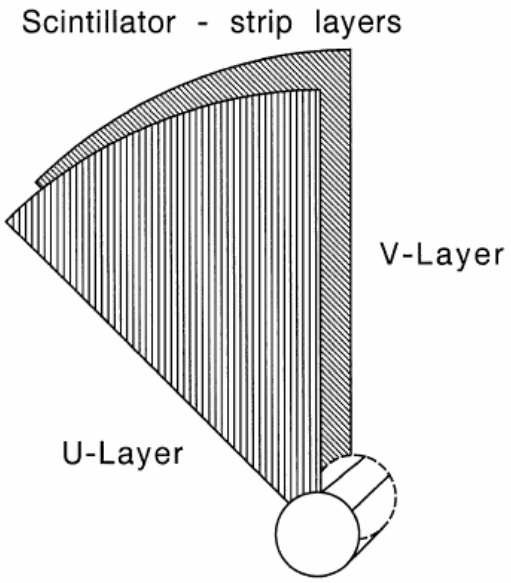

Figure 3.8: Layout of the PES detector. One can see the scintillator bars in u and v direction respectively forming an angle of 45 degrees.

	CMU	CMP	CMX	IMU
coverage	$\|\eta\| < 0.6$	$\|\eta\| < 0.6$	$0.6 < \|\eta\| < 1.0$	$1.0 < \|\eta\| < 1.5$
drift tubes	2304	1076	2208	1728
counters		269	324	864
min P_T	1.4 GeV/c	2.2 GeV/c	1.4 GeV/c	1.4 - 2.0 GeV/c

Table 3.2: Design parameters of the CDF II muon detectors.

The first two triggers are hardware triggers, the block diagram is shown in figure 3.10. The last step is a software trigger running on a Linux PC farm. Level-1 uses custom designed hardware to find physics objects based on a subset of the detector. The hardware consists of three parallel synchronous processing streams: one to identify calorimeter-based objects, another one to identify muons while the third one does tracking in the COT using the *eXtremly fast tracker* (XFT). The decision is done by simple counting these objects (e.g. one electron with 12 GeV). If an event is accepted by the Level-1 trigger, the data are moved to one of the four on-board Level-2 buffers, to average out the rate fluctuations. The typical rate of the Level-1 triggers is at present 24 kHz accept rate.

The Level-2 trigger do a limited event reconstruction using a custom-designed hardware. The hardware consists of several asynchronous subsystems, e.g the hardware cluster finder using calorimeter information. In addition, data from the shower maximum detector (CES) can be used to improve the identification of electrons and photons. The most challenging addition for the Level-2 trigger is the *Silicon Vertex*

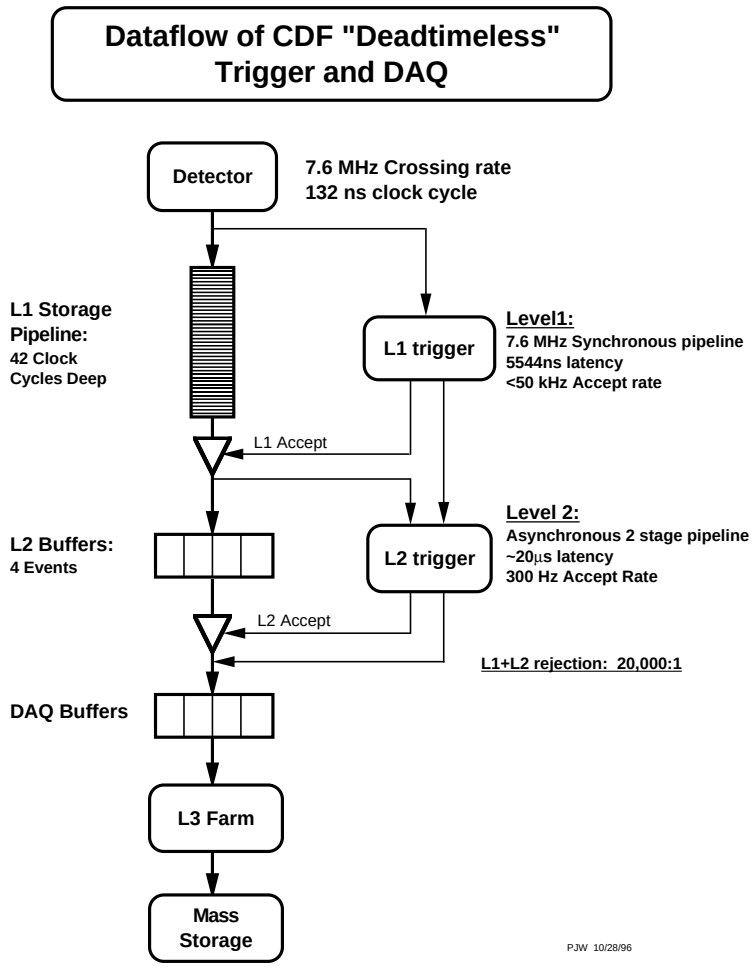

Figure 3.9: Block diagram of the CDF II data flow. The indicated crossing rate of 7.6 MHz is the maximum crossing rate and corresponds to the clock cycle of 132 ns. However, bunches are separated in space, so not at all clock cycle, a collision occurs. The bunch distance is 396 ns, resulting in a bunch crossing rate of 2.5 MHz. As not all bunches are filled to leave space to abort the beam, the mean crossing rate is reduced to 1.7 MHz. The trigger must, however, be able to handle a crossing rate of 7.6 MHz.

Tracker [38]. The SVT allows to select tracks with large impact parameter, which opens a complete new window for physics measurements at a hadron collider. The level-2 trigger accepts 300 events per second, which are transferred to the Level-3 processor farm [39].

At the processor farm the events are reconstructed and filtered, using the algorithms run in the "offline" reconstruction, and are written to permanent storage with approximately 75 Hz at present. To facilitate the handling of the huge data volumes collected with the CDF, events passing the Level-3 trigger are split into eight different streams. The triggers an event has passed decide to which stream this event belongs e.g. all events passing any of the highly energetic lepton triggers end up in "stream B".

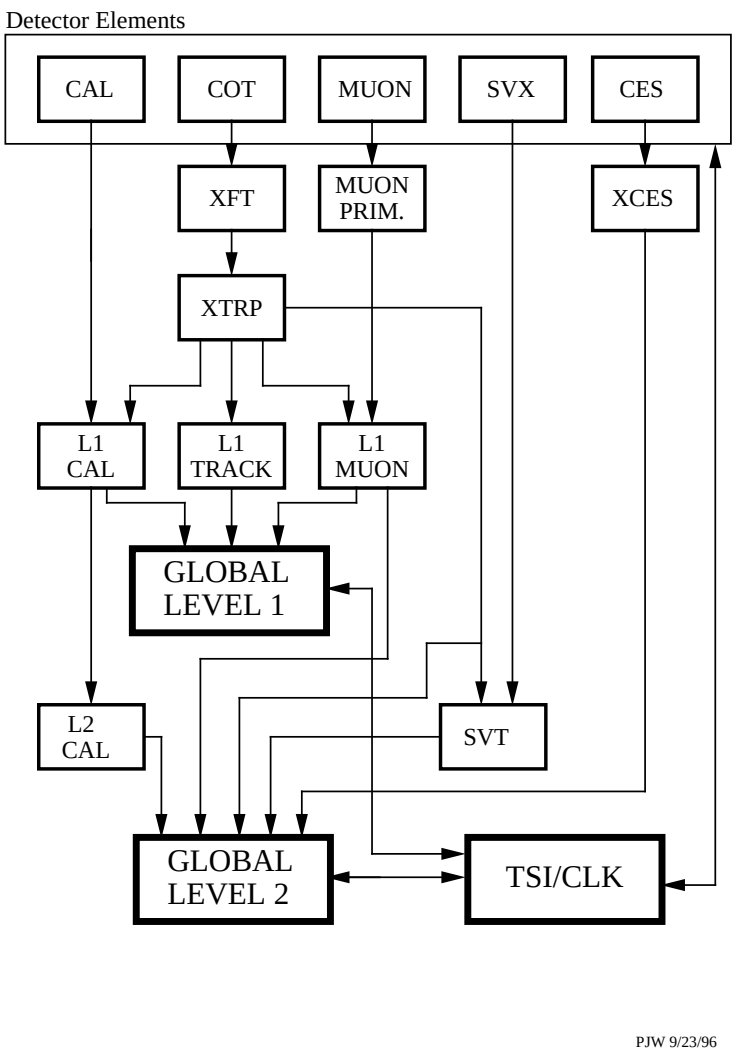

Figure 3.10: Block diagram of the CDF hardware trigger system in Run II.

In figure 3.11 the Event Display of a W boson candidate event with two jets is shown.

3.2.6 Online Monitoring of Data Taking

A complex multi-purpose detector, like the CDF, consists of many different detector systems. To take data with high efficiency and high quality it is necessary to quickly spot problems with one of the subdetectors. This can be achieved by monitoring the data during data taking. At CDF, all processes receiving data from the *Data AcQuisition* (DAQ) are called *consumers*.
For this purpose the so-called *Consumer Framework* [40] was developed based on the ROOT package [41]. A schematic view of the framework is shown in figure 3.12. The most important feature is that the part which displays the monitored results is

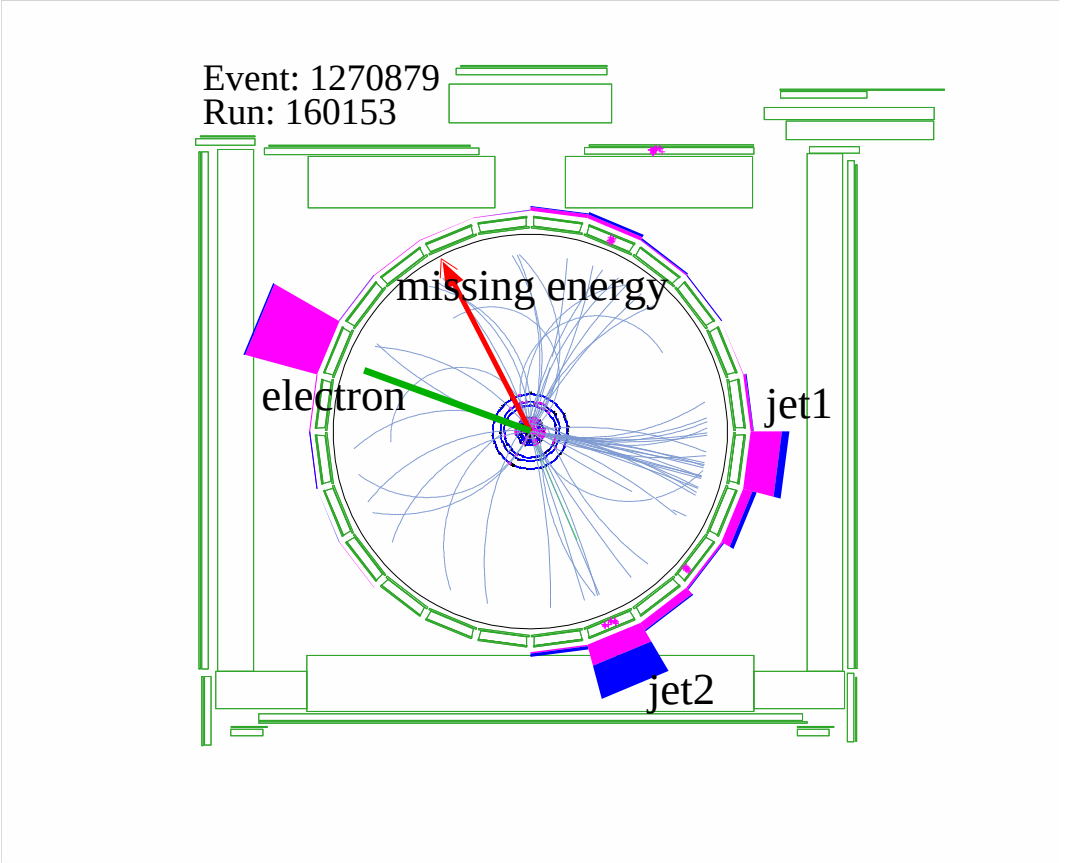

Figure 3.11: Event display of a W boson candidate with two jets. The energy of the electromagnetic calorimeter is drawn magenta, the energy of the hadronic calorimeter blue. The size of the cluster is proportional to the measured energy, in this case the highest cluster contains 97.64 GeV. As most energy is in the electromagnetic calorimeter and an isolated track is pointing to the energy deposition, this cluster is probably an electron. The vector sum of all cluster points to the lower right corner, therefore the $\not\!\!E_T$-vector points to the higher left corner, representing the neutrino from the W boson decay.

separate from the actual consumer programs.

The framework has three main components :

- **Consumers:** These are the modules which monitor and analyze objects in the event stream. They provide the connection to the rest of the CDF online framework.
- **Display Server / Display Viewer:** The Display Server is a ROOT-based program that allows the display viewer programs to connect to it as a client. Multiple display viewer programs can connect to one display server.
- **Error Handler:** This process receives the error messages from the different consumers and communicates with runcontrol so that appropriate action can be taken (e.g. reset an SVX CHIP).

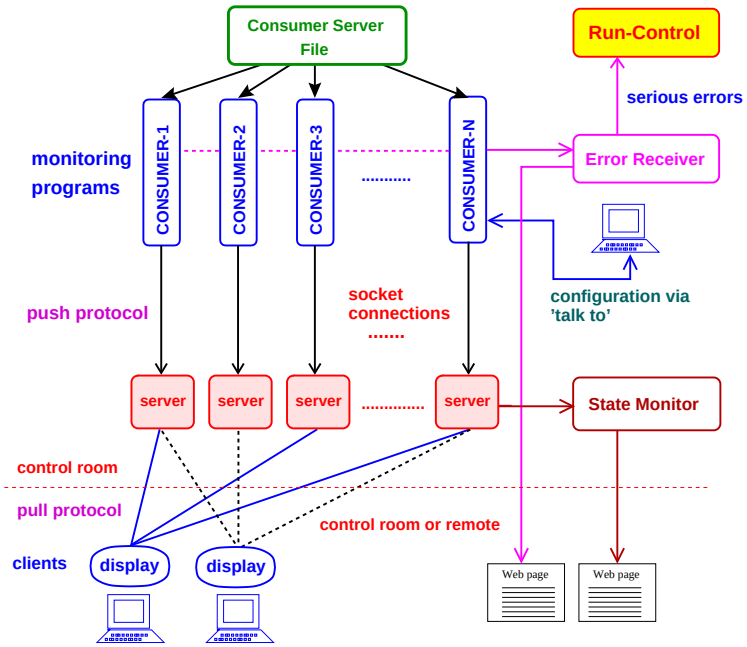

Figure 3.12: Overall design of the consumer framework

With these tools, it is possible to run the CDF with a minimum of operators. A typical shift crew consists of a scientific coordinator, two specially trained monitoring and DAQ operators and a consumer operator monitoring online the quality of the data. A technician takes care of the high voltage, cryogenics and gas systems. The operations manager ensures the continuity over the different shifts and coordinates all changes to the detector. In case of major problems, the shift crew can contact specific experts for each detector component.

Chapter 4

The EKPplus Cluster

The data samples used in CDF analyses easily reach several TByte in size. This analysis uses, among others, the forward electron data sample, which is 2 TByte large. The size of the data sets will even increase in next generation experiments at the Large Hadron Collider (LHC) at CERN.
The complexity of analyses also increases: novel analysis methods like neural networks allow a more efficient usage of the data, but also need a higher computing power than traditional methods. Therefore, for a long time already, the personal desktop has not been the place anymore where the main part of the analysis is done. A system with only few central, very powerful machines has shown to be by far not sufficient to serve the needs of a large community. At CDF, the original plans foresaw that all analysis should be done on one machine: the 128 processor machine `fcdfsgi2`. Rapidly, the resources were not sufficient anymore, an enlargement of the system was too expensive. Therefore, CDF has moved to the CAF concept: CDF Central Analysis Farms, based on PCs connected via Ethernet [42].
To cope with a larger amount of data, it is crucial to develop new methods of data processing, to test them and gain experience.

The LHC computing model, as sketched in figure 4.1, foresees a hierarchical model in which CERN forms the main center for raw data storage. However, due to the large amount of data and the large number of users, a second row of smaller computing centers is needed. One of these Tier-1 centers is located at the Forschungszentrum Karlsruhe. These centers replicate some data for a local user group. Still, these centers are not meant for daily users' work: in the LHC computing model, Tier-2 and Tier-3 centers will play these roles.
The Tier-2 and Tier-3 centers are located at universities and institutes, and will store the data needed for development of local analyses. They also provide sufficient computing power needed by the members of the institute, and aim for a short response time.
To be usable by everyone, these Tier-centers need to have a common hardware and software basis. Most programs are written in Fortran and C++, two languages that compile into machine-dependent code. When a user submits job (i.e. an executable) to the grid, he always expects the same CPU-architecture and operating

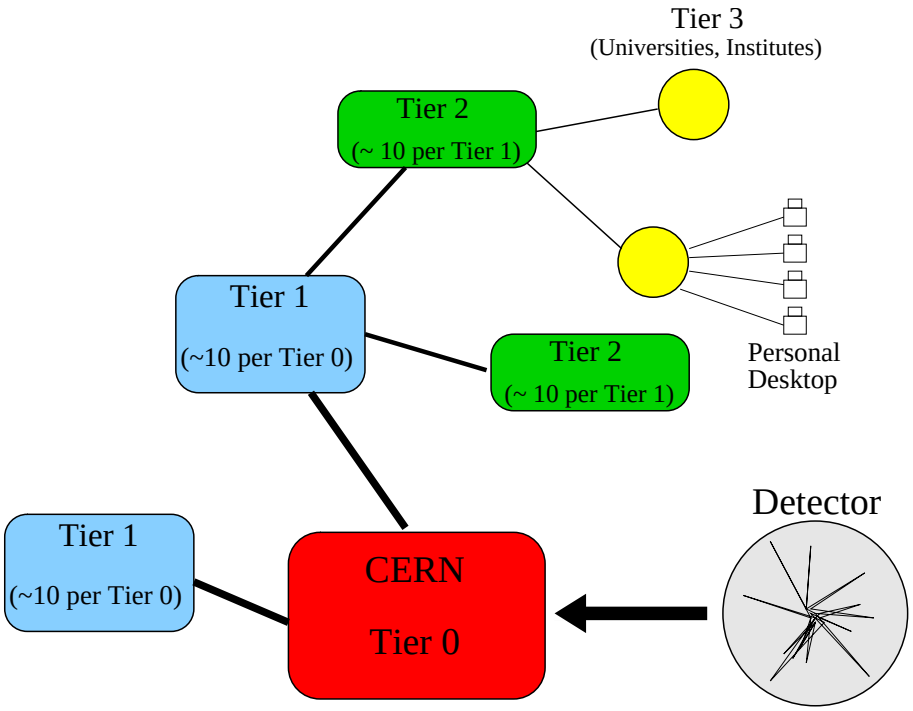

Figure 4.1: Sketch of the LHC computing model. The connections between Tier centers are not shown.

system. Switching to a language like Java, which compiles to a platform-independent bytecode language and is executed by a platform dependent interpreter, would be very time-consuming. The execution of the program would also take a longer time. Therefore, such a switch is not planned. Due to the complexity of the code, recompiling the programs on the target machine before execution is not feasible either. The solution is a homogeneous environment, so that the only remaining question is the choice of the hardware and operating system.
Since the Intel Pentium was launched, the performance gap between the x86 architecture and others like the Alpha architecture has become less important[1]. At the same time, the vendors of computers using a non-x86 architecture increased their prices, so that in experimental High Energy Physics, computers with x86 architecture have become the preferred platform.
In parallel with this hardware evolution, a free operating system was developed by Linus Torvalds and others, which was meant as "UNIX for x86". The name of this POSIX-compliant [43] operating system is "Linux". Linux became the operating system of choice for the x86 platform in High Energy Physics for the following reasons: its network abilities are good, the handling is familiar to the user from the large UNIX machines, it is freely available and good development tools like the GNU compiler collection [44] exist.

[1] x86 is used as a generic name for the processors based on the i386 by Intel. This also includes processors from other manufactures that use the same instruction code. The Alpha processor was developed by Digital Equipment Corporation and uses a different instruction set.

A single processor cannot provide the computing power necessary for a whole analysis. The traditional answer to this problem were large multiprocessor machines like `fcdfsgi2`. These machines are undeniably better suited for computational problems which can be parallelized but where the different threads need a large interprocess communication, like a weather simulation. A typical analysis in High Energy Physics has a totally different use-case: jobs can be trivially parallelized into separate jobs without any interprocess communication. This allows for cheaper solutions with the same performance: clusters of interconnected (x86-)computers.

4.1 The EKPplus Cluster

At the EKP, a cluster, "EKPplus", was built in 2001 by Dr. Patrick Schemitz [45]. It is a typical Tier-3 cluster and a prototype for a larger Tier-2 cluster. It consists of the following components:

4.1.1 Computing Nodes

Users can submit their analysis jobs to these nodes via a batch queuing system to these nodes. They are not meant for interactive work. They are usually single or dual processor machines. At the moment, there are 27 nodes totaling 34 processors, some of them are 64-bit processors. One hard drive per node of typically 100 GByte essentially provides temporary storage for user jobs. Each node is equipped with 1 GByte RAM per CPU.

4.1.2 Portal Nodes

Portal nodes hold the experiment specific software. They offer access to the users and allow them to submit their jobs to the batch queuing system. They are usually dual processor machines, with a larger amount of storage space. The storage must be fast, since the portals are used to compile the software, and it must be secure, as users hold their analysis code on them and the operation should be reliable. IDE-disks in a striping-mirroring array RAID-10 [46] have proven to be the solution that works best. The LHC experiment CMS group at the EKP has one portal while the larger CDF group has three.

4.1.3 Storage Nodes

Fileservers hold the largest amount of data and make it available to the computing and portal nodes via the Network File System (NFS) [47] or other protocols based on the Internet Protocol (IP) over fast- and Gbit-Ethernet. The file servers should be reliable with speed being less important. This is ensured when putting IDE or SATA disks in a RAID-5 array, which means that one disk can fail in the array without the data being lost. Users are not supposed to log in to these fileservers. At the moment, five fileservers are in use with a total capacity of 15 TByte.

4.1.4 Management Nodes

Control machines are necessary for managing user access and distributing user login directories. They also manage the queuing system, export the operating system of the nodes and control vital parameters like power supply and room temperature. If the room temperature exceeds 40°C or the temperature in one rack exceeds 45°C, they trigger an automated shutdown. This prevents damages by a failing air conditioning unit or failing rack ventilation.

4.1.5 Network Components

In order to connect all these systems, networking hardware is needed. A mixed fast- and gigabit-Ethernet switch based on copper wires has proven to be the best solution for a cluster of the size of the EKP grid cluster. These provide for a maximum bandwidth of 10 resp. 100 MByte/s. However, for a future enlargement, Ethernet could become a bottleneck, thus requiring costlier solutions like Infiniband [48], which would allow up to 30 Gbits/s of bandwidth.

4.1.6 Operating System

While the software on the portals is experiment dependent, the operating system on all the machines is based on the Linux kernel. The distributions are not identical on all the components, as some are better adapted to different tasks. The fileservers run Debian stable release, which is easy to maintain. The CDF portals run a distribution which is based on RedHat 7.3, but modified by the Fermilab computing division. The CMS portal as well as the computing nodes run a derivative of the RedHat Enterprise Server called Scientific Linux. Running different Linux distributions in the same cluster has not lead to major problems. Linux was chosen as it is the only operating system under which all the CDF, CMS and Grid software runs. The different distributions were chosen as they are the only certified platforms for the respective experiment software.

4.2 Operation Experience

The design and the construction of the EKPplus cluster are described in the PhD thesis of Dr. Patrick Schemitz, written in 2002. I have been in charge of consolidation and expanding the cluster as well as the daily administration and planning since that time. From this point of view, it might be useful to provide the reader with the experience gained since 2002 and the lessons learned from daily use.

4.2.1 Storage Issues

While the performance of the fileservers turned out to be as expected in the beginning, it degraded with time. The reading performance was bad, especially when

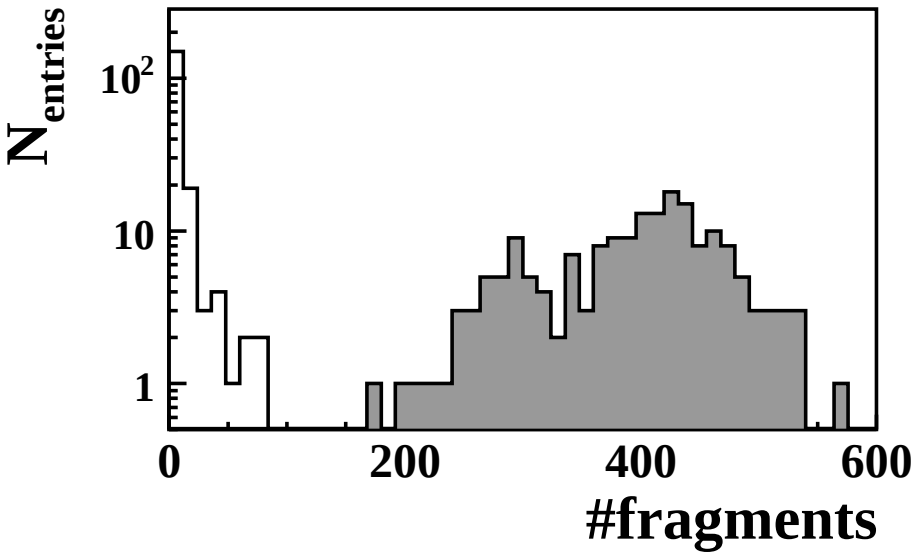

Figure 4.2: The axis of abscissae shows the number of fragments a file is fragmented in (i.e. its fragmentation). The axis of ordinates shows the occurrence of files with a certain fragmentation. The files written on an XFS filesystem (white area) are fragmented in mean into 10 fragments. The files written on an ext3 filesystem (gray area) are fragmented in mean into 392 fragments. All files are 500 MByte in size and written according to the described test setup.

many users were accessing the same device via NFS. The reason was long unknown: finally, however, it turned out that a very fragmented filesystem slowed down the read access. The reason for the very heavy fragmentation was the write pattern of the CDF data management tool SAM [49]: it writes data to disk in multiple streams, thus mixing blocks of the different files. The solution was to change the filesystem from ext3 [50] to XFS [51]. The latter is much less susceptible to fragmentation and, if fragmentation occurs, it can defragment files on-line. Another solution would be to rethink the write pattern of SAM.

Large files were fragmented in very small fragments, fragments of 12 kByte were not seldom, resulting in up to 10.000 fragments per file of one GByte on an ext3 filesystem. In order to be able to compare different filesystems, the author wrote a small script [52] which fragments files in an almost reproducible way by writing them in parallel streams to the disk and deleting randomly if free space is no more available. Using this tool, it was possible to demonstrate that files written to an XFS filesystem tend to fragment less than files written in an identical manner to an ext3 filesystem. During 24 hours, files were written to a 100 GByte-large filesystem formatted with XFS respectively ext3. About 25000 files of 500 MByte each could be written to XFS, only 8000 files were written to the ext3 filesystem. In both cases, 200 files remained and were checked for fragmentation. Figure 4.2 shows the number of fragments one file is splitted in. The files written to XFS had 10 fragments per file in mean, the files written to ext3 had 392 fragments per file in mean. Already

while writing the files, a lower writing rate was noticed when writing to a fragmented filesystem. Only one third of the writing rate was achieved for the ext3 filesystem compared to the XFS filesystem. The XFS filesystem has an online defragmentation tool, which rewrites files which are heavily fragmented. This utility was not used during our test. One could, however, expect to even lower the fragmentation of files in the XFS filesystem using this tool. There is no utility which could defragment a mounted ext3 volume.
One apprehension was that the network speed would not be sufficient for delivering the data to the nodes, but it turned out that the limiting factor was the speed of the disks and the RAID controllers respectively. The maximum internal speed observed was 110 MByte/s, which is also the maximum a Gbit Ethernet card can deliver to the net. With disks and controllers getting faster, the network could become a bottleneck. However, a drastic improvement is not to be expected, so a simple solution like channel bonding which would double the bandwidth could be sufficient to cope with the anticipated improvement of disk and controller speed.
As the number of fileservers increases, the bookkeeping of the different storage areas becomes more and more difficult for administrators and users when these areas are simply mounted via NFS. One solution could be a cluster RAID filesystem like Lustre or a unified data access system like dCache. Both would then show all the fileservers as one single volume to the user. dCache would have the advantage of load balancing: if files are solicited very often, they are automatically replicated to different fileservers.

4.2.2 Portal Issues

The internal storage area of the portals has been changed from RAID-0 (striping) to RAID-10 (mirroring+striping). It has shown that IDE disks are not reliable enough to run them as RAID-0. RAID-5 (distributed checksums) was not an option on the portals, as the area is used to compile and link, which asks for fast disk access.
The presence of three portals for the CDF experiment caused problems. Cross-mounting the home and scratch areas of the portals made a clean reboot after an uncontrolled shutdown almost impossible so that these areas are no longer cross-mounted. This has now led to the situation that users only reticently change the machine when one machine is in heavy use and others are not. They have to copy their code to the unoccupied machine, with the danger that different versions of code exist on different machines. There is no simple solution to this: users would have to change their habits and make more use of code management software like CVS [53]. Alternatively, a separate, common code server could be introduced with a very fast connection probably not relying on Ethernet but on more expensive techniques. Using such a technique, the home areas would be visible to all the portals and provide for a disk fast enough for compiling. An even costlier option would be the purchase of a machine with more than two processors.

4.2.3 Connection to the Desktop Cluster

In the beginning, the EKPplus cluster was completely separated from the desktop cluster. This meant doubling the user administration efforts. It was therefore decided to merge the user administration of the two clusters. A complete merger of the two clusters, however, is not imaginable at the moment. The desktop cluster is meant for local (those sitting in front of the desktop computer) users only, whereas the EKPplus is meant for "local" (those sitting in the physics building) and external users submitting jobs via grid. If the fileservers export their data to the desktop cluster, it must be ensured that external grid users still get a defined throughput. Simply exporting via NFS would not ensure this.
Integrating the desktop nodes into the queuing system available for external grid users is difficult for different reasons:

- The desktop PCs do not have access to the fileservers, at least not at this stage. Integrating the desktop PCs into the queuing system at this stage would make the cluster heterogeneous and more difficult to maintain.

- The hardware used in the desktop cluster is more fault-prone than the one used for the EKPplus cluster. There is also a larger variety of hardware employed. Administration as well as troubleshooting in case of problems would be made more difficult again.

- The software environment is different. While the desktop PCs offer newer software for multimedia or desktop applications, the Linux distribution on the EKPplus computing nodes has been put together with stability in mind. However, solutions like XEN [54] or VMware [55] could run a virtual Scientific Linux system on a host machine running a modern Linux distribution.

- As shown in section 4.2.1, integrating the storage into the desktop cluster is not easy. dCache [56] could solve these problems also for the integration of storage into the desktop cluster. The use of resilient dCache would ensure that the desktop computers do not use too much bandwidth of the main fileservers to the computing nodes.

- Desktop PCs by definition have physical users sitting in front of them. The work of the users must not suffer from jobs running on their machine. The scheduler must be configured such that jobs are submitted to the desktop cluster only at nighttime, or a job migrating mechanism must be found, which would, however, only work for light-weighted jobs.

- It must not be forgotten that desktop users can by mishap shut down a desktop PC with a job running on it. The unreliability of the desktop cluster must be taken into account when one plans its integration into the EKPplus cluster.

A well thought-out integration plan can increase the total computing power without affecting the local users. At the moment of writing this thesis, the prerequisites are, however, not given for such an integration.

4.2.4 Network Issues

The original logical network layout consisted of three subnets, `192.168.101.0/24`, `192.168.102.0/24` and `192.168.103.0/24` in the same collision domain. The reasons for this were that some mainboards did not cope with gigabit Ethernet cards, so instead three fast Ethernet cards were built in. Another reason was to perform some kind of traffic shaping through different subnets, two subnets would share the traffic for the fileserver and the third subnet would be dedicated for portal and user directories. The first point became obsolete with improved board BIOS and gigabit Ethernet cards. The traffic shaping point remained, but it turned out that, due to a design flaw in the network stack of the Linux kernel, a race condition occurs when different network interfaces configured to listen to different subnets are in the same collision domain. Therefore, the separation into different subnets was eliminated, which also simplified the administration of the cluster.

4.2.5 Security Issues

The EKP has no dedicated administrators, its computers are run mainly by PhD students. A lot of administrative tasks require superuser access to the machines. In the beginning, the software `sudo` [57] was used to give special privileges to different people. `sudo` enables users to perform predefined tasks as if they were superuser on the machine. It has shown that this software has some drawbacks: the user needs to have an unprivileged account on the machine, which is a problem for a firewall or a web server. The user cannot edit files, as all editors known to the author give the user the possibility to run any given program.
It became a common practice to give some users the root passwords to some machines. This practice is, of course, against all administrative rules that state that the number of people knowing the root password should be small. Two incidents convinced us to change this policy: one user was using root privileges to speed up the execution of his jobs. The other case was that people outside of the institute had the root password of one central machine.
At present, the policy is the following: every machine has its own password, which is created by a password generator. It is written on paper and stored at a secure place, but accessible to the main administrators. All people that need root access to a machine will have their ssh-key with password encryption stored in the authorized keys of the root account. For normal use, people can log in via their ssh-key. In case of network problems, the main administrators can access the machine via the console and the password looked up in the list. If a user no longer has administrative tasks, his keys can be removed easily.
This raises the security to a higher level without removing the ease of administration. A very precautious administrator will nevertheless criticize this scheme, as still too many people will work as root. The question is whether one wants to work in a high security environment. This would mean that users will no longer have physical access to the machines and to the network, and that no laptops will be allowed. To give an example, a simple attack with a laptop can reveal all the private data

from all EKP members within minutes. The risk of such an attack is small as it is assumed that no member of the institute wants to harm another member in such a way. However, with the enlargement of the institute, the problem of accessibility to its resources, also to third persons, should not be neglected. Also, with grid jobs running on the cluster, the number of people with potentially harmful intentions is increasing. Alternatives to a cluster based on Network Information Service (NIS) and NFS should be taken into account. An easy replacement of NFS could be Secure-NFS, NIS could be substituted by the Lightweight Directory Access Protocol (LDAP). The price of these measures could, however, be a drop in performance.

4.2.6 Architecture of CPU

In 2003, AMD launched the Opteron, a processor based on the 32-bit x86 architecture, but with a 64-bit extension, an integrated connection to other processors and memory controller (HyperTransport technology). In cooperation with AMD, the EKP tested a system equipped with two Opteron processors. There was the need to port some of the software used in High Energy Physics to this new architecture, but the effort needed for porting was surprisingly low. Some packages were even already ported, like the analysis framework ROOT. The detector simulation package GEANT, ZFITTER, a program to fit the Higgs mass to measurements of electroweak parameters were ported in a very short time. The CERN Program Library, a large collection of general purpose libraries and modules, was ported partially, the ZEBRA module making difficulties.
The benchmark results proved that some applications benefited from the 64-bit architecture [59]. Other applications could benefit from the very good connection of the CPU to the memory, even while running a 32-bit operation system. As regards performance and stability, the Opteron was at least comparable to 32-bit processors based on x86-architecture. For lack of a test system, no evaluation of other 64-bit processors could be made. However, third-party benchmarks show that for some 64-bit applications, especially FPU-demanding, other 64-bit processors like the Itanium perform better than the Opteron. The design of the Itanium lacks a full compatibility with the x86-architecture, and running 32-bit executables on the Itanium is very inefficient. Beside this, the cost/performance ratio of the Opteron as well as the thermal properties are better. The AMD64 is therefore the choice the EKP has made for future computing and portal nodes. Because of the excellent thermal properties, the AMD64 will replace older desktop PCs in the future, enabling very silent but still powerful desktop machines.

4.2.7 Cooling and Power

When building even a small-scale cluster like the one at the EKP, the cooling of the nodes and of the whole system must be thoroughly planned. Especially the air flow can cause a problem in small or inadequate rooms. Together with a suboptimal performance of the heat exchanger, cooling can limit the operation and scalability

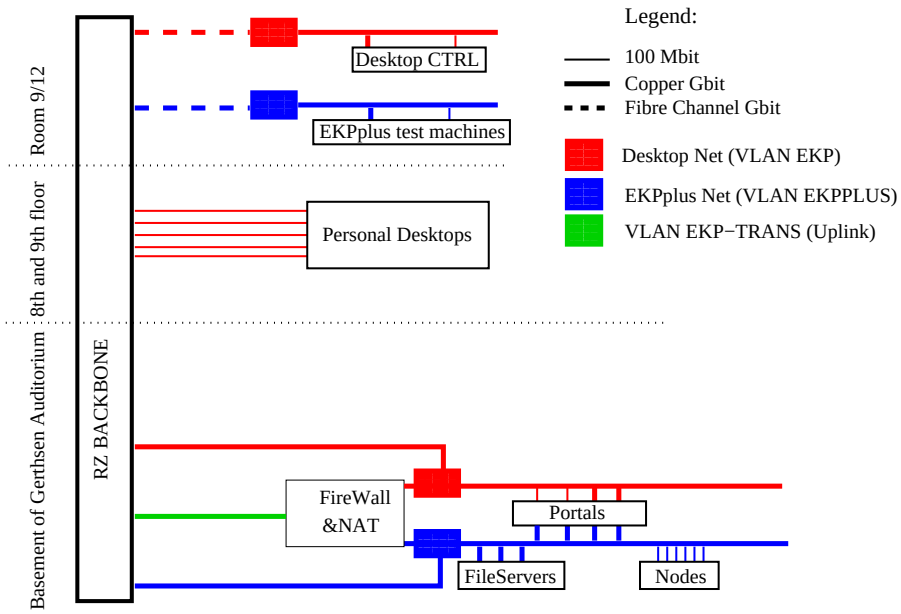

Figure 4.3: Schematic overview of the different computer categories and the collision domains of the EKPplus and EKP desktop cluster.

of a cluster. Of similar importance is a temperature monitoring system which shuts down the cluster in case of failure of the cooling system. Unfortunately, this occurred way too often in the computing room of the EKP. A solution was provided by the faculty which offered a room in the basement with a new ventilation and air conditioning. All machines of the EKPplus cluster were moved to this room, and only the machines necessary to run the desktop cluster stayed upstairs on the ninth floor. Whereas this is comprehensible for the computing nodes and the fileservers, the decision was not easy for the portals, the control machines and the central firewall. For performance reasons, these machines were also moved to the basement. The Virtual LAN (VLAN) enabled switches of the Universitätsrechenzentrum were used to connect the machines. Figure 4.3 shows the different categories of computers and the collision domains they are in and, in addition, the type of connection is detailed. Mostly the cabling is based on copper wires, some Fibre Channel (FC) optical links enable Gbit ethernet in the ninth floor. The additional VLAN EKPPLUS was introduced, enabling the administrators to run a machine which should be in the collision domain of the EKPplus cluster elsewhere than in the basement. This is useful when installing, testing or troubleshooting machines. The firewall inspects traffic between the inner net and the VLAN EKP-TRANS, which is the uplink to the Universitätsrechenzentrum. At the same time, the firewall serves as a Network Address Translation machine (NAT) [60], thus enabling the access of computing and storage nodes with private addresses to the Internet.

Electric power was not a limiting factor but should nevertheless be planned together with the whole cluster. If the power circuit of the power provider is not designed for uninterruptible power supply (UPS), one should use battery packs for

the most important machines to keep them running–when a short outage happens– and trigger an automated shutdown if the outage turns out to be longer than, in our case, two minutes. The machines powered by a UPS are the control machines, the fileservers and the portals. Connecting the nodes to the UPS was considered to be too expensive in comparison with the potential damage.

4.3 The Future of EKP Computing

High Energy Physics has already made heavy use of modern and novel computing techniques for a long time. Therefore, computing in High Energy Physics is a rapidly evolving field and is now of large importance. Especially the Tevatron experiments and the coming LHC experiments will set new milestones in computing.
Computing is also an important return from High Energy Physics to society: techniques which will change daily life in future times are developed and tested in High Energy Physics.
The EKP is well-positioned in this field with its own facilities which can be seen as a Tier-3 center in the LHC computing model. A strategical advantage is the proximity of and the connection to the German Tier-1 center GridKa at the Forschungszentrum Karlsruhe. The EKPplus project has been very successful in the past and will, hopefully, also be successful in the future.

Chapter 5

Analysis Prerequisites

In order to be able to perform any analysis on data taken by the CDF detector, this data has to be processed. The raw information coming from the detector must be interpreted in order to enable a reasonable analysis. For example, information about the charge deposition in the tracking detectors is useless unless a tracking algorithm combines and reduces this information to an object called "track", where the only relevant information for this analysis is the four-momentum, the charge of the particle track and the quality of the performed fit. The tracking algorithm is just one example of an algorithm used to interpret the raw information.
It is in the nature of a large collaboration that not everyone does everything by himself. However, it is of importance to know about the different steps made when transforming and reducing the information.
The same is true when using simulated events. The generators have been written by other people, but a good understanding of the different generators with respect to the physical processes studied is important.
This chapter will outline some technical prerequisites in order to give the reader the opportunity to better understand the analysis.

5.1 The CDF Software Framework

The CDF Software Framework is an application framework in the context of a HEP experiment. It allows physicists to develop code and combine it with code developed by other people. The framework is written in C++ to profit from the advantages of object-oriented programming and is called AC++ [61]. This software runs online during data-taking to make the first interpretation of the raw data. The primary vertex is found, the tracking is made, jets are reconstructed, electrons and muons as well as b quarks are identified. When using Monte Carlo generators, this software puts the output of these generators into a detector simulation out of which the physical objects are reconstructed.
Some physics groups like the b quark working group base their analysis on the data format written out by this software. This has the advantage that they have access to any detector information and that they can, if necessary, change the algorithms to identify objects like tracking, for example. The drawback is, however, the large size

of these files. It has shown that such detailed information is not necessary for most analyses in the top physics group, therefore a lossy compression of data is performed during the conversion with the TopEventModule.

5.2 Track Reconstruction

Using information from the tracking detectors, particle trajectories can be reconstructed. Inside the solenoid, charged particles travel on a helix with its axis parallel to the magnetic field. Five parameters describe this helix [62]. These parameters are defined with respect to the point of minimum approach to the origin, the perigee.

- $\cot\theta$: the cotangent of the polar angle at the perigee
- C : the half-curvature (same sign as the charge of the particle)
- z_0 : the z position at the perigee
- d_0 : the signed impact parameter; the distance between the helix and the perigee
- ϕ_0 : direction of the track at the perigee

5.2.1 Tracking in the Central Outer Tracker

In a first step, tracks in the Central Outer Tracker (COT) are reconstructed. The drift chamber is the tracking detector with the largest distance from the beam axis. Due to the fact that its occupancy is lower and that the tracks are more isolated, the reconstruction is easier for this detector in comparison to the silicon detectors. There are two different algorithms in use to reconstruct tracks in the COT. One algorithm is based on the code used in Run I to reconstruct tracks in the Central Tracking Chamber (CTC, now replaced by the COT) [63]. In this approach, segments are reconstructed in the super-layers. These segments are then linked together to reconstruct the trajectory.
The second algorithm [64] uses one segment in the outer super-layers and the expected beamline to construct a reference track. The distances of the hits in the other super-layers from this reference are filled into a histogram. This histogram is used to determine the track parameters. This involves that the tracks are already beam-constrained, which improves the momentum resolution. However, the exact position of the beamline is not known when the reconstruction is done and the tracks reconstructed by this algorithm have a bias towards the assumed beam position used in the construction of the reference tracks.

5.2.2 Silicon Tracking

There are three different approaches to reconstruct tracks in the silicon system: $outside-in$, $inside-out$ and $stand-alone$ tracking. The $outside-in$ tracking

algorithm propagates a track found in the COT into the silicon system and tries to add hits to the track. After a hit has been added, the track parameters are recalculated using this additional information. In the CDF software, there are two implementations of this algorithm. One is based on the Run I code and uses a progressive fitter [65]. The other uses a Kalman fitter, which is the optimal fitter for this task, since it naturally takes dE/dx and multiple scattering effects into account. This fitter and the algorithms based on it have been developed at the Institut für Experimentelle Kernphysik in Karlsruhe [66].
The $stand-alone$ tracking algorithm is based as well on this Kalman fitter. The COT does not cover the forward and backward regions ($|\eta| > 1.1$). Thus, only the information of the silicon detectors can be used to find tracks up to $|\eta| < 2.0$. This is the task of the $stand-alone$ algorithm. To reduce combinatorics, the algorithm uses only hits not used by the two $outside-in$ strategies. The position of the beamline is needed for the construction of the track candidates causing a small bias towards the assumed beam position.
The $inside-out$ tracking algorithm uses silicon $stand-alone$ tracks to define a search road for hits in the COT detector. The hits in the road form a COT track that is fitted using the silicon track information as constraints. The silicon hits are finally refitted using the new COT track as a seed.

5.3 Primary Vertex Reconstruction

Many analyses like life time measurements and analyses which need a b tag require the precise measurement of the primary vertex position for every event. The primary vertex is the point from which all prompt tracks originate. In many applications, the position of the beamline can be used to estimate the primary vertex position in x and in y, if the z coordinate is known. This method is limited by the size of the collision region, the beam width, but proved to be sufficient for most applications in b physics. For events with high multiplicity (e.g. $t\bar{t}$) the primary vertex can be found with a better precision than the beam width. To achieve this goal Vxprim [67] was developed. The Vxprim program fits the primary vertex using reconstructed tracks. Vxprim is used to fit the beamline on a run by run basis [68].

The Vxprim algorithm is run on production level. The results of this algorithm are used to determine the "beamline" positions [69]. The beamline is defined by the locus of all reconstructed primary vertices. Thus, the beamline is the position of the luminous region. The Vxprim algorithm uses all tracks fulfilling certain quality requirements. A track is accepted if, for example, at least two stereo and two axial super-layers with at least six COT hits each have been assigned to this COT track. Silicon tracks reconstructed by an outside-in algorithm are required to have at least four $r-\varphi$ hits. In a first step all tracks passing the quality cuts are fitted to a common vertex. In an iterative "pruning" process every track is removed from the vertex fit and a χ^2 of this track with respect to the vertex is calculated. If the highest χ^2-value of these tracks exceeds a certain threshold, the track is removed. Then the vertex fit is repeated using the remaining tracks and this pruning procedure

is continued until all tracks pass the χ^2 cut. The vertex is accepted if a minimum number of tracks is assigned to the vertex.

5.4 Electron Reconstruction

High momentum electrons leave isolated energy deposits in adjacent towers in the electromagnetic calorimeters. These towers can be identified and merged into one electromagnetic cluster. Electrons are then identified in the central electromagnetic calorimeter (CEM) as isolated clusters which match an XFT track in the pseudorapidity range $|\eta| < 1.1$. The corresponding energy deposition in the hadronic calorimeter should be low. The electron hardware trigger requires the assigned XFT track to exceed a transverse momentum of 8 GeV and an electromagnetic transverse energy of the cluster $E_T > 16$ GeV. The ratio of energy depositions in the hadronic and the electromagnetic calorimeter has to be less than 0.125.

5.5 Muon Reconstruction

Muon candidates are identified as isolated tracks which can be extrapolated to muon stubs. Muon stubs are reconstructed track segments in one of the four-layer stacks of the muon chambers (CMX,CMU,CMP). The muon hardware trigger requires an XFT track with $p_T > 8$ GeV matched to such a track segment or stub in the joint CMUP configuration or in the CMX.

5.6 Jet Reconstruction

The hadronization of a final state quark creates a jet of hadrons. Hadronization describes the transition from colored partons to color neutral objects. These particles then form a particle jet. The energy of the hadrons is measured in the calorimeters. The momentum of the initial quark can be reconstructed by combining the energy measurements in the calorimeter towers that belong to the jet. Figure 5.1 illustrates the transition of fundamental particles to calorimeter jets. The out of cone particles correspond to particles originating from the parton but their energy deposit is not assigned to the calorimeter jet.

Three different algorithms to reconstruct jets are implemented in the CDF software: JetClu, Midpoint and K_T-Clustering algorithm. The latter two are seldom used, the reader is referred to [70] for a description.
The most used algorithm, the JetClu algorithm, was the standard algorithm in Run I. Thus, its systematics are very well understood. First, this algorithm selects a seed tower. Then it draws a cone around this tower with a fixed radius in the $\eta - \varphi$ plane. All calorimeter towers inside this cone are combined to form the jet. The axis of this jet is used as the new direction of the cone axis in the next iteration of this algorithm. If the jet axis stays stable, the reconstruction of this jet is finished. Seed towers are all calorimeter towers with a measured energy above a certain threshold.

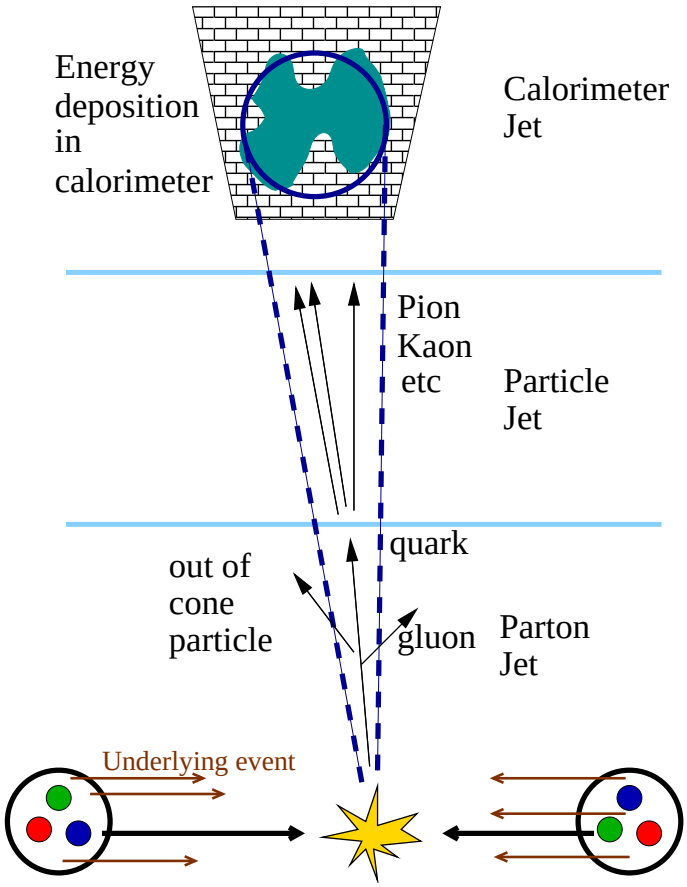

Figure 5.1: An illustration of the transition from partons to calorimeter jets.

Although this algorithm works very well in the dense environment of hadron-hadron collisions, it has two problems when it is applied to partons in order to derive theoretical predictions:

- A single parton with energy above threshold will serve as a seed. But if the momentum of this parton is distributed among two partons each having half the energy, both might fail the threshold cut. This is called the collinear problem.
- If there are two high energetic partons that have a distance in the $\eta - \varphi$ plane that exceeds the cone size, two jets will be formed by the algorithm. A gluon emitted by one of the partons might move the jet axis in a way that now both partons and the gluon form just one jet. This is called the infrared problem.

5.7 Jet Energy Corrections

The primary goal of the jet energy corrections group is to determine the energy correction to scale the measured energy of the jet back to the energy of the final state particle level jet [71]. Additionally, there are corrections to associate the measured jet energy to the parent parton energy, so that direct comparison to the theory can be made. Currently, the jet energy scale is the major source of uncertainty in the top quark mass measurement and inclusive jet cross section.
The CDF jet energy corrections are divided into different levels to account for different effects that can distort the measured jet energy, such as response of the calorimeter to different particles, non-linearity response of the calorimeter to the particle energies, uninstrumented regions of the detector, spectator interactions, and energy radiated outside the jet cone. Depending on the physics analyses, a subset of these corrections can be applied.

5.7.1 Relative Scale Corrections

The central calorimeter is best calibrated and understood, a relative correction is applied to jets in the forward calorimeters. This correction is obtained using Pythia and data di-jet events. The transverse energy of the two jets in a $2 \rightarrow 2$ process should be equal. This property is used to scale jets outside the $0.2 < |\eta| < 0.6$ region to jets inside the region. This region is chosen since it is far away from the cracks or non-instrumented regions. This results in a correction as a function of pseudorapidity and P_T. After corrections, the response of the calorimeter is almost flat with respect to pseudorapidity, as can be seen in figure 5.2.

5.7.2 Correction for Multiple Interactions

The energy from different $p\bar{p}$ interactions during the same bunch crossing falls inside the jet cluster, increasing the energy of the measured jet. This correction subtracts this contribution on average. The correction is derived from minimum bias data and

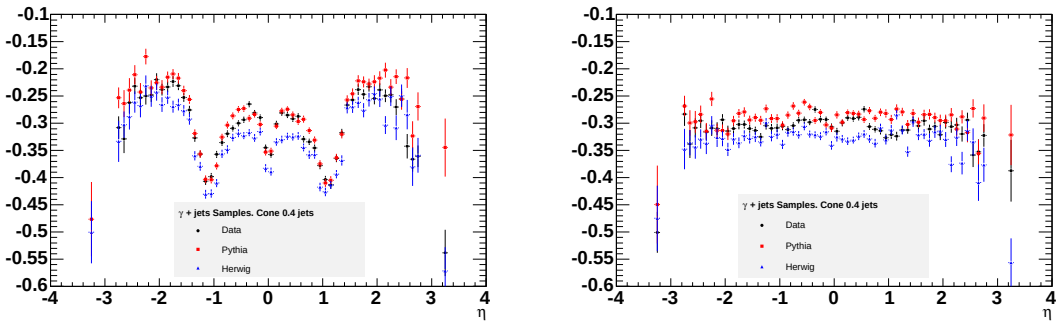

Figure 5.2: The η-dependence of the relative calorimeter response in arbitrary units before (left) and after applying the relative corrections (right).

it is parametrized as a function of the number of vertices in the event. The amount of the corrected energy can be seen in figure 5.3.

5.7.3 Absolute Scale Corrections

The jet energy measured in the calorimeter needs to be corrected for any non-linearity and energy loss in the uninstrumented regions of each calorimeter. Since there are no high statistics calibration processes at high E_T, this correction is extracted from Monte Carlo. The simulation of the calorimeter needs to accurately describe the response to single particles (pions, protons, neutrons, etc). The fragmentation in Monte Carlo events needs to describe the particle spectra and densities of the data for all jet E_T. The fragmentation and single particle response is measured in data and the Monte Carlo simulation tuned to describe it. The correction is obtained by mapping the total P_T of the hadron-level jet to the P_T of the calorimeter-level jet. The hadron-level jet consists of particles within a cone of the same size as and within $\Delta R < 0.4$ of the calorimeter-level jet. The correction factor as a function of P_T can be seen in figure 5.4.

5.7.4 Correction for Underlying Event

The underlying event is defined as the energy associated with the spectator partons in a hard collision event. Depending on the details of the particular analysis, this energy needs to be subtracted from the particle-level jet energy.

5.7.5 Out-of-Cone Correction

The out-of-cone correction corrects the particle-level energy for leakage of radiation outside the clustering cone used for jet definition, taking the "jet energy"back to "parent parton energy". The correction is derived from measurements of the energy flow between cones of size 0.4 and 1.3. The correction factor for the jets used in this analysis is shown in figure 5.5.

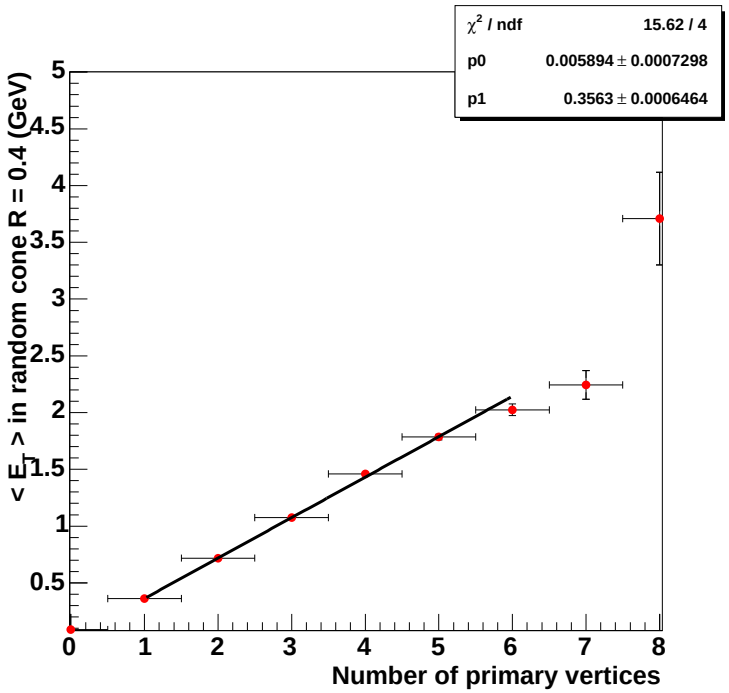

Figure 5.3: Average correction for multiple interactions as a function of the number of primary vertices.

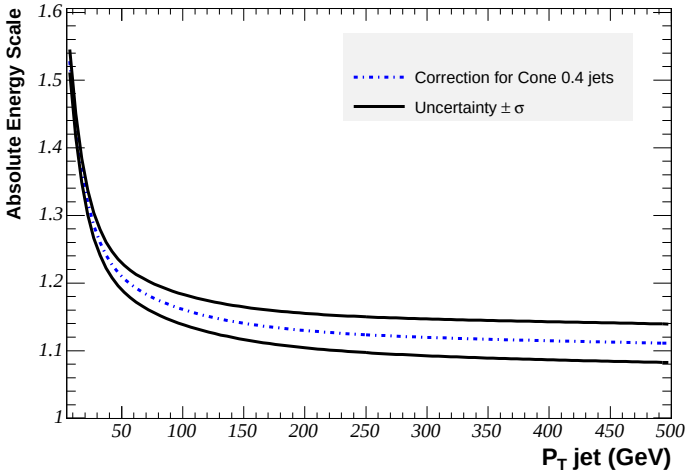

Figure 5.4: Correction factor for absolute energy scale as a function of jet p_T.

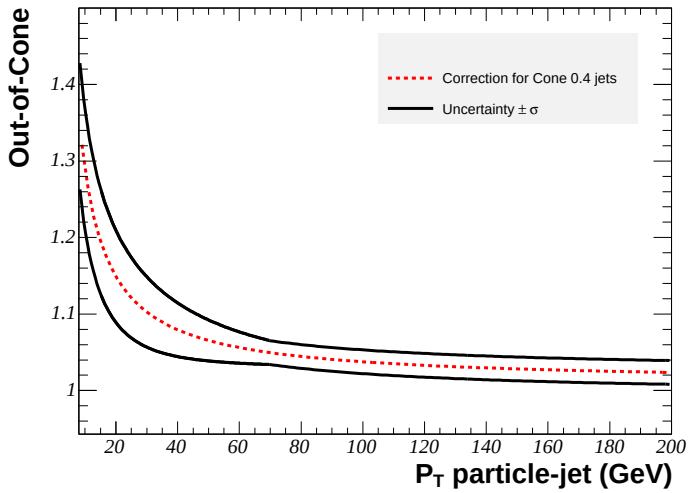

Figure 5.5: Out-of-Cone corrections for cone 0.4 jets.

5.7.6 Application of the Jet Corrections

Depending on the analysis, the corrections are applied to some level:

0. No corrections
1. Relative energy corrections
2. Previous + time-dependent corrections (Not existent in version 5 of CDFsoft)
3. Previous + energy-scale corrections (Not existent in version 5 of CDFsoft)
4. Previous + multiple interaction energy corrections
5. Previous + absolute energy corrections
6. Previous + underlying event corrections
7. Previous + out-of-cone corrections

In an analysis which would need a good reconstruction of the final state partons, one would correct up to level 7. For other analyses like the search for electroweak top quark production, a correction up to level 4 is deemed sufficient. The determination of the systematic uncertainties is easier for a lower correction level.

For the computation of $\not\!\!E_T$ in the analysis detailed later, jet corrections up to level 4 are applied. However, because of technical reasons, the raw $\not\!\!E_T$ corrected this way is not always accurate: electrons are also contained in the jet list at first. When tagging electrons, the corresponding jet is removed from the jet list, the corrected $\not\!\!E_T$ is computed thereafter. This analysis, however, uses another definition of the electron criteria in the plug region, so for some events the corresponding jet is not removed before the correction is applied, hence resulting in an overcorrection of $\not\!\!E_T$. This is accounted for by recorrecting $\not\!\!E_T$ after the electron identification.

5.8 The Identification of Bottom Jets

In many physics analyses, it is crucial to know the flavor of a quark producing the jet to extract a signal. It is possible to discriminate jets originated by a bottom quark from jets originated by lighter quarks or gluons. Due to the relatively large mass of the bottom quark, the bottom hadron carries most of the momentum of the original quark. The hadron is boosted and, due to its lifetime of approximately 1.5 ps, it travels a sizable distance before it decays.
The algorithm mostly in use at the CDF is called *SecVtx*. This algorithm searches for a secondary vertex directly. SecVtx is essentially unchanged from Run I, only the track selection has been retuned to match the improved CDF II detector. A detailed description of the algorithm can be found in reference [72].

The first step of the algorithm is to identify the primary vertex of the event. If a high momentum lepton is identified in the event, the vertex with the smallest distance to the lepton is used. In the absence of such a lepton, the vertex with the highest total scalar sum of transverse momentum of associated tracks is used. The position of the primary vertex is then refitted by using all the tracks that are found within a window of ±1cm around the z-position of this vertex and fulfilling the requirement to have an impact parameter significance $|d_0/\sigma_{d_0}| < 3$ relative to the beamline. In a pruning process all those used tracks are removed which contribute a $\chi^2 > 10$ to the fit. If no tracks survive the beamline profile is used for the primary vertex position estimate. The next step of the algorithm is the actual reconstruction of the secondary vertex. Since the algorithm operates on a per-jet basis, the tracks within the jet cone are considered for each jet in the event. All tracks not passing quality cuts on the number of silicon hits assigned to the track, the quality of those hits and the χ^2-value of the track fit are rejected. Only jets to which at least two such good tracks have been assigned are called "taggable". Based on the impact parameter significance with respect to the primary vertex displaced tracks are then selected and serve as input for the algorithm. *SecVtx* uses a two-pass approach to find displaced vertices. In the first pass the algorithm uses all tracks with $P_{\mathrm{T}} > 0.5$ GeV/c and $|d_0/\sigma_{d_0}| > 2.5$. In this pass at least three tracks are required to form a secondary vertex. If this first pass fails, the track requirements are tightened ($P_{\mathrm{T}} > 1$ GeV/c and $|d_0/\sigma_{d_0}| > 3$), but also two-track vertices are accepted. Once a displaced vertex is found in a jet, certain criteria are applied to the vertex to enrich vertices originating from b and c hadron decays. One requirement is for example that $L_{2d}/\sigma_{L_{2d}} > 3$. Here L_{2d} denotes the two-dimensional decay length of the secondary vertex that is calculated as a projection onto the jet axis of the vector pointing from the primary to the secondary vertex in the $r-\phi$ view only. The algorithm parameters can be tuned such that a tight and a loose SecVtx tag can be deduced.
The left plot in figure 5.6 shows the efficiency to tag a fiducial b jet in simulated $t\bar{t}$ events in the central region of the tracker depending on the transverse jet energy. The right plot in this figure denotes the pseudorapidity dependence of the tagging efficiency of jets with $E_{\mathrm{T,jet}} > 15\ GeV$. This efficiency is $\approx 45\%$ for the tight SecVtx tag and $\approx 50\%$ for the loose SecVtx tag in the central region for $t\bar{t}$ events. As the

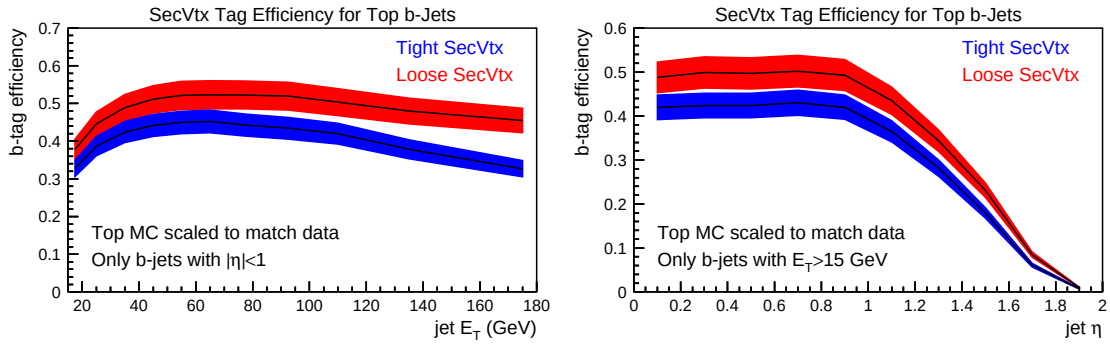

Figure 5.6: Efficiency to tag a b jet as function of the transverse jet energy for $t\bar{t}$ events.

efficiency is not identical in simulation and in data, a scale factor is applied on the tag rate in simulated events of 0.909 ± 0.06 for the tight tagger and 0.927 ± 0.066 for the loose tagger. This scale factor accounts for the fact that in simulated events the quality of the tracks is overestimated.

Unfortunately not only secondary vertices originating from heavy quark decays are identified but also so-called mistags. These mistags correspond to wrongly assigned vertices fulfilling all required vertex quality criteria. Sources for mistags are light flavor jets, where by accident a random combination of tracks form a secondary vertex. Figure 5.7 shows the mistag efficiency for the algorithm as a function of $E_{\mathrm{T,jet}}$ and η_{jet}. Although these numbers are quite small (0.1-0.4%) compared to the tagging efficiency, the reader should keep in mind that the production rate of light flavor jets is much higher than the one for heavy flavor jets. Thus, mistags are a significant source of background events for any analysis using this tagger.

In order to estimate the kinematic properties of such mistags, a mistag matrix has been developed within the CDF collaboration that provides a mistagging probability depending on the number of tracks assigned to the jet, $E_{\mathrm{T,jet}}$, η_{jet} and the azimuthal angle ϕ_{jet}. In addition, the scalar sum of the transverse momentum of all taggable jets is considered.

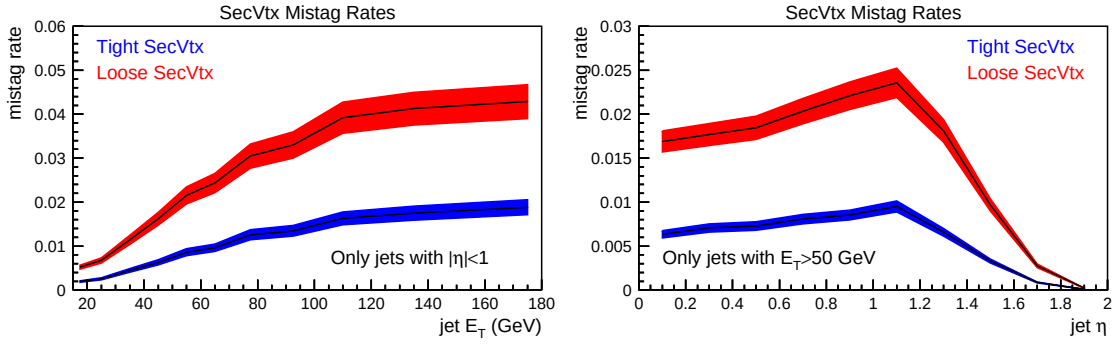

Figure 5.7: Efficiency to misidentify a heavy flavor jet as function of the transverse jet energy and the jet pseudorapidity.

5.9 Detector Simulation

The understanding of efficiencies, acceptances and the kinematic properties of signal and background processes requires a deep knowledge of the physics processes and the detector response. Therefore, Monte Carlo generators are used that randomly generate hard parton interactions according to the probability density of phase space. The resulting partons are then processed by a parton showering to simulate gluon radiation and fragmentation. The resulting particles are then handed to the detector simulation. The detector response is modeled on a detailed simulation with the GEANT3 [73] package. Most of the time, the standard GEANT algorithms are employed. To speed up the simulation, the charged particle ionization and drift properties in the COT are parametrized and also tuned to data. The development of showers in the various calorimeters is simulated by GFLASH [74], a shower development package. The GFLASH parameters for electromagnetic and hadronic showers are tuned to data. A detailed description of the CDF II simulation can be found elsewhere [75]. The simulation is an integral part of the CDF software framework. In order to further analyze the simulated data, it is subject to the same reconstruction algorithms which perform tracking or do jet clustering. This step is called "production". Now, the format of the simulated data is, with the exception of the HEPG bank containing the information from the Monte Carlo generator, identical to the data format EDM.

5.10 The TopNtuple Format

Not all people working in the top physics group need all the detailed information contained in the EDM format. Therefore, a smaller ntuple is created. Before this ntuple is written some of the high-level objects are remade with latest calibrations. This "remake" is done by the executable `TopFind` which also runs the TopEventModule. In this module the events are classified: the lepton identification is run, Z^0 and cosmic muon tagger are applied. This additional information is then added to a copy of the original EDM files. In parallel, this information is filled into a ROOT tree with several branches for different objects. This format is called the TopNtuple format [76]. Table 5.1 gives an overview of the most important classes (objects) in the TopNtuple, which is the format used by most of the top group.

The top physics group provides an official macro which can be used as a starting point for analyzing TopNtuples. The Karlsruhe top physics group has developed a small framework which provides additional object-oriented features as well as ready-made physics tools like jet corrections on TopNtuples. Most of the code written in the context of this analysis used the Karlsruhe framework for analyzing TopNtuples. All data and Monte Carlo samples used in this analysis are in the TopNtuple format.

Class name	Short explanation
evt	Summary of event and run information
summary	Summary of physical objects in the TopNtuple
privertex	Primary vertex
zvtxs	Z-vertices
secvtxtrack	Secondary Vertex Track
jetprobtrack	JetProb information: used for tagging b quarks
trigInfo	Trigger bits passed by the event
trigName	Names of these trigger bits
missingEt	Missing transverse energy in the event
hepg	Information from the Monte Carlo generator
obsp	Links from the tracks to the hepg information
electron	Properties of the electrons in the events
muon	Properties of the muons in the events
tau	Properties of the tau leptons in the events
jet	Properties of the jets in the events, different jet algorithms
offltrack	Properties of the tracks

Table 5.1: A selection of the most important classes in the TopNtuple tree.

5.11 NeuroBayes

The neural network used for the electron identification is the NeuroBayes package [77] provided by the company phi-t. NeuroBayes combines a three-layer feed forward neural network as seen in figure 5.8 with a complex robust preprocessing of the input variables. There is one input node for each input variable plus one bias node. The number of nodes in the hidden layer can be freely chosen by the user. There is one output node which gives a continuous output in the interval [-1,1].

The nodes of two consecutive layers are connected with variable weights. For each node i, a biased weighted sum of the values of the previous layer x_i is calculated

$$a_j(\mathbf{x}) = \sum_i \omega_{i,j} x_i + \mu_{0,j} \tag{5.1}$$

and passed to the transfer function which gives the output of the node. The bias $\mu_{0,j}$ (which is calculated for each input) implements the thresholds of the several nodes: if the input to a node is larger than its threshold, the node will send an input to the next layer. The output of each node is determined by a transformed sigmoid function

$$S(\mathbf{x}) = \frac{2}{1 + e^{-a(\mathbf{x})}} - 1 \tag{5.2}$$

which gives an output of -1 for background and +1 for signal. As can be seen in figure 5.9, the sigmoid function is only sensitive to a relatively small range around zero. Outliers in the original distribution are mapped by this transformation to the

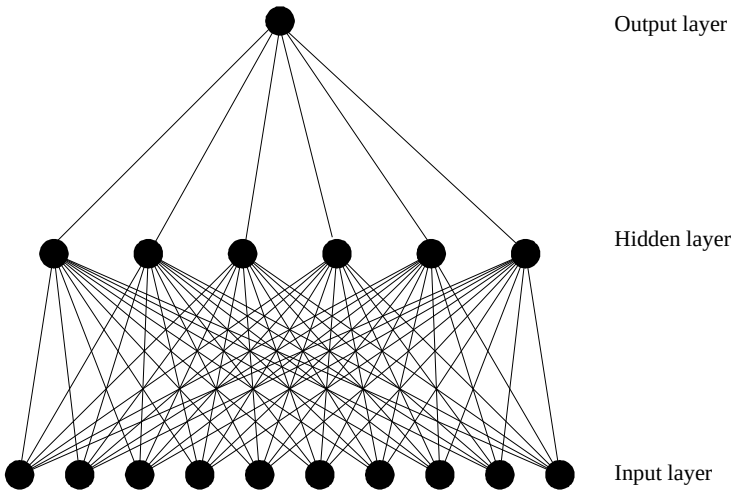

Figure 5.8: Example of a geometry of a three layer neural network. There is one input node for each variable plus one bias, an arbitrary number of hidden nodes and one output node. The nodes in two consecutive layers are catenated with variable connections.

interval [-1,+1], leading to a saturation effect. The bias mentioned above shifts the mean of the input data distribution to the linear part of the sigmoid function.

The output of the neural network is calculated by

$$o_k = S(\sum_{j=0}^{M} \omega_{j,k} \cdot S(\sum_{i=0}^{d} \omega_{i,j} x_i + \mu_{0,j})) \tag{5.3}$$

where d is the number of input and M the one of hidden nodes. ω_{ij} denotes the weights from the input to the hidden layer, ω_{jk} the weights from the hidden to the output layer. $\mu_{0,j}$ is the weight that connects the bias node with the hidden nodes.

5.11.1 The Training Process

The training of the neural network is done by minimizing the deviation between the true output and the one calculated by using the actual weights. The error function minimized in this neural network is the entropy error function, which is essentially given by the sum of the logarithms of the output values. The aim of the training of the neural network is to find the minimum in the multidimensional structure of the error function with many peaks and valleys. As this task can be difficult to resolve, the training process is done by the combined method of gradient descent and backpropagation. The neural network is trained with regularization techniques to improve generalization performance and to avoid overtraining. During the training process, weights and nodes whose significance is below a certain threshold are pruned away. This reduces the number of free parameters and hence improves the signal–to–noise ratio by removing the cause of the noise. This leads again to an improved generalization ability.

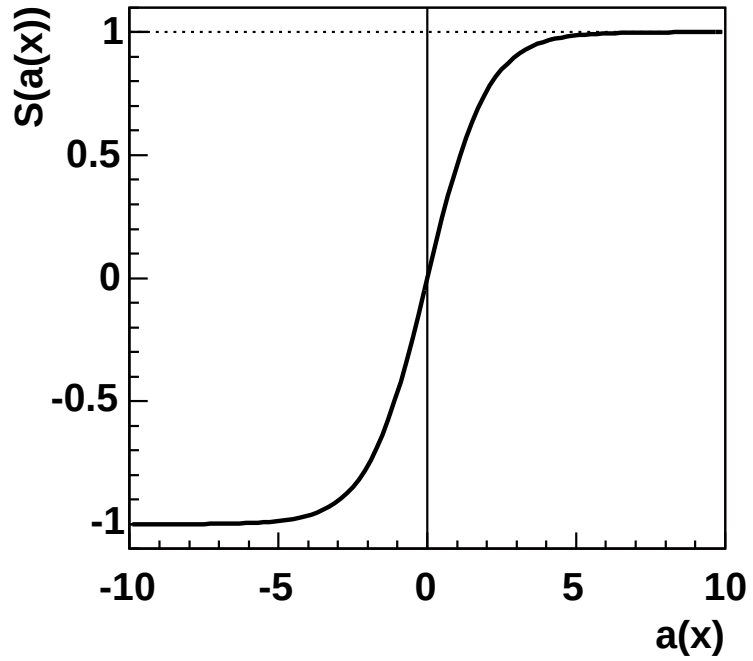

Figure 5.9: The transformed sigmoid activation function S(a(x)) as given by formula 5.2.

5.11.2 Preprocessing of the Variables

To find the optimal starting point for minimizing the error function, the input variables are preprocessed. This preprocessing is done in a completely automatic way. To care for extreme outliers, the input distributions are equalized to lie between -1 and 1. Those flattened distributions are then converted into a Gaussian distribution, centered at zero with standard deviation 1. This avoids saturation of the nodes due to the above mentioned shape of the activation function in figure 5.9.
After this transformation, the input variables are linearly decorrelated diagonalizing and rotating the covariance matrix into a unit matrix. This unit matrix is again rotated until one variable includes the complete linear correlation to the target and all other correlations are zero.
The above mentioned transformation to a Gaussian distribution may be altered by individual variable preprocessing like fitting the flattened distribution with a spline if this is considered to be sensible. In addition, discrete variables can be treated as members of classes. The preprocessing of those kinds of variables can also deal with a certain order of values if this is important, e.g. the number of tracks in a jet.
The preprocessing is also capable of dealing with variables that are not given for every event by assigning the missing values to a δ–function.
It is important to mention that the preprocessed input variables do not, unlike the original ones, have any physical meaning.

5.11.3 Automatic Variable Selection

The significances of the training variables are determined automatically during the preprocessing in NeuroBayes. This is done by removing each variable one after another and checking their correlations to the target. This correlation as well as the size of the training sample determine the training significance of each variable. After the preprocessing process, it is possible to cut on the significance to take into account only variables that include enough information that is not already incorporated by other variables.

Chapter 6

Forward Electron Identification

As detailed in the theory chapter 2.2.2, the single-top analysis can benefit from the inclusion of electrons detected by the forward detectors of the CDF experiment. It is therefore mandatory to have good identification criteria to distinguish electrons from background.
The kinematics of the electrons from W boson decays is such that electrons have a relatively high transverse momentum and energy, typically over 15 GeV. The momentum can be measured by the tracking detectors, the energy is measured with the calorimetry system.
At these energies, the only difference in the experimental signature between electrons and positrons is the curvature of their track. The term *electron* is therefore employed for both the negatively charged electron and the positively charged positron.
The background to the electrons comes mainly from QCD jet production: especially heavy flavor jets can include an electron, these are, however, considered background as one can see from identification variables like the relative isolation.
In this chapter, I will present a new method of identifying electrons in the forward region of the CDF detectors. This method is developed with an application to the single-top analysis in mind, but should be universally applicable to the identification of isolated electrons resulting from the decay of a heavy gauge boson.

6.1 Identification Variables

Several variables are used to distinguish electrons from heavy gauge boson decays and QCD background. This background has different origins:

- As already mentioned, real electrons can be included in a jet in which heavy flavor quarks decay semileptonically. This electron is, however, considered background as it is not isolated.

- Also a π^0 can fake an electron as it causes an electromagnetic shower in the calorimeter. If a charged particle like a $\pi^{\pm}$ is produced in coincidence, a track to the cluster can be found. Pions are produced through strong interactions,

thus having a much higher cross section at a hadron collider than processes that lead to isolated electrons.

- Since the development of showers in the electromagnetic and hadronic calorimeter is a stochastic process, it can sometimes happen that a jet showers early, thus depositing most of its energy in the electromagnetic part of the calorimeter. Because of the large cross section of hadronic events, this contribution is not negligible.

As one can see from these considerations, the variables which are used to separate signal from background fall into two categories: calorimeter-based variables and tracking-based variables [79, 80].

6.1.1 The Calorimeter-Based Variables

- **$\mathbf{E_T}$**: The transverse energy is defined as: $E_T = E_{\mathrm{EM}} \times \sin(\theta)$ where E_{EM} is the electromagnetic energy of the cluster, θ is the polar angle of the associated track or the position given by the shower maximum detector. The vertex along the beam line is chosen to be the Z-vertex with highest sum P_T and a quality flag ≥ 12.

- **Had/Em:** This is the ratio of the energy in the hadronic calorimeter to the energy in the electromagnetic calorimeter.

- **Isolation ratio:** The ratio of the energy inside a cone of radius $\Delta R = \sqrt{\Delta\eta^2 + \Delta\varphi^2} = 0.4$ around the electron cluster to the energy of the electron cluster. The energy of the electron cluster is subtracted from the numerator and the energy in the cone is corrected for calorimeter leakage. This quantity is often referred to as Isolation, but it is the ratio that is always meant.

- **PEM3x3 FitTower:** This is the number of towers used by the 3x3 PEM cluster fit to data from test runs with 57 GeV positrons [81].

- **PEM3x3 Fit χ^2:** This is the χ^2 value of the 3x3 PEM cluster fit to data from test runs.

- **PES 5by9 u:** The PES 5by9 u and v variables describe the shower profile measured by the PES detectors. For the PES 5by9 u variable, a ratio is formed between the energy measured with five scintillator bars in u direction and nine bars in u direction, with the center being the best matching 2d PES cluster.

- **PES 5by9 v:** Identical to the PES 5by9 u variable except that scintillator bars in v direction of the PES detectors are used to form the ratio.

- **PES-PEM ΔR:** The position of the maximum energy deposition in the PES detector is determined from the PES 5by9 u and v positions. The position of the maximum energy deposition in the PEM calorimeter is extracted from the PEM3x3 fit. ΔR is the difference between these two positions in the $\eta - \varphi$ plane.

- **Fiducial η:** Also referred to as detector η. This is the pseudorapidity of the best matching 2d PES cluster.

6.1.2 The Track-Based Variables

Tracking in the forward region $|\eta| > 1.2$ of the CDF detector is more difficult than in the central region as the full coverage of the COT drift chamber only extends to $|\eta| < 1.0$ (see section 3.2.1). Silicon standalone tracking is possible with the ISL up to about $|\eta| < 2.0$. The difficulty is that silicon standalone tracking creates a helix only from silicon measurements and for this it requires information of at least two 3D hits and one 2D hit [66]. However, due to the detector geometry, the efficiency of the unmodified algorithm is rather low at higher η. To improve the efficiency for electrons, a special algorithm that resembles the COT outside-in tracking algorithm is used. Instead of a COT seed track, an energy deposition in the forward calorimeter is taken "as seed": two helices are computed with the primary vertex as a starting point. The curvatures are derived from P_T of an electron and a positron respectively corresponding to the measured energy deposit. These two helices are then fitted to the silicon hits. The user can decide upon the χ^2 of the fit which track he wants to use for his analysis. Electrons which have such a track from this modified algorithm are called PHX or Phoenix electrons [82].
The tracking criteria for a PEM cluster to be an electron are:

- Link to a track (Standard tracking or PHX tracking).
- Number of hits in silicon ≥ 3.
- Distance on the beam axis between track and the origin of the CDF coordinate system ≤ 60 cm.

6.2 Datasets

Determining efficiency and purity of any method of identification implies the use of two sets of events: a set containing signal events and a set containing background events. These sets could be derived from Monte Carlo simulation. This however has proven to be impractical for two reasons: All samples gained this way must be compared to data to see if the underlying theoretical model describes the data. If a comparison has to be made to data, one could also use data right from the start. Furthermore, only very few background events can be obtained from Monte Carlo events. The number of simulated background events is high enough to test some hypotheses but not high enough to train a network with, as one can see later in this chapter. This is why the signal sample has been made out of the `bhel0d` data sample and the background sample out of the `ptop00` data sample. All samples were processed with version 5.3.1 of the CDF software and put into the TopNtuple format with the software version 5.3.3_nt. The size of the data samples is equivalent to an integrated luminosity of 320 pb^{-1}. The samples are presented in more detail in the following subsections.

6.2.1 The Signal Sample

Using $Z \rightarrow e^+e^-$ with one electron in the central region and one in the forward region, one can make a signal sample for plug electrons by applying the very tight central electron cuts.
The `bhel0d` sample is the central electron stream [83, 84]. It mainly contains events fulfilling the requirements of the `ELECTRON_CENTRAL_18` trigger path. The requirements of this trigger path are basically:

- A track obtained from COT tracking with $P_T \geq 9$ GeV.
- $E_T \geq 18$ GeV calculated from the energy measured in the central calorimeter system($|\eta| < 1.1$). E_T is then calculated with respect to z_0 of the track.
- $\Delta Z < 8$ cm between primary vertex and track z_0.
- A maximum of two primary vertices in the event.
- Lateral shower profile LShr < 0.4
- An additional requirement is put on the central electrons Had/Em ≤ 0.125.

The events that fulfill these requirements must have been recorded with the necessary parts of the detector being in a good state. In CDF, data-taking is divided into so-called runs that extend from a few minutes up to several hours in which the beam and detector properties are not changed. The good run list is a collection of runs which contain at least a certain number of events and for which the detector was in a good state. Version 7 of this list is used for this thesis, this includes runs from March 2002 until September 2004.
If the events are in this list, they are subsequently subject to the following cuts:

- The event must have exactly one tight central electron with an associated track, no jet should be in the central region of the detector.
- This electron must additionally fulfill harder constraints on the lateral shower profile LshrTrk < 0.1 (instead of < 0.2), on the relative isolation < 0.05 (instead of < 0.1) and $E_T > 20$ GeV. This makes relatively sure that the central electron is not a fake electron. This electron is called the central leg.
- There should be exactly one electron candidate in the forward region of the CDF detector. No additional jet should be in the forward region of the detector.
- This electron candidate must have a track associated with it fulfilling the requirements mentioned in section 6.1.2. The kinematic requirements on the electron are $E_T > 20$ GeV and $1.2 < |\eta| < 2.0$ with η being the detector η.
- Additionally, we ask for Had/Em < 0.1, isolation < 0.25, PEM 3x3 FitTowers $\neq 0$, and PEM $\chi^2 < 15$. This electron candidate is called the plug leg.

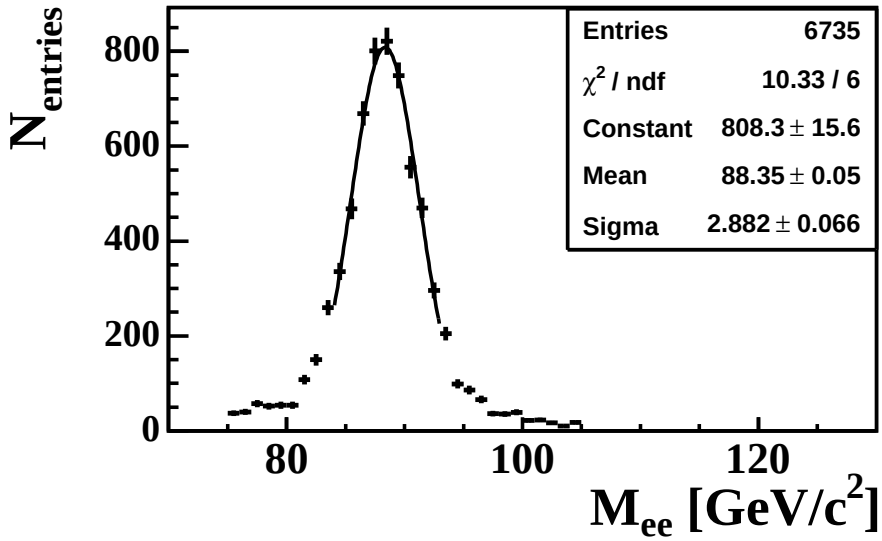

Figure 6.1: M_{ee} for events fulfilling the requirements detailed in section 6.2.1. A Gaussian is fitted to determine the mean value of the distribution. Specific corrections as used by CDF in [85] e.g. are applied neither to the central nor to the plug leg to stay compliant with the single-top analysis.

- The invariant mass M_{ee} of the central and plug leg must be in the range 75 GeV/c^2 < M_{ee} < 105 GeV/c^2. Figure 6.1 shows M_{ee} for this sample. The shape of this distribution indicates that most of the events are Z bosons decaying into electron and positron. A Gaussian is fitted to the peak region. The mean value of the fitted Gaussian satisfactorily reproduces the value for M_Z = 91.2 GeV/c^2 known from literature [13]. The deviation can be explained as the specific corrections applied for example in the measurement of $\sigma \cdot \mathrm{Br}(Z^0 \to e^+e^-)$ [85] are not applied, as these are not applied in the single-top analysis.

Applying the standard CDF cuts on the plug electron variables, 6204 out of 6735 events remain, as can be seen in table 6.1.

Cut	#Signal	#Backgr.	Signal eff.%	Backgr. eff.%
None	6735	10126	100	100
Had/Em < 0.05	6699	8655	99.47	85.47
+ Isolation < 0.1	6583	6244	97.74	61.66
+ PEM χ^2 < 10	6417	5429	95.28	53.62
+ PES 5/9 u > 0.65	6342	5153	94.16	50.89
+ PES 5/9 v > 0.65	6277	4883	93.2	48.22
+ ΔR (PES-PEM)< 3.0	6204	4391	92.12	43.36

Table 6.1: Cut flow for selection cuts on signal and background sample, as asked by the electroweak and top physics working groups.

The identification efficiency of the signal sample is $(92.1 \pm 1.2)\%$, which is in good agreement with the one obtained by other analyses.

6.2.2 The Background Sample

The background sample used for training is also derived from data. The idea is to use a dijet sample, in which one jet is measured with the central calorimeter and the forward jet is an electron candidate.
The background sample is derived from the `ptop00` dataset which is stripped from `bpel0d`. The `bpel0d` dataset is the plug electron stream [83, 84]. It mainly contains events passing the requirements of the `PLUG_ELECTRON_20` trigger path. The requirements of this trigger path are basically:

- Energy deposition $E_T > 20$ GeV in the plug region of the calorimeter.
- Detector η: $1.1 < |\eta| < 3.6$ at trigger level
- Had/Em < 0.125

Additional cuts are applied to the events while stripping `bpel0d` down to `ptop00` [86]:

- $E_T > 10$ GeV (This cut should not affect any events, as the trigger cut is higher. However, this requirement was added to be compliant with other datasets for which no requirement is put on E_T at trigger level)
- Had/Em < 0.1
- Isolation ratio < 0.2
- PEM $\chi^2 < 15$

Events which pass these trigger and stripping cuts as well as the good run requirements are then subject to the following preselection cuts:

- Exactly one jet with $E_T > 15$ GeV in the central part of the detector, no electron in the central part.
- The fraction of the energy deposition in the CEM versus the total calorimeter energy deposit should be < 0.8.
- Exactly one plug electron candidate. The additional requirements for the kinematic and calorimeter variables are the same as for the signal sample.
- No track should point to the electron candidate. Requiring a track would diminish the number of training events and would make a training impossible. The distributions of the calorimeter variables do not differ largely between electrons with a track and electrons without a track.
- The central jet and the plug electron candidate should be back to back in φ: $\cos \Delta\varphi < -0.99$.

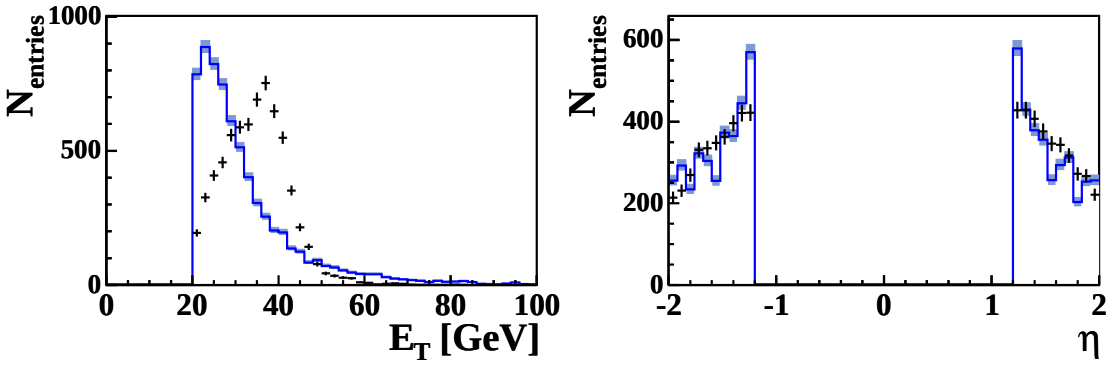

Figure 6.2: Electron E_T [GeV] and detector η. The signal is represented by crosses with error bars and the background by a solid line. The background is scaled to the number of signal events.

- E_T of the central jet and the plug electron candidate should be balanced: the difference in E_T should be smaller than the energy resolution of the central calorimeter $0.8\sqrt{E}$.

Applying the selection cuts results in the cut flow listed in table 6.1. From an initial number of 10126 events, one ends up with a total of 4391 events, which is an efficiency of $(43.4 \pm 0.7)\%$ on this background sample.

6.3 Distributions of Selection Variables

To better compare the variables used for the selection, the distributions are presented. The quantities are plotted after preselection, so the distributions total 6735 signal and 10126 background events. Every plot shows signal (crosses) and background (full line), with background scaled to the number of signal events.

Figure 6.2 shows the E_T and pseudorapidity (η) distributions. η looks quite the same for signal and background. The E_T distribution of the signal peaks at about 30 GeV, whereas the background shows a typical falling QCD distribution. As shown in figure 6.3, the mass of the reconstructed Z boson clearly peaks at around 90 GeV while the background shows a typical QCD decrease. Also in figure 6.3, the selection variable Had/Em is plotted. The reason for the long tail of the background distribution is the much higher energy deposition in the hadronic calorimeter for background events than for signal electrons. Figure 6.4 shows the distributions of isolation and PEM χ^2. The small isolation for real electrons can be explained by the fact that the events come from Z boson decays and are therefore isolated, whereas the electromagnetic cluster of the background are part of a hadronic shower. The PEM χ^2 values should be small as the fit is performed on real electrons. The PES 5by9 u and PES 5by9 v distributions in figure 6.5 again show that showers produced by real electrons are smaller than those produced by jets. Real electrons do not show a large difference in the $R = \eta - \varphi$ plane between the fitted position in the PEM

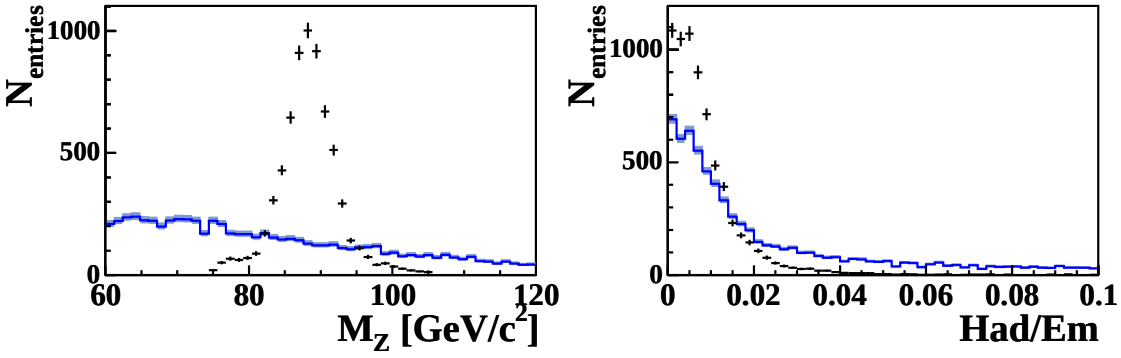

Figure 6.3: Mass of the reconstructed Z candidate and electron Had/Em. The signal is represented by crosses with error bars and the background by a solid line. The background is scaled to the number of signal events.

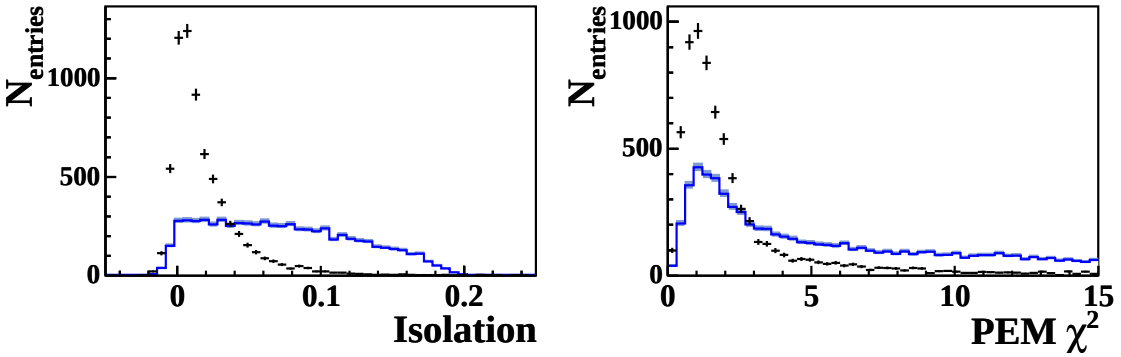

Figure 6.4: Electron isolation and electron PEM χ^2. The signal is represented by crosses with error bars and the background by a solid line. The background is scaled to the number of signal events. The small negative values of the isolation values can be explained as the energy of the cluster is subtracted from the leakage corrected energy in a cone of $R = 0.4$.

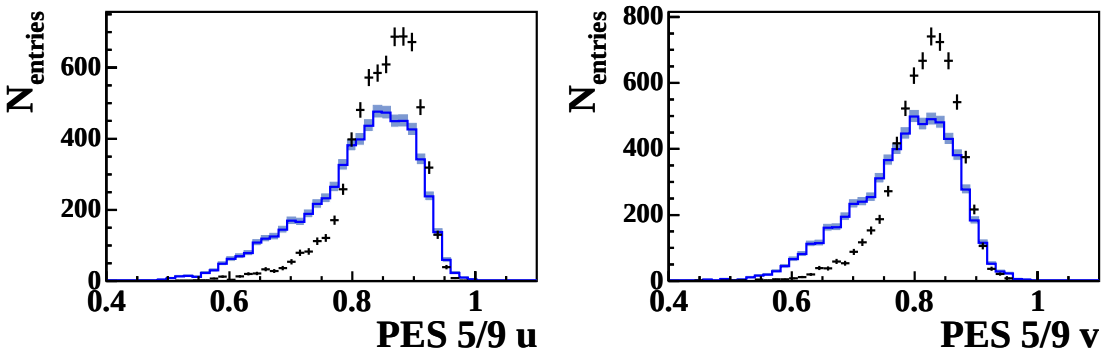

Figure 6.5: Electron PES 5/9 u and electron PES 5/9 v. The signal is represented by crosses with error bars and the background by a solid line. The background is scaled to the number of signal events.

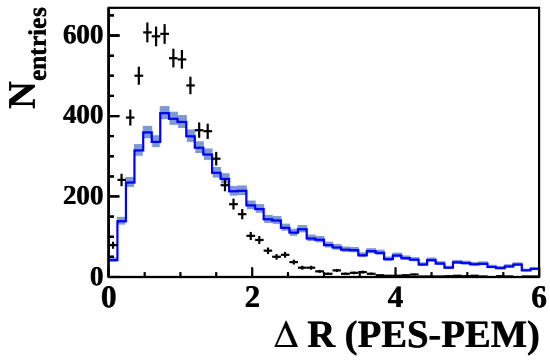

Figure 6.6: Difference in the $R = \eta - \varphi$-plane between the fitted position in the PEM and the position in the PES detector. The signal is represented by crosses with error bars and the background by a solid line. The background is scaled to the number of signal events.

and the position in the PES detector. This distance is larger for fake electrons, as can be seen in figure 6.6.

Performing sequential cuts is an optimal strategy to distinguish between signal and background only if the variables are not correlated among each other. This is, however, not the case, as one can see from table 6.2. The correlations in the case of the two PES variables are due to the geometry of the detector: the u and v directions are not perpendicular, but 45 degrees. Additionally, until the end of 2003, cross-talk between the photomultipliers used to read out the u and the v detector enhanced the correlations among these variables [87]. Another example for correlations is the correlation between the PES-PEM ΔR variable and the PEM χ^2. If the PEM fit performs well, the position in the PEM detector can be best measured, thus resulting in only a small distance between the PES and the PEM position.

Correlations in the signal sample						
Variable	$\frac{Had}{Em}$	isolation	PEM χ^2	PES $\frac{5}{9}$ u	PES $\frac{5}{9}$ v	ΔR
$\frac{Had}{Em}$	100	28	16	-7	-6	3
Isolation		100	31	-14	-15	7
PEM χ^2			100	-23	-22	21
PES $\frac{5}{9}$ u				100	52	-18
PES $\frac{5}{9}$ v					100	-20
PES-PEM ΔR						1.00
Correlations in the background sample						
Variable	$\frac{Had}{Em}$	Isolation	PEM χ^2	PES $\frac{5}{9}$ u	PES $\frac{5}{9}$ v	ΔR
$\frac{Had}{Em}$	100	13	4	6	8	-12
Isolation		100	29	1	-1	0
PEM χ^2			100	-4	-5	11
PES $\frac{5}{9}$ u				100	58	-10
PES $\frac{5}{9}$ v					100	-12
PES-PEM ΔR						100

Table 6.2: The correlation matrix of the training sample (in %). (Upper part: signal sample, lower part: background sample)

The same argumentation holds for the negative correlation between the PES-PEM ΔR variable and the two PES variables: if the deposition in the PES is point-like (resulting in large PES 5by9 u resp. v values), the ΔR value will be small.

6.4 Artificial Neural Network Technique

The samples used for training are the same as those used to determine the cut efficiencies. As described, the signal and the background sample are derived from data which is novel for training of a neural network. Usually, to make sure that the respective signal or background sample is pure, one uses simulated events. It was, however, difficult to find a Monte Carlo simulation that well describes the variables used for the training in the data sample.

The input variables are the same as for the sequential cut analysis:

- Had/Em
- Isolation
- PEM χ^2
- PES 5/9 u
- PES 5/9 v
- PES-PEM ΔR

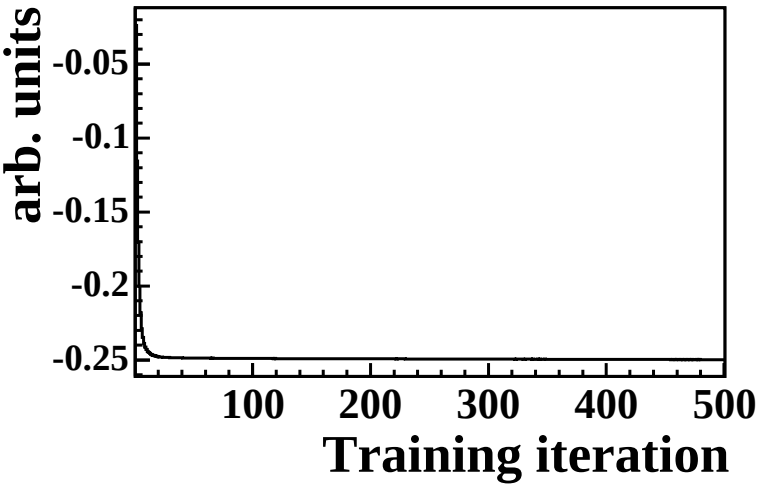

Figure 6.7: Error on the training sample as a function of the training iteration.

Usually, when training a neural network, the signal and background sample should be such that their relative weight matches the a-priori composition of the data analyzed. For the purpose of this thesis, however, the data to be analyzed is rather different from the training data: the single-top working group is interested in events which are mostly in the $W+2$-jet bin and $W+3$-jet bin. Furthermore, other physics groups might be interested in the tool, so training with a certain weight might be better for one case but worse for another. For this reason, I made the decision to weight signal and background events such that the composition is 50:50. The network used is the NeuroBayes package developed by phi-t described in 5.11. This package not only offers a neural network but also various preprocessing options. The setup of the network used in this analysis is such that variables are decorrelated and normalized and then transformed to a Gaussian distribution. The training is better if the variables are subject to a regularized spline fit. While 7 nodes are used for the input layer (6 variables plus one bias node), 8 nodes are used for the hidden layer. However, the network output is not too sensitive to the number of nodes in the hidden layer. To ensure that the network is well trained, is is trained over 500 iterations. Figure 6.7 shows the error as a function of the iteration. A value of less than 100 iterations would have been enough as the error is not diminishing after these iterations. Since overtraining is not a problem and also the time spend for the training is not a problem, the network is trained for such a high value of iterations.

The output distribution of the training sample can bee seen in figure 6.8. The background distribution has a peak at around 0.7, which is probably due to signal contamination or irreducible signal-like background in the background sample. This is also the reason why the network output is not larger than 0.8: the network cannot distinguish any event with absolute certainty between signal and background.

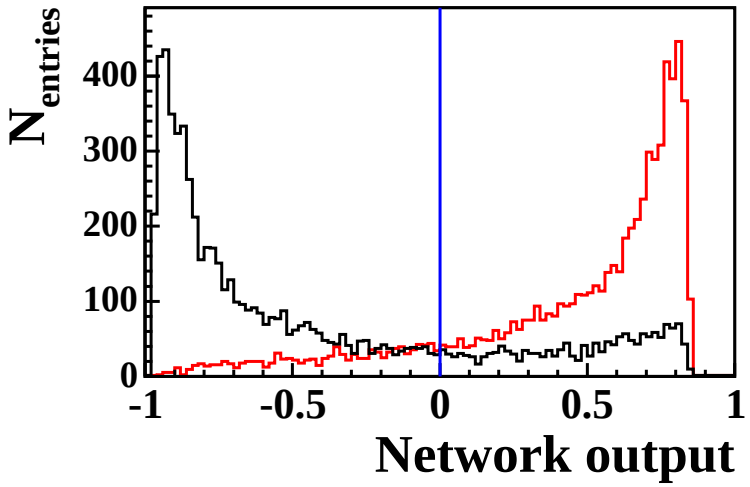

Figure 6.8: Output of the neural network, signal (line peaking at around 0.8) and background (line peaking at around -1).

The correlations between the variables in the training sample have been computed in table 6.3 and can be seen in figure 6.9. The target variable is 1 in case of a signal event, -1 in case of a background event. The correlations are different in this table compared to table 6.3 as both categories are included. Furthermore, the variables have been preprocessed, in this case, all distributions have been flattened and exchanged by the result of a regularized spline fit to means of the target.

Variable	Target	$\frac{Had}{Em}$	Isolation	PEM χ^2	PES $\frac{5}{9}$ u	PES $\frac{5}{9}$ v	ΔR
Target	100	38	60	42	25	25	33
$\frac{Had}{Em}$		100	43	28	14	14	14
Isolation			100	53	22	22	27
PEM χ^2				100	24	23	39
PES $\frac{5}{9}$ u					100	45	23
PES $\frac{5}{9}$ v						100	25
PES-PEM ΔR							100

Table 6.3: The correlation matrix of the training sample. The target is 1 for signal, -1 for background. (correlation in %, only positive values are given by NeuroBayes due to the preprocessing of the variables).

One can derive from Bayes theorem that the output of a well-trained neural network is a measure of the purity. One can see in figure 6.10 that, in the limits of statistics, the network output populates the area around the diagonal line, which indicates that the network is well-trained.

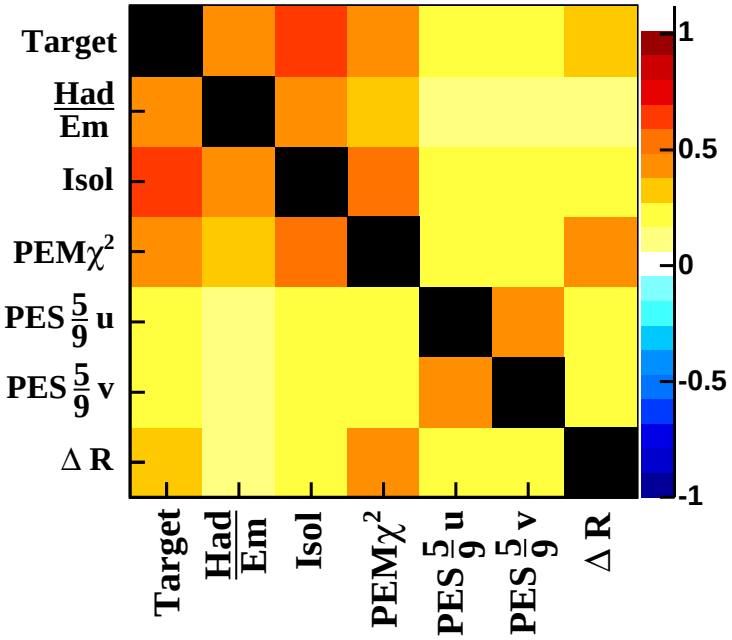

Figure 6.9: Graphical representation of the correlations among variables in the training sample. The target variable is -1 for background and 1 for signal.

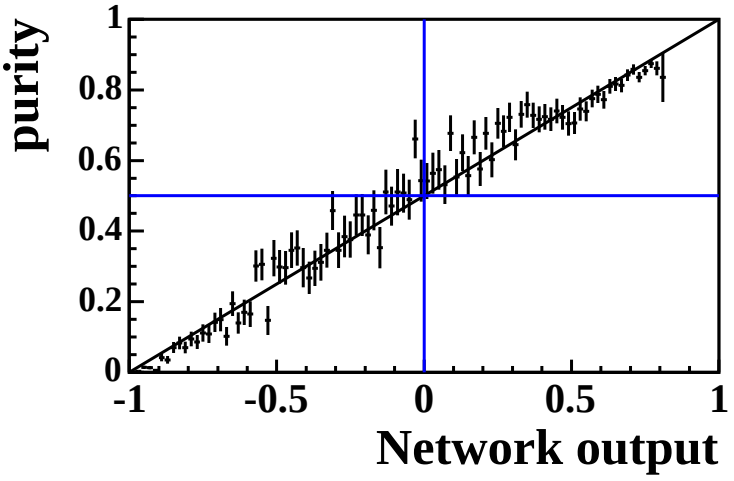

Figure 6.10: Purity versus network output.

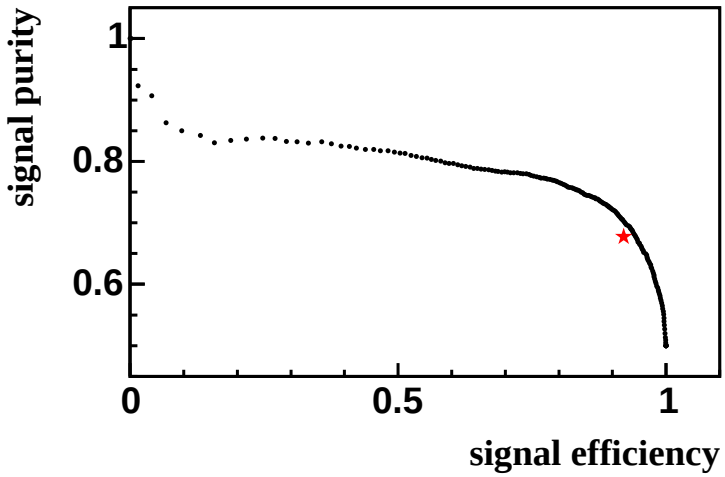

Figure 6.11: Signal purity versus Signal efficiency. The dots represent different network output cuts, the star represents the CDF cuts.

Overtraining, i.e. the danger that the neural network learns the individual events by heart, is not a problem for NeuroBayes. Moreover, if the neural network would learn the training sample by heart, the performance on the training sample would be good, but the performance on the completely different W+jets sample would be worse. The reader will see in the following chapter that this is not the case.

To determine a selection cut and to compare the performance of the neural network with the CDF cuts, one can plot the signal purity versus the signal efficiency for different cuts on the neural network output and the CDF cuts. This is done in figure 6.11. One can tune the network output such that the signal efficiency is the same as for the CDF cuts. One then has 3948 background events instead of 4348, which is 13% less background. Alternatively, one could choose a neural network cut such that the signal purity is the same. One would then have 6309 signal events instead of 6204, which is an increase in efficiency of 1.7%.
In chapter 7, the reader will see how the neural network performs on the sample used for the single-top analysis.

6.5 Comparison of Signal with Simulation

In this section, the signal obtained from data is compared to simulated events. The Monte Carlo sample is part of the `ztop7i` dataset, which contains $Z \to e^+e^-$ events generated with the Pythia Monte Carlo generator [88]. The preselection cuts are the same as for the signal with the difference that no good run criterion is applied. The data is represented by crosses with error bars. The Monte Carlo distributions are represented by a solid line.

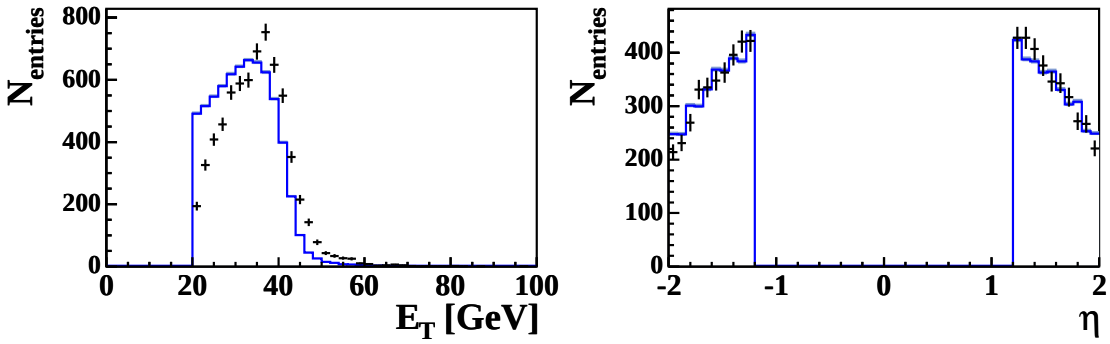

Figure 6.12: Electron E_T [GeV] and detector η. The signal distribution extracted from data is represented by crosses with error bars. The signal distribution extracted from Monte Carlo events is represented by a solid line.

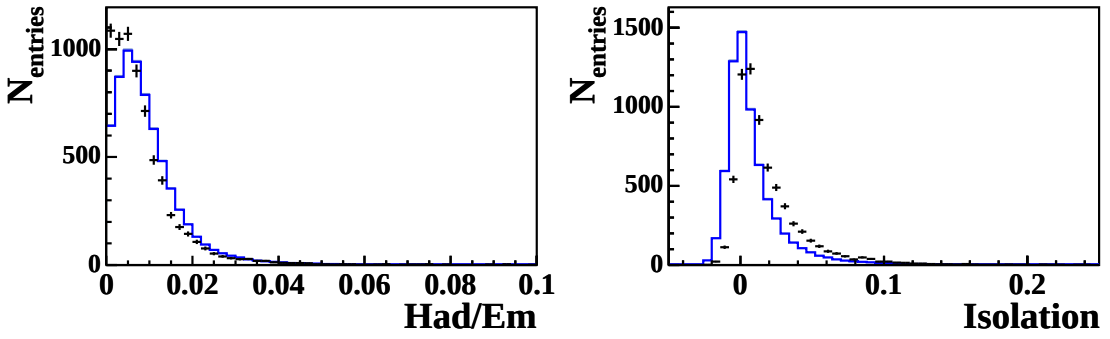

Figure 6.13: Electron Had/Em and isolation. The signal distribution extracted from data is represented by crosses with error bars. The signal distribution extracted from Monte Carlo events is represented by a solid line.

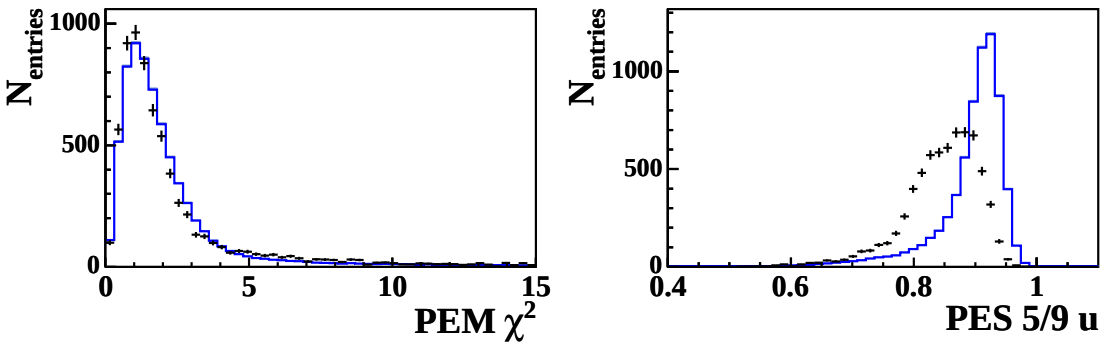

Figure 6.14: Electron PEM χ^2 and electron PES 5/9 u. The signal distribution extracted from data is represented by crosses with error bars. The signal distribution extracted from Monte Carlo events is represented by a solid line.

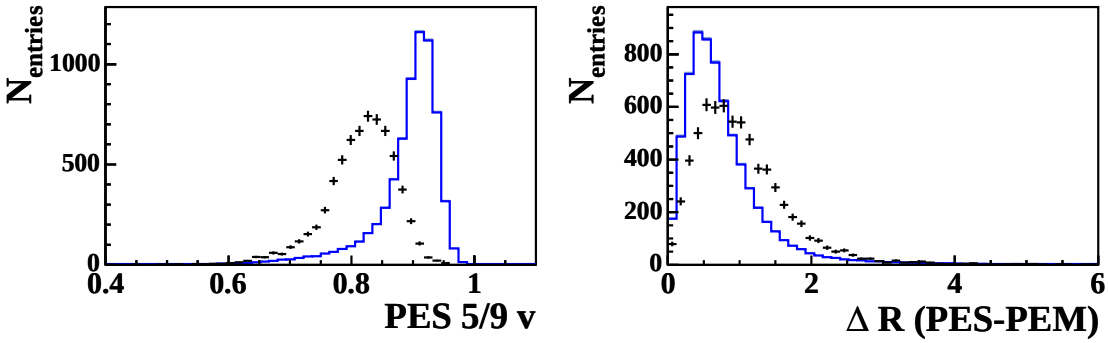

Figure 6.15: Electron PES 5/9 v and ΔR (PES-PEM). The signal distribution extracted from data is represented by crosses with error bars. The signal distribution extracted from Monte Carlo events is represented by a solid line.

The difference in E_T in figure 6.12 between data and simulated events can be explained by the missing trigger requirements in the simulation. In real data the trigger, which demands an E_T of above 18 GeV has a turn on curve, therefore the trigger only becomes fully effective at larger values of E_T. In the same figure, one can see that η is very well simulated. Small differences can be seen in the Had/Em distributions and the PEM χ^2 fit value in figure 6.13 and 6.14 respectively. The isolation in figure 6.13 and especially the two PES variables in this figure and figure 6.15 show noticeable differences between data and the simulation. It is known that the PES variables are not well simulated. Furthermore, until the 2003 shutdown, cross-talk between the two PES detectors smeared out the distributions. For the same reason, ΔR (PES-PEM) peaks at smaller values for simulated events.

Table 6.4 shows the cut flow for the standard cuts applied to the Monte Carlo sample. The efficiency is 94.3%, which is somewhat higher than the $(92.1 \pm 1.2)\%$

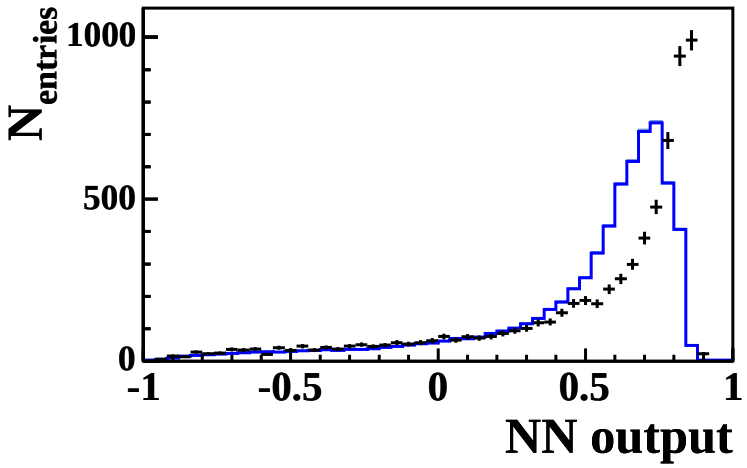

Figure 6.16: Output of the neural network. The signal distribution extracted from data is represented by crosses with error bars. The signal distribution extracted from Monte Carlo events is represented by a solid line.

efficiency of the cuts on data. One would expect the network output for the Monte Carlo sample to be higher than the output for the data sample, this is, however, not the case, as seen in figure 6.16.

Cut	#Signal	Signal eff.%
None	338908	100%
Had/Em < 0.05	336829	99.39%
+ Isolation < 0.1	333113	98.29%
+ PEM $\chi^2 < 10$	327498	96.63%
+ PES 5/9 u > 0.65	325081	95.92%
+ PES 5/9 v > 0.65	323581	95.48%
+ ΔR (PES-PEM)< 3.0	319636	94.31%

Table 6.4: Cut flow for selection cuts on the ztop7i MC sample, as asked by the Electroweak and top working groups.

As a conclusion to the comparison between the signal samples extracted from data and the signal sample extracted from simulation, I can say that my decision to train the neural network on data is justified. Training on simulated events will only be feasible if the simulation is tuned to match the data distributions.

6.6 Comparison of Background with Simulation

In this section, the background obtained from `ptop00` data is compared to a Monte Carlo sample. As no specialized description of electron-fakes exists, the Monte Carlo

events have to be derived from other samples. The idea is to look at jet production and apply the electron identification algorithm to these jets. Three categories of Monte Carlo samples are taken into account:

- W+jet production where $W \rightarrow \mu\nu_\mu$. The muons are not identified by the algorithm, so only jets can be misidentified as electrons.
- Di-jet production.
- W+jet production where $W \rightarrow e\nu_e$. Electrons that are found in the region of the electron from the W decay as know from the Monte Carlo truth information are rejected. Thus, only jets can be misidentified as electrons.

Sample	Simulated process	Generator	Processed events
atopcb	$W \rightarrow \mu\nu_\mu$+1p	Alpgen [89] + Herwig [90]	196285
atopdb	$W \rightarrow \mu\nu_\mu$+2p	Alpgen + Herwig	252395
atopeb	$W \rightarrow \mu\nu_\mu$+3p	Alpgen + Herwig	291199
atopfb	$W \rightarrow \mu\nu_\mu$+4p	Alpgen + Herwig	287271
atopib	$W \rightarrow \mu\nu_\mu + b\bar{b}$+1p	Alpgen + Herwig	201897
atopkb	$W \rightarrow \mu\nu_\mu + c\bar{c}$+0p	Alpgen + Herwig	291858
atoplb	$W \rightarrow \mu\nu_\mu + c\bar{c}$+1p	Alpgen + Herwig	288863
atopmb	$W \rightarrow \mu\nu_\mu + c\bar{c}$+2p	Alpgen + Herwig	254511
atopwb	$W \rightarrow \mu\nu_\mu + b\bar{b}$+0p	Alpgen + Herwig	219002
btop5a	dijet	Herwig	84455
btop6a	dijet	Herwig	78897
btop7a	dijet	Herwig	151802
btop8a	dijet	Herwig	151673
ltop0m	$W \rightarrow \mu\nu_\mu$+0p	Alpgen(v1.3.3) +Herwig	300296
ltop1m	$W \rightarrow \mu\nu_\mu$+1p	Alpgen(v1.3.3) +Herwig	82074
ltop2m	$W \rightarrow \mu\nu_\mu$+2p	Alpgen(v1.3.3) +Herwig	187851
ltop3m	$W \rightarrow \mu\nu_\mu$+3p	Alpgen(v1.3.3) +Herwig	156384
ltop4m	$W \rightarrow \mu\nu_\mu$+4p	Alpgen(v1.3.3) +Herwig	100808
ltop3b	$W \rightarrow \mu\nu_\mu + b\bar{b}$+0p	Alpgen(v1.3.3) +Herwig	248380
ltop3c	$W \rightarrow \mu\nu_\mu + c\bar{c}$+0p	Alpgen(v1.3.3) +Herwig	312372
ltop4b	$W \rightarrow \mu\nu_\mu + b\bar{b}$+1p	Alpgen(v1.3.3) +Herwig	283728
atop5a	$W \rightarrow e\nu_e$+2p	Alpgen + Herwig	180208
atopaa	$W \rightarrow e\nu_e$+1p	Alpgen + Herwig	208351

Table 6.5: Monte Carlo samples used to extract electron fakes. In total, 4.8 million events are processed out of which 1804 events are found with a jet identified as an electron.

Table 6.5 lists all the Monte Carlo samples used to derive a simulated background. In total, 4.8 million events are processed out of which only 1804 events are found in which a jet is misidentified as an electron. These are too few events to train the neural network with. One can however perform crosschecks on this sample to see if the background sample made from data agrees with the simulated background

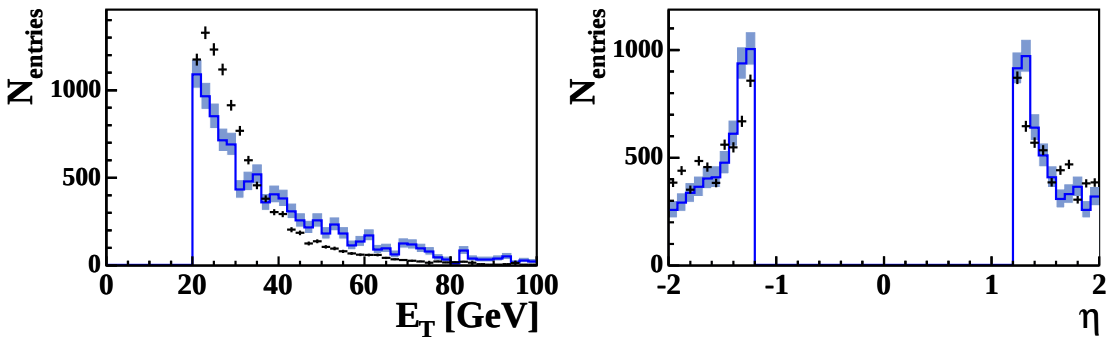

Figure 6.17: Electron E_T [GeV] and detector η. The background distribution extracted from data is represented by crosses with error bars. The simulated background events are represented by a solid line with error bands.

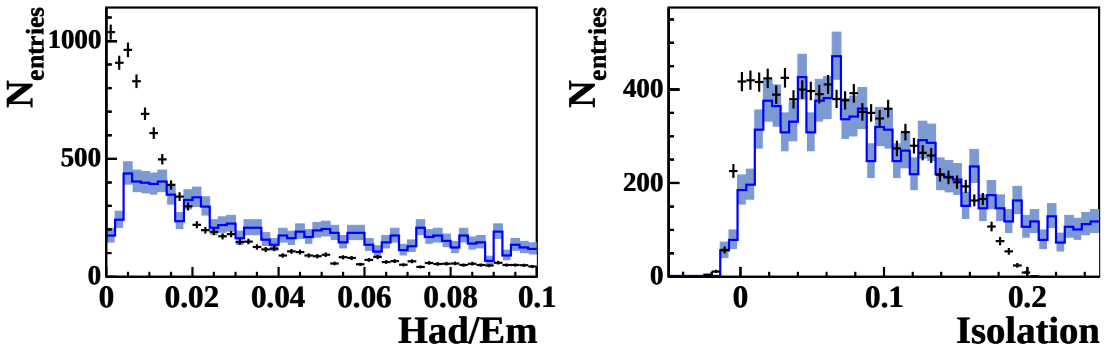

Figure 6.18: Electron Had/Em and isolation. The background distribution extracted from data is represented by crosses with error bars. The simulated background events are represented by a solid line with error bands.

distributions. As shown in figure 6.17, small differences can be seen in the E_T distributions. These are due to trigger effects, which only play a role on the data events. Also on figure 6.17, one can see that the η distributions differ. This is due to cuts made at generator level on some Monte Carlo samples, the dijet Monte Carlo events have only very few jets (i.e. fake electrons) with $|\eta| > 1.5$. Some of the simulated events contain heavy flavor quarks which decay semileptonically. It has not been checked whether the fraction of such events in the simulated dataset is the same as in data events.

Differences can also be seen in the Had/Em distributions and the PEM χ^2 fit value in figure 6.18 and 6.19. This shows that the background sample obtained from data is either not pure or other objects deposit energy in the electromagnetic part of the calorimeter. One good candidate for such objects could be pions, especially π^0 which decay into two photons. The preselection cuts for the background are such that explicitly no track is asked for, which could be an indication for π^0. To further

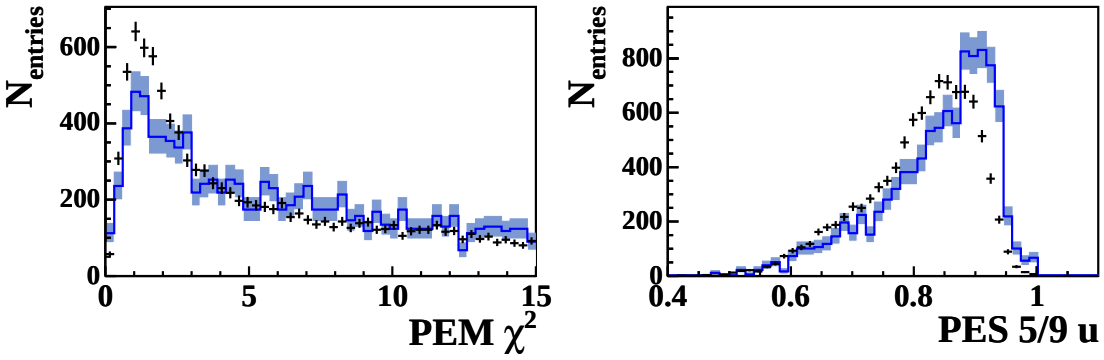

Figure 6.19: Electron PEM χ^2 and electron PES 5/9 u. The background distribution extracted from data is represented by crosses with error bars. The simulated background events are represented by a solid line with error bands.

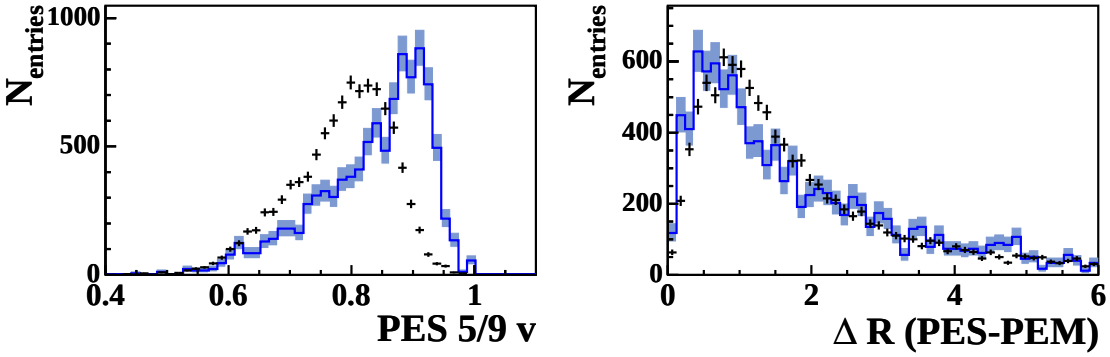

Figure 6.20: Electron PES 5/9 v and ΔR (PES-PEM). The background distribution extracted from data is represented by crosses with error bars. The simulated background events are represented by a solid line with error bands.

test this hypothesis, one would need, however, more Monte Carlo events.
The two PES variables in figure 6.19 and 6.20 show noticeable differences between data and simulation, this for the same reasons as for the signal sample. Again, this also affects the ΔR (PES-PEM) distributions.

The output distribution of the neural network in data has two peaks, a larger one as expected at lower values and a smaller one at higher values, as shown in figure 6.21. This shows that some events in the background sample are very signal-like. The hypothesis is that these signal-like events would have been removed if a track is asked for. Due to lack of statistics, this hypothesis cannot be tested. The shape of the Monte Carlo distribution is as expected, no peak at higher output values can be seen. In the next chapter, this distribution is used as a template distribution for fits, one has therefore different possibilities:

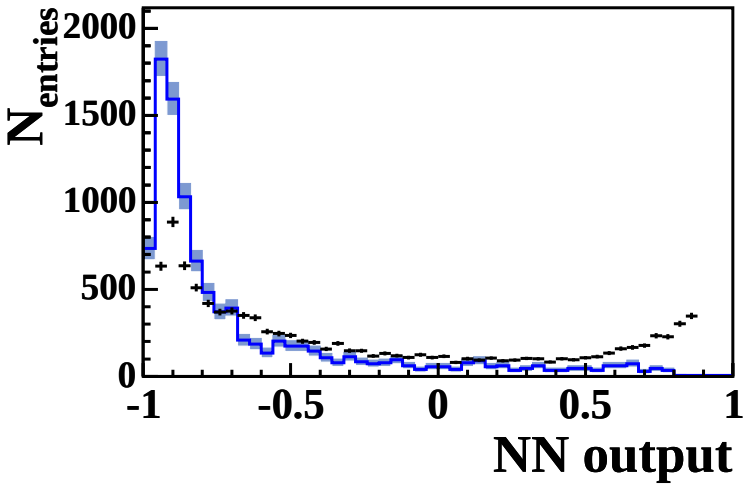

Figure 6.21: Output of the neural network. The background distribution extracted from data is represented by crosses with error bars. The simulated background events are represented by a solid line with error bands.

- Take the output distribution from data as a template, with the knowledge that the description is not adequate. This would systematically overestimate the background content in a fit.
- Modify the data output distribution such that the peak at larger output values matches the expected distribution. This could be done with a linear or exponential fit in the middle range of the distribution and exchanging the bin contents in the higher output values range by the prediction of the fit.
- Take the Monte Carlo distribution as a template. This has the drawback that statistics is small. However, the shape appears to be the most realistic one for higher values of the network output.

In the next chapter, these Monte Carlo distributions are used when a background template is needed.
The efficiency for the CDF cuts on Monte Carlo background events is $(27.5 \pm 1.3)\%$, as can be seen in table 6.6. This is smaller than the efficiency of the CDF cuts on the background sample extracted from data.

Cut	#Signal	Signal eff.%
None	1804	100%
Had/Em < 0.05	1142	63.3%
+ Isolation < 0.1	694	38.47%
+ PEM $\chi^2 < 10$	599	33.2%
+ PES 5/9 u > 0.65	577	31.98%
+ PES 5/9 v > 0.65	561	31.09%
+ ΔR (PES-PEM)< 3.0	496	27.49%

Table 6.6: Cut flow for selection cuts on the background MC sample, as asked by the Electroweak and top working groups.

Chapter 7

W+jets Events

In this chapter, I will use the electron identification algorithm developed in chapter 6 to select $W+n$-jet events. I will use the standard CDF 4-sector method to determine the background contamination of the signal and compare it with two novel methods of determining the QCD background contamination. The first method fits two $\not\!\!E_T$ templates to the data to extract the background content. The second method fits a signal and background template for the output of the neural network to the data in order to determine the background contamination.

7.1 Datasets

7.1.1 The W+jets data sample

Again, the `ptop00` data sample described in chapter 6 is used. However, the selected events are different. This starts with the trigger requirement: instead of the `PLUG_ELECTRON_20` path asked for in section 6.2.2, the `MET_PEM` path is required. These two are essentially identical in their requirements on the electron. The `MET_PEM` trigger puts an additional cut on the missing transverse energy (MET or $\not\!\!E_T$) above 15 GeV at trigger level. This cut reduces the number of (mainly background) events for any analysis in which a leptonic decay of the W boson is involved.

The selection criteria are also different: to select W+jet events, exactly one plug electron candidate is required. If jets are present in the sample, they should be more than 0.4 away from the electron candidate in $\eta - \varphi$-plane. Various cuts are applied on the plug electron candidates, either the standard CDF cuts or cuts on the neural network output.
The jet multiplicity is the number of jets reconstructed using the cone algorithm with radius 0.4. The jets must have a minimum E_T of 15 GeV after level 4 correction and satisfy $|\eta| < 2.8$.

Sample	Process	Generator	σ, Events	MC version
wtop1i	$W \to e\nu$	Pythia	1961 pb, 2M	5.3.2
atopaa	$W \to e\nu$+1 parton	Alpgen+Herwig	682.4 pb, 0.2M	5.3.3
atop5a	$W \to e\nu$+2 parton	Alpgen+Herwig	246.7 pb, 0.2M	5.3.3

Table 7.1: List of MC samples. Pythia is described in reference [88], Alpgen in reference [89] and Herwig in reference [90].

7.1.2 The Monte Carlo Sample

Different Monte Carlo samples are used for different jets bins and different processes. The samples are listed in table 7.1. These are official Monte Carlo samples from the CDF top group, which are available in the TopNtuple format.

7.1.3 The Background Samples

The different methods for estimating the background content need different descriptions of background. Common to all of them is the interpretation of what is background. In this chapter, background to the W+jets sample means QCD production in which a jet is faking an electron and energy is missing due to detector effects, mismeasurements or neutrinos coming from semileptonic decays of B mesons.
For every method described, the reader will find information on how exactly the background template was obtained.

7.2 The 4-Sector Method

The traditional way of evaluating the QCD background contamination of the data sample is the so-called 4-sector method. It was developed for the central electrons and is adapted to the plug electrons. All cuts but the isolation cut are performed on the data sample. Then, the isolation is plotted versus the missing transverse energy. By a simple geometrical consideration one can subsequently determine the QCD background contamination in the signal region of lower isolation and higher missing transverse energy. The idea behind this method is that the background is uniformly distributed in isolation, and the signal is concentrated at large $\not\!\!E_T$ and low isolation values. In figure 7.1 this region is denoted sector D. The number of background events in sector D is then given by

$$N_D^{\text{bac}} = \frac{N_C}{N_A} N_B$$

where $N_{A,B,C}$ are the number of total events in the respective sector.
One could say that the background is represented by data in the sideband regions A, B and C.

In a second iteration, the number of signal events in region D is used to correct the background regions A, B and C for signal content. A W+jets Monte Carlo

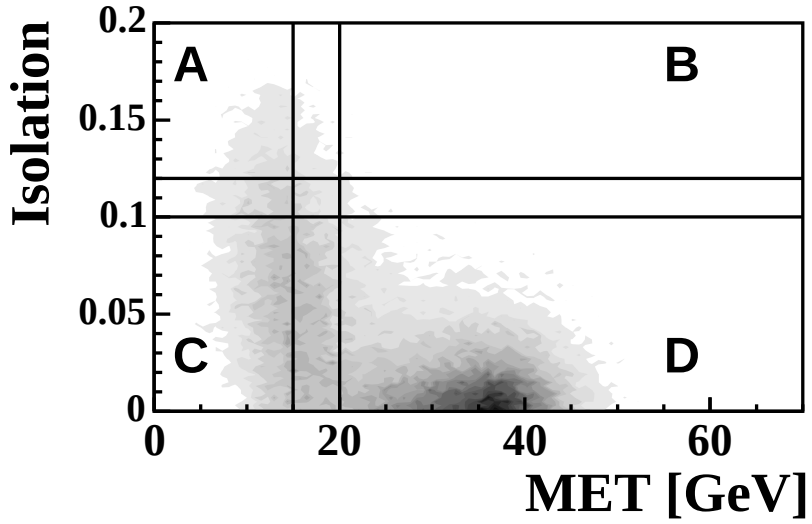

Figure 7.1: The distribution of events passing the CDF cuts except the isolation cut in the isolation versus $\not\!\!E_T$ plane. The darker the color, the higher the event density. The four sectors are indicated. Sector D is the region in which the signal is expected.

sample is normalized such that the number of signal events in the signal region is identical to the number of signal events in the data signal region. The number of simulated events in the three other regions are then subtracted from the respective regions in the data. Finally, the signal and background fraction in the signal region is recomputed.
If the 4-sector method is used in an analysis, the background fraction is always computed using this second iteration. For the sake of completeness, I will also indicate the background fractions obtained by only using the first iteration without the correction with simulated events. This method is then denoted by "uncorrected 4-sector method". The complete method with the correction in the second iteration is denoted by "MC-corrected 4-sector method" or, if no confusion is possible, simply by "4-sector method".

One of the drawbacks of the 4-sector method is that it does not allow to reconstruct the shape of the background distribution. The method cannot be applied when one uses a neural network cut instead of the sequential cuts for the identification of electrons: the cut on the isolation cannot be separated from the neural network cut, therefore the four sectors cannot be populated with reasonable statistics. In order to estimate the background fraction when using a neural network cut, other methods had to be developed, they will be presented in section 7.3 and section 7.4 and consist of templates fit to the data.

7.2.1 Exact Method

One should iterate the 4-sector method until the fractions in the signal region become stable. If one assumes that the variables $\not\!\!E_T$ and Isolation are uncorrelated, one can also compute directly the signal and background content in the region.
In this thesis, the following definitions are used:

- Region A: number of events in the region $\not\!\!E_T < 15$ and Isolation > 0.2 (0.12 for plug electrons)
- Region B: number of events in the region $\not\!\!E_T > 20$ and Isolation > 0.2 (0.12 for plug electrons)
- Region C: number of events in the region $\not\!\!E_T < 15$ and Isolation < 0.1
- Region D: number of events in the region $\not\!\!E_T > 20$ and Isolation < 0.1 (Signal region)
- N_i the number of data events in region i (uncorrected)
- M_i the number of Monte Carlo events in region i (unweighted)
- S_D and B_D are the number of signal resp. background events in the signal region D. $S_D + B_D = N_D$.

One can see that

$$B_D = \frac{N_C - M_C \frac{S_D}{M_D}}{N_A - M_A \frac{S_D}{M_D}} \left(N_B - M_B \frac{S_D}{M_D} \right)$$

With $B_D = N_D - S_D$, one obtains a quadratic equation in S_D which can be resolved and gives two solutions:

$$S_D = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

with

$$a = M_B M_C - M_A M_D$$
$$b = N_A M_D^2 - N_C M_B M_D - N_B M_C M_D + N_D M_A M_D$$
$$c = N_B N_C M_D^2 - N_A N_D M_D^2$$

The unphysical solution is eliminated.

7.2.2 Results Obtained with the 4-Sector Method

In table 7.2, for jet multiplicities zero to three in the `ptop00` sample, the number of events in the four regions is counted. From this number, the background contamination in sector D using the three variants of the 4-sector method is extracted. For the sake of completeness, the number of scaled Monte Carlo events subtracted in the second iteration of the MC-corrected 4-sector method is also indicated.

	jet 0	jet 1	jet 2	jet 3
Data: A	316	507	92	22
Data: B	155	110	41	9
Data: C	2209	2590	560	93
Data: D	30240	4149	921	174
MC: A	19	16	4	0
MC: B	1170	255	63	6
MC: C	2285	2620	443	61
MC: D	196818	33704	5850	828
MC (scaled): A	2.81	1.7	0.46	0
MC (scaled): B	173.3	27.1	7.2	0.98
MC (scaled): C	338.5	279	50.8	10
MC (scaled): D	29156	3587	671	135.9
Background content in sector D with different variants:				
uncor. 4-sector	1083.5	562	249.6	38
uncor. 4-sector (%)	3.58%	13.5%	27%	21.9%
MC-cor. 4-sector	-109.4	379	187.8	30.2
MC-cor. 4-sector (%)	-0.36%	9.1%	20.4%	17.4%
Exact	-152.2	370	182	29.8
Exact (%)	-0.5%	8.9%	19.8%	17.1%

Table 7.2: Background estimation obtained with the different variants of the 4-sector method.(Number of events in the respective sector unless indicated otherwise).

Between the uncorrected 4-sector method and the MC-corrected 4-Sector method, large differences can be seen. As the MC-corrected 4-sector method is more realistic, it is used for all background estimation in W+jets analyses. No noticeable difference is seen, however, between the MC-corrected 4-sector method and the exact calculation. The number of estimated background events in the $W + 0$-jets bin is negative for the 4-sector method and the exact method. This cannot be explained at this point: in section 7.3.1, the reader can find an explanation for the negative values which refers to figure 7.3.
The difference between the MC-corrected 4-sector method and the exact method is very small compared to the 25% systematic error of the method [91].

7.2.3 Consistency Check

In order to control the validity of a method, one must test it with known input and check whether the output matches the input. In the case of the 4-sector method, I select only background events and apply the method on them. The method should then return 100% background content. As no Monte Carlo sample is available to simulate the QCD background, I have to take the background events from data. The background selection cut is an inverted cut on the output of the neural network: NN output < -0.85. I then apply the same procedure as that described before on the three variants of the method. The Monte Carlo simulation of signal events is identical to the one described above. One can object that the neural network cut implicitly changes the isolation distribution. This should, however, not affect the results: one of the hypothesis of the method is that isolation and $\not\!\!E_T$ are not correlated, hence the neural network cut affects sector A and sector B on the one hand and sector C and D on the other hand in an equal manner. As the method consists of forming the ratios N_C/N_A and N_D/N_B respectively, the method should work even if an impact from the neural network cut on the isolation distribution exists. Of course, the statistics will be lower but still enough to perform a consistency check.

	jet 0	jet 1	jet 2	jet 3
Data: A	105	239	49	13
Data: B	22	50	21	4
Data: C	101	164	38	9
Data: D	53	84	28	9
Background estimation in sector D:				
uncorr. 4-sector	40%	40.8%	58.2%	31%
MC-corr. 4-sector	39.4%	39.5%	56.4%	29%
Exact	39.4%	39.5%	56.5%	29%

Table 7.3: Consistency check for the different variants of the 4-sector method. The number of data events is given in row 2-5. The number of simulated events are the same as the (unscaled) simulated events in table 7.2. The last three lines show the background fraction obtained by the different variants of the 4-sector method. The background fractions should all be 100% because only background events are selected for this consistency check.

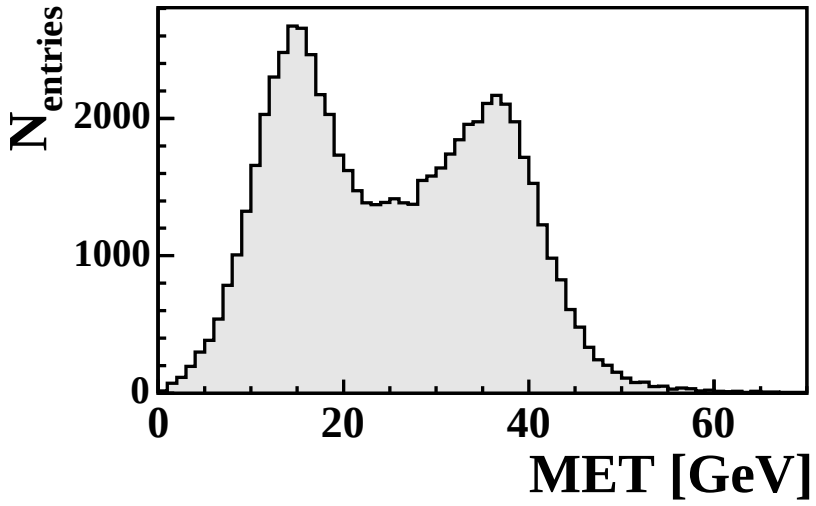

Figure 7.2: The distribution of $\not\!\!E_T$ for all events in the `ptop00` data sample before any cut. Two different components are clearly visible: a background component peaking at $\not\!\!E_T \sim 15$ GeV and a signal component peaking at $\not\!\!E_T \sim 37$ GeV.

Table 7.3 shows the background fraction obtained using the different variants of the 4-sector method. No large difference is seen between the three methods. The computed fraction of background events in region D differs significantly from the expectation of 100%. This can have two reasons: the background sample used for cross-checking still has signal contamination. The other reason would be that the method itself is not self-consistent but systematically underestimates the background content. At this point, no decision can be made whether the method itself or the background sample is invalid. A good Monte Carlo description of the background events could resolve this open question, but this is out of the scope of this thesis.
Another point has been seen in 7.2.2, table 7.2. The method should always produce a background fraction between zero and 100%. This is, however, not the case for the $W + 0$-jets bin, where a negative number of background events is found. This is clearly unphysical and can only be explained by a bad background model. A method relying on a fit would have the advantage that it gives both a background fraction and a consistency check via the χ^2 of the fit or other methods.

7.3 $\not\!\!E_T$ Fit-Method

Another method of estimating the background content is the $\not\!\!E_T$ fit method. The basic idea behind this method is that two components are in the data: signal and background. This is clearly visible in figure 7.2: the signal component peaks at $\not\!\!E_T \approx 37$ GeV and the background component peaks at $\not\!\!E_T \approx 15$ GeV.
The signal and background template used in the fit are derived from different sub-

samples. Their modeling is described in the following sections.
The fit determines a scale factor which represents the best fit of the two scaled templates to the data. The fit is performed in a restricted $\not\!\!E_T$ range. After this, the events are counted in the scaled background and signal templates which pass the cut $\not\!\!E_T > 20$ GeV, which is a requirement for the single-top quark search and other W+jet analyses. This method was employed for $W \to e\nu$ with electrons in the central region in reference [92]

7.3.1 Background Template

As no Monte Carlo description for background is available, I use again the data sample `ptop00` to model the background. Instead of performing the standard CDF cuts, I invert the cut on the neural network output. In figure 7.3, the shapes are shown for cuts on NN output < -0.8, NN output < -0.85, and NN output < -0.9. Additionally, a template is shown in which the the standard CDF cuts are performed with the exception that the isolation cut is inverted (Isolation > 0.1). This would be equivalent to the background modeling of the 4-sector method.

One can see that the shapes of the distributions obtained by the different cuts on the neural network are identical in the limits of statistics. Of course, the number of events decreases with harder cuts. Except for the $W + 0$-jets bin, the inverted cut is also very similar to the neural network distributions. The number of events in this distribution is, however, much smaller than in the neural network distributions. The inverted isolation background model is inadequate for the $W + 0$-jets bin, as one can see from the peak at around 30 GeV. This is an indication of signal content in this region, which is also the explanation of the negative number of background events given by the MC-corrected 4-sector method and the exact method for this jet bin: these signal events are in sector C. They affect the weighting of the simulated events such that too much signal is subtracted from the other regions, and the background estimate turns into negative.
In the following, I will choose the background template obtained with the cut on the neural network output < -0.85. This cut has an appropriate purity and still enough statistics to perform a fit with. The distribution has the expected shape from QCD, folded with the typical acceptance function of the missing transverse energy trigger. The reason for values in $\not\!\!E_T$ smaller than the trigger threshold of 15 GeV is the following: one does not trigger on the $\not\!\!E_T$ used for analyses and corrected for muon energy or jet energy, but on the raw missing transverse energy. In an analysis like the single-top quark analysis, one would cut on the corrected $\not\!\!E_T > 20$ GeV.

7.3.2 Signal Template

The signal template is derived from Monte Carlo events, the `wtop1i`, `atop5a` and `atopaa` samples are used. In figure 7.4, the $\not\!\!E_T$ distribution for the jet multiplicities zero to three are shown. In the figure, no cut on the neural network or the CDF

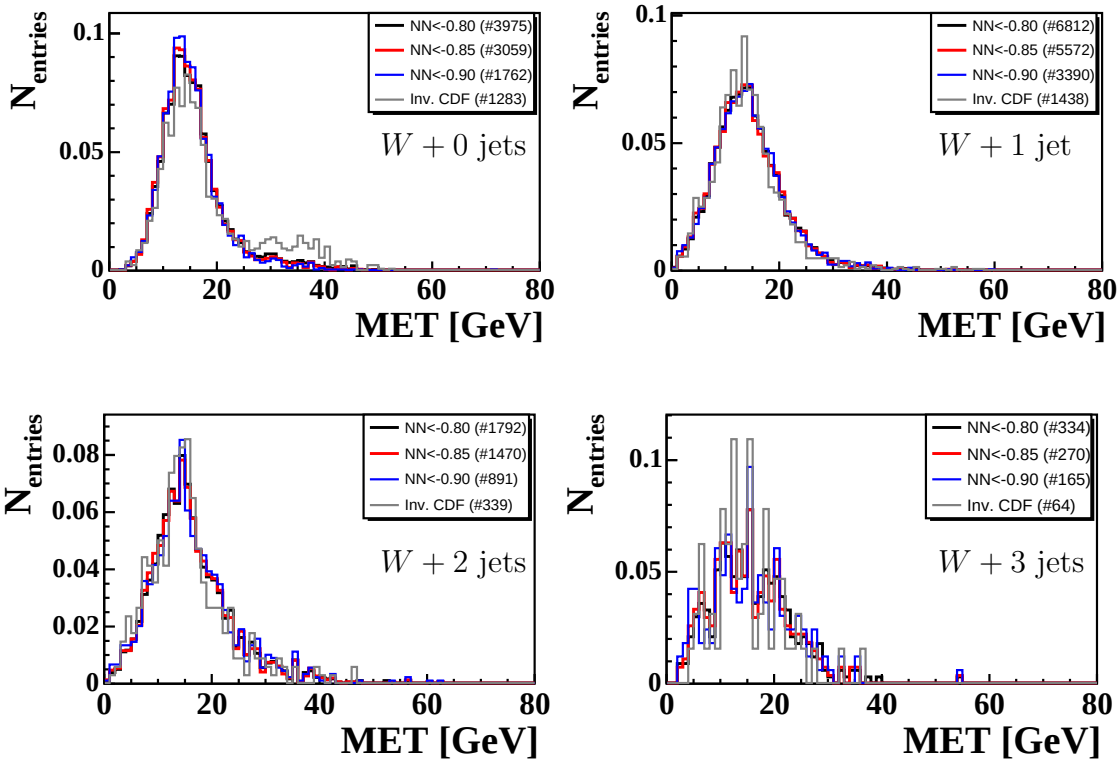

Figure 7.3: $\not\!\!E_T$ background templates obtained from data for different jet multiplicities and different cuts normalized to unit area. The number of events contributing to one distribution is given in the legend as well as the respective cut.

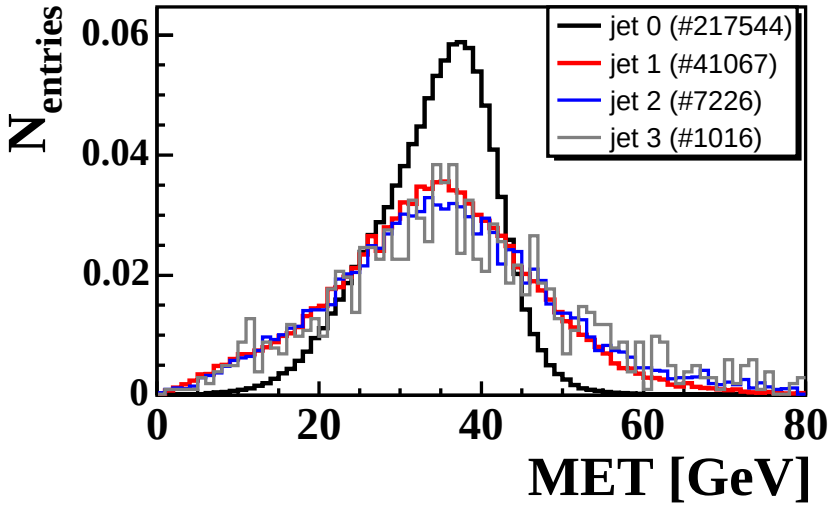

Figure 7.4: Monte Carlo samples for different jet multiplicities, normalized to unit area. The number of events for each jet multiplicity is given in the legend of the plot. No identification cuts are performed in this plot.

cuts is done. For the fits, the same cut is performed on the signal template that for the data sample.

One can see that the the $W + 0$-jets distribution is different from the three other distributions, which are essentially identical in the limit of statistics. Especially in the $W + 0$-jets distribution, one can see that the neutrino takes away half of the energy of the W boson, as the $\not\!\!E_T$ distribution peaks at 40 GeV, half of the W mass. With jet activity in the event, this argumentation still holds but the measured missing energy distribution has also a component from mismeasured jets or neutrinos from semileptonic decays of B-Hadrons. This smears out the distribution.

7.3.3 Data Distribution

The data is derived from the `ptop00` data sample. The cut scenarios applied on the sample vary: the CDF cuts have been performed as well as different cuts on the output of the neural network. In figure 7.5, different scenarios are compared. For the cuts on the neural network output, three scenarios are exemplified in this figure: NN > -0.80, NN > -0.65 and NN > -0.50. All of them are approximately as hard as the CDF cuts, which are also shown in the figure. In the subsequent analysis, 50 different cuts have been made on the output of the neural network from -1 to 1.

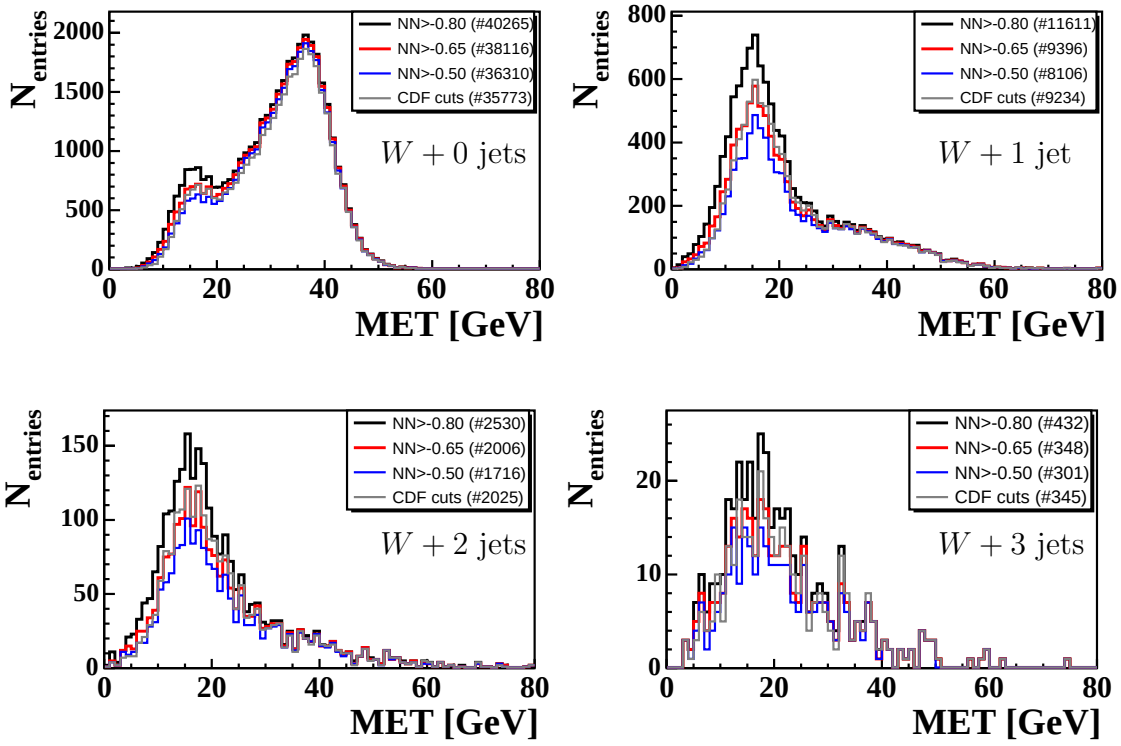

Figure 7.5: Distributions of the data for different jet multiplicities and different cut scenarios. One can see that by varying the network cut the background contribution also varies. With the same signal as for the CDF cuts the background contribution is lower for network cuts.

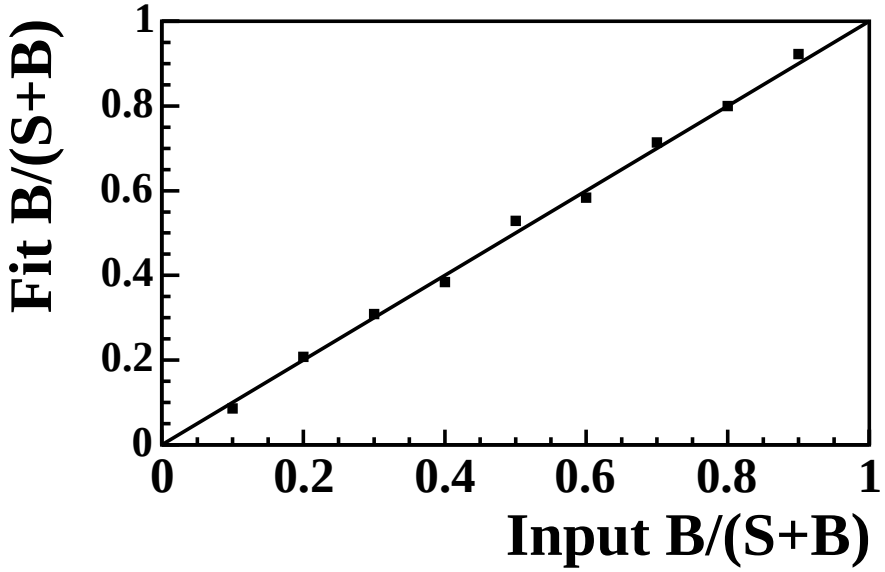

Figure 7.6: Consistency check for the $\not\!\!E_T$ fit method, exemplified on W + 2-jets events. A test sample is made by adding up signal and background events for a given background fraction. The fit results are compared to different input background fractions: one would expect these to be equal, i.e. the points lying on the diagonal line. The fit range for this fit was $15 < \not\!\!E_T[\text{GeV}] < 40$

.

7.3.4 Fit Method

For each cut scenario, the $\not\!\!E_T$ distribution for data and the two templates is prepared. Using a standard likelihood fit using Poisson statistics in which the template predictions are also varied within statistics [93], a scaling factor for the background and the signal template is obtained. The method then counts the events in these scaled distribution with $\not\!\!E_T > 20$ GeV and gets the number of signal and background events from which one can then derive other quantities like purity or efficiency.
One important point is that the fit is not done over the whole $\not\!\!E_T$ spectrum but only in a restricted range. This range is determined from a consistency check of the method. In this consistency check, a test sample is made by adding up background and signal events in a predefined fraction. The fit range is varied and the range is chosen for which the difference between the input background fraction and the output background fraction given by the fit method is smallest. Figure 7.6 shows the output background fraction given by the fit versus the input background fraction for events with jet multiplicity equal to two. One can see that all points are very well on the diagonal line which proves that the method performs well. This consistency check is also done for other jet multiplicities, the deviations from the diagonal line are similar to those of the W + 2-jets events when using the following fit ranges:

- W + 0 jets events: 15 GeV $< \not\!\!E_T <$ 35 GeV.
- W + 1 jet events: 15 GeV $< \not\!\!E_T <$ 40 GeV.

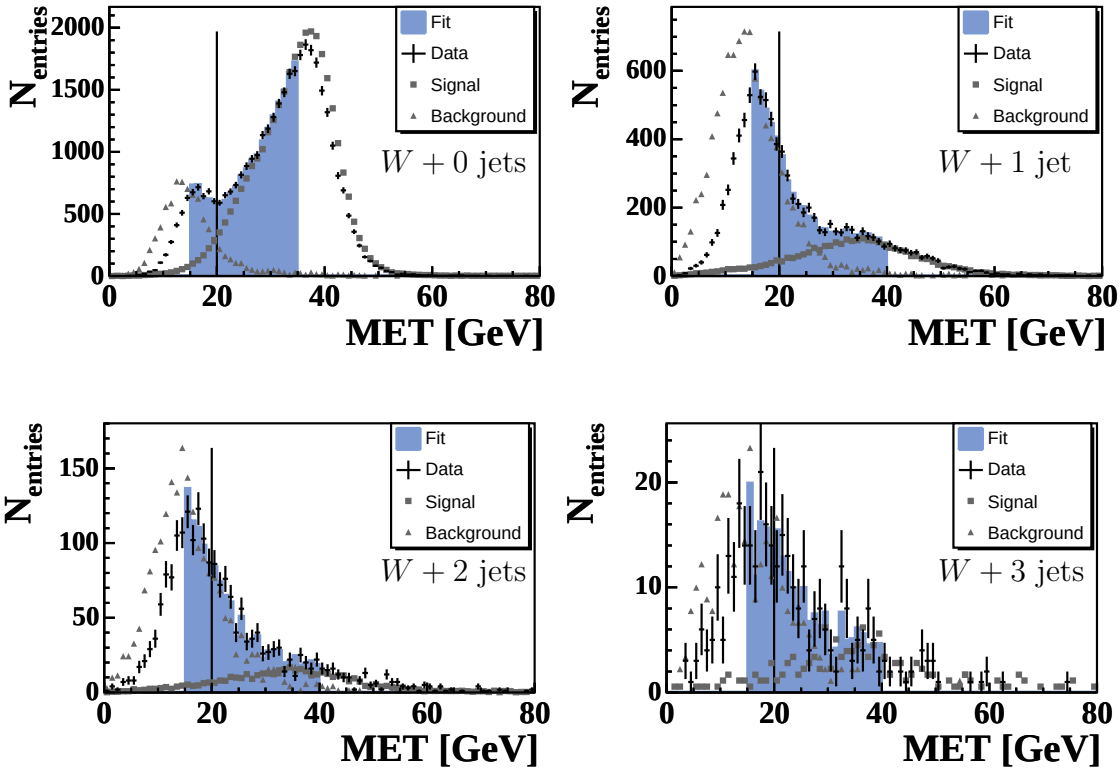

Figure 7.7: Background template (triangles) and signal template (boxes) fit to the data (crosses). The fit region and the fit result is shaded. Different jet multiplicities are fitted, signal and data are obtained by performing the CDF cuts. The vertical line indicates the cut which is performed to count signal and background events.

- $W + 2$ jets events: 15 GeV $< \not\!\!E_T <$ 40 GeV.
- $W + 3$ jets events: 15 GeV $< \not\!\!E_T <$ 40 GeV.

7.3.5 $\not\!\!E_T$ Fit Results

The $\not\!\!E_T$ fit is performed for jet multiplicities 0, 1, 2 and 3. Figure 7.7 shows the fit with data and signal obtained by performing the CDF cuts. For 50 different cuts on the neural network output, signal and background content is also obtained by similar fits. These fits are not shown, instead, the signal and background fraction obtained with each cut on the network output is shown and compared to the fractions obtained with the CDF cuts.

For 50 cuts on the neural network output and the CDF cuts, the fit is done for every jet multiplicity considered. Different quantities like $S/S + B$ which will help to determine the optimal cut on the network output for a certain analysis. The first distribution the reader might want to have a look at is the number of events in the background sample versus the number of events in the signal sample in figure 7.8. The points in the lower left part result from hard cuts on the neural network output

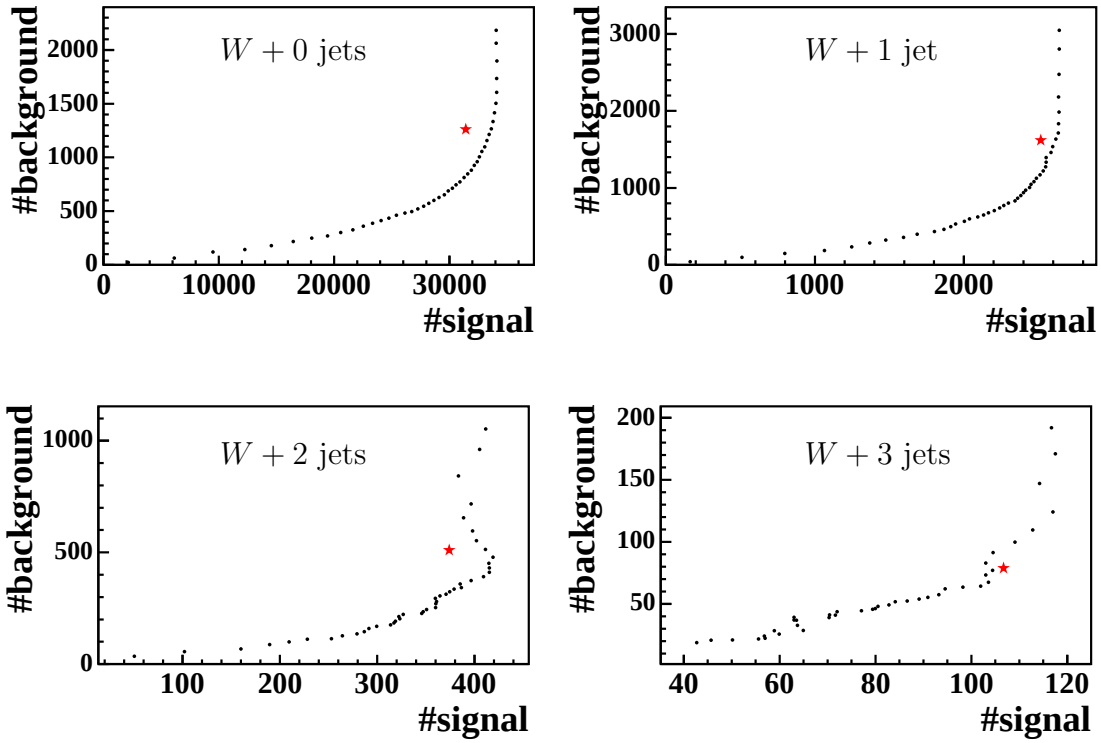

Figure 7.8: Number of background events versus number of signal events obtained by fitting distributions with different cuts on the neural network output (dots), compared to the CDF cuts (star). The cut on the network output is softer in the upper right corner than in the lower left corner.

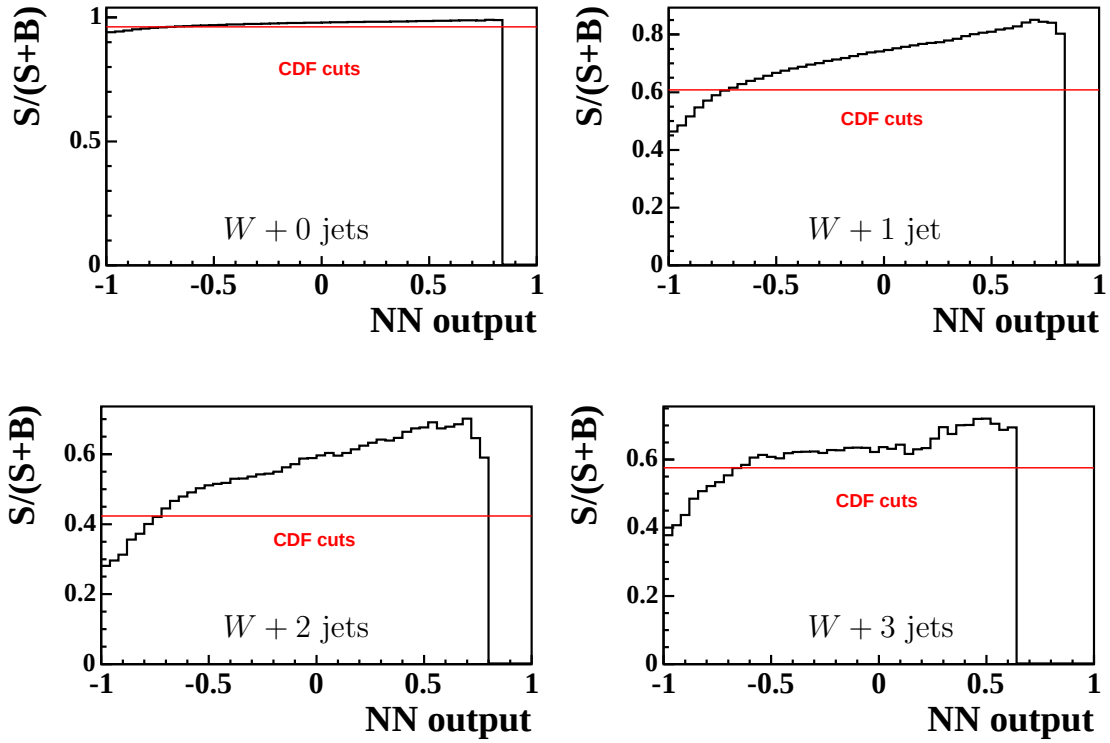

Figure 7.9: Purity $S/(S+B)$ as a function of the cut on the network output. As a comparison, the purity obtained by the CDF cuts is indicated as a red line.

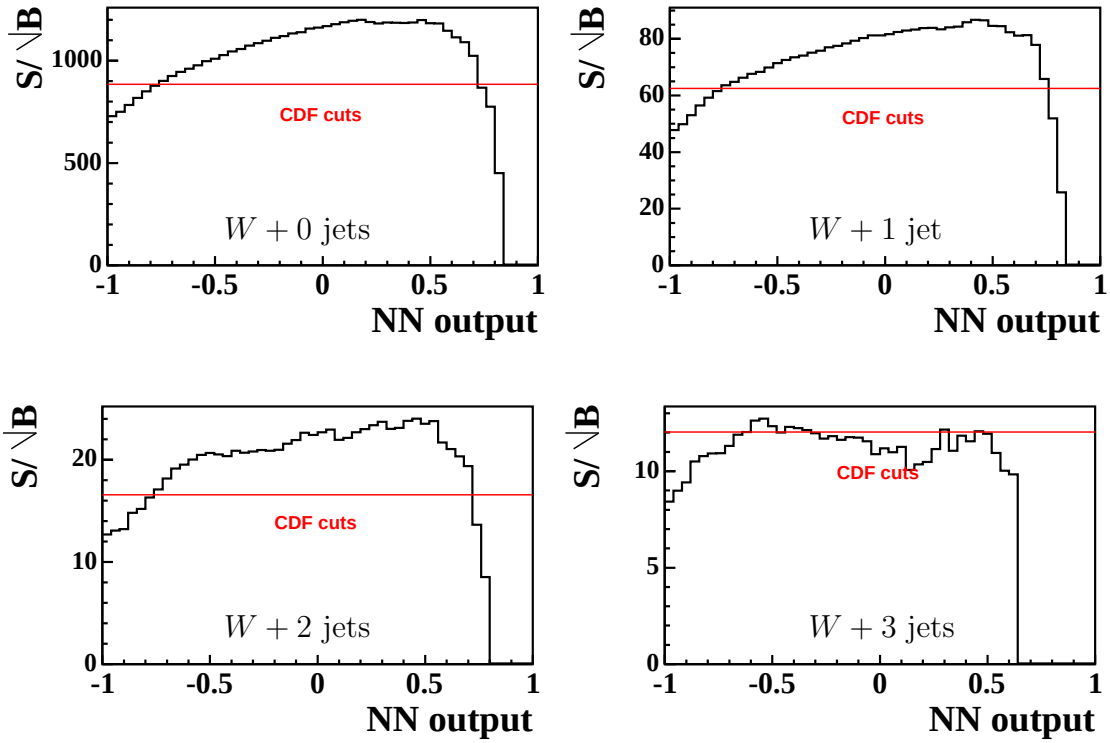

Figure 7.10: The significance $\sigma = S/\sqrt{B}$ as a function of the cut on the network output. As a comparison, the significance obtained by the CDF cuts is indicated as a red line.

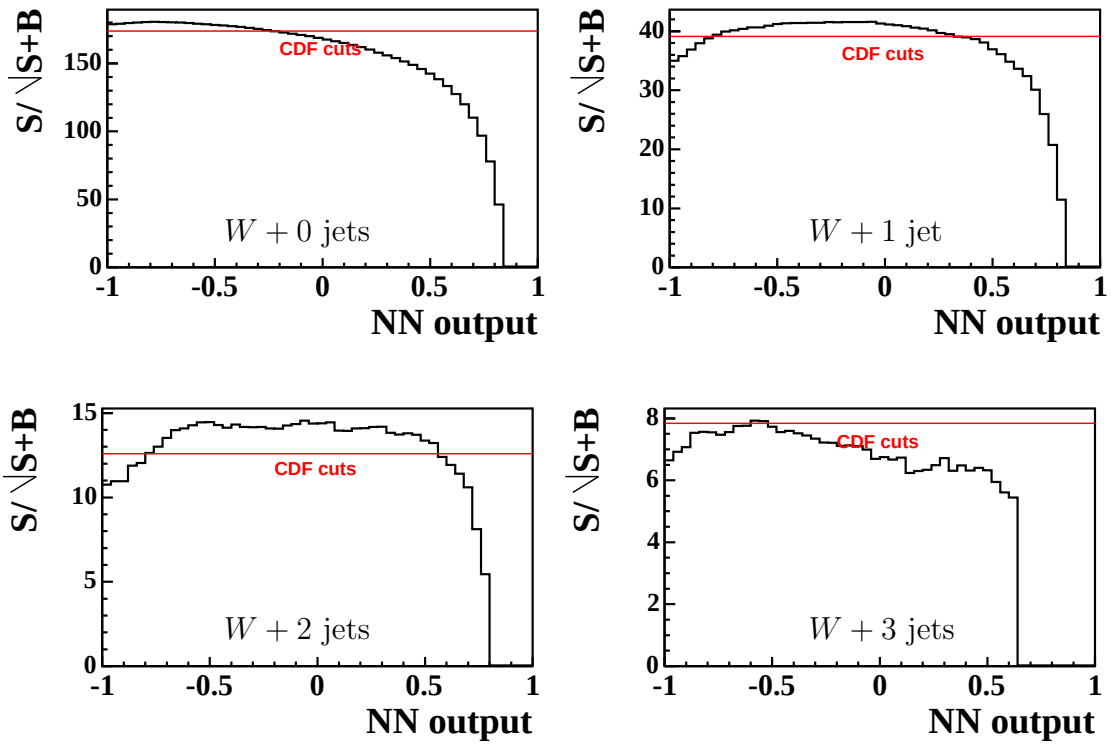

Figure 7.11: The significance $\sigma = S/\sqrt{S+B}$ as a function of the cut on the network output. As a comparison, the significance obtained by the CDF cuts is indicated as a red line.

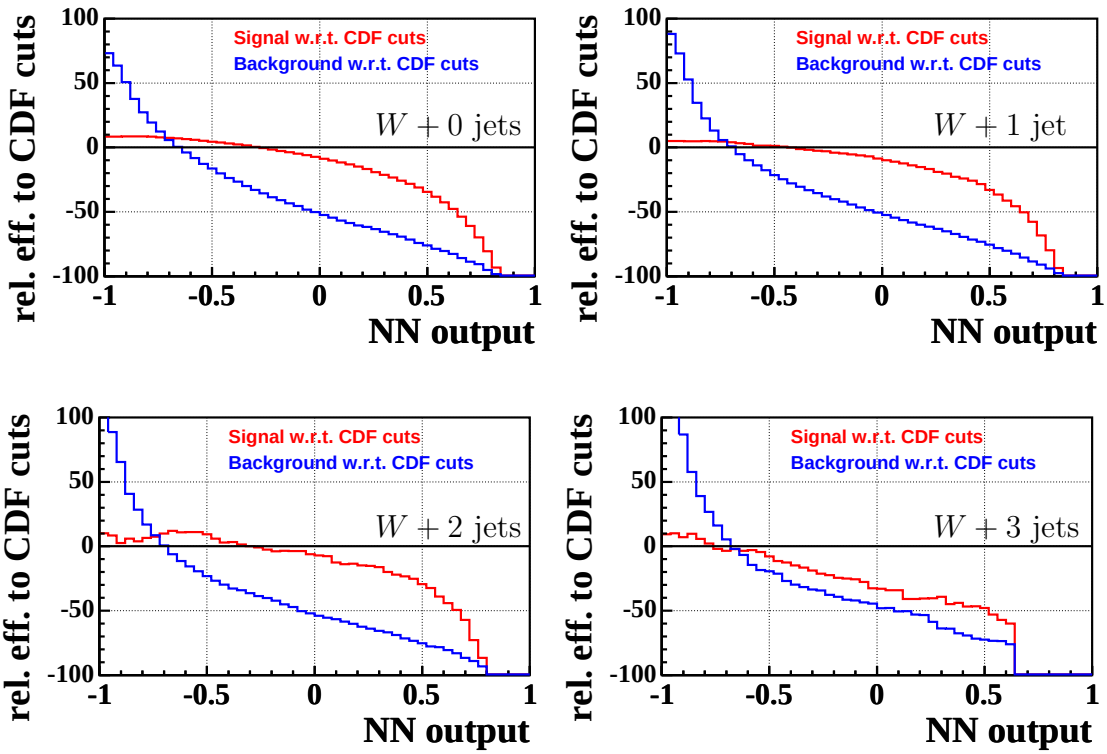

Figure 7.12: The relative difference between the number of signal (background) events obtained by a certain cut on the network output and the number of signal (background) events obtained by the CDF cuts as a function of the cut on the network. The blue line is the signal difference, the red line the background difference.

(close to 1) whereas the points in the upper right part of the distribution result from softer cuts on the network output (closer to -1). One can see that W+0-jets and W+1-jet are modeled very well. W+2-jets is starting to be problematic due to low statistics while the W+3-jets plots is problematic due to low statistics. This also translates into the purity distributions in figure 7.9. One can see that the neural network cuts perform better over a large range than the CDF cuts except for $W+3$-jets. The two definitions of significance, $\sigma = S/\sqrt{B}$ and $\sigma = S/\sqrt{S+B}$ are plotted in figure 7.10 and 7.11 respectively. Depending on the kind of analysis, one might want to optimize one or the other quantity. The last analyzed quantity is shown in figure 7.12, which shows the relative difference between the number of signal (resp. background) events obtained with a certain network cut and the number of signal (resp. background) events obtained with the CDF cuts. Except for effects due to the lower number of events with a higher jet multiplicity, the shapes of the distributions are identical for $W+0$-jets, $W+1$-jet, and $W+2$-jets events. The shape of the $W+3$-jets events look different, as no or only very little gain in signal can be obtained. However, the fluctuations in the distribution itself are so large that this could also be a statistical effect.
The relative similarity of the shapes leads to the conclusion that one cut can be chosen for all multiplicities and with roughly the same efficiency and purity gain. This was not expected, because naively, one would expect a different performance for different jet multiplicities. The isolation of the electron is strongly correlated to the network output, and one could expect that in higher jet multiplicities the isolation for electrons is more "background-like". The independence of the jet multiplicity shows the robustness of the method.

7.4 *In-Situ* Fit Method

One of the drawbacks of the $\not\!\!E_T$ fit method is that for every single cut on the neural network output, a separate fit has to be done. This can only be done automatically, and the control over the fit is not guaranteed. Furthermore, when fitting jet multiplicities larger than two, the statistics are rather low and a fit is difficult. The fit itself is rather sensitive to the shape of the templates.
In the *In-Situ* fit-method, the output of the neural network from the data sample is fitted with a signal template and a background template. These templates are derived from the templates used for the training of the neural network, and are shown in chapter 6. The background sample is obtained from the simulated background events described in 6.6.
The data distribution is the output of the neural network of events having a given jet multiplicity. If needed for the analysis (as it is the case for the single-top analysis), a cut on $\not\!\!E_T > 20$ GeV is also performed. The templates remain the same for all jet multiplicities, so for higher jet multiplicities only the binning should be adapted. For jet multiplicities zero to three, the data is filled into a histogram with 50 bins. In my studies, I have seen that the *In-Situ* methods can even be applied to $W+4$-jets when the binning is reduced from 50 to 15 bins.
This method removes the disadvantages of the $\not\!\!E_T$ fit method: only two fits have

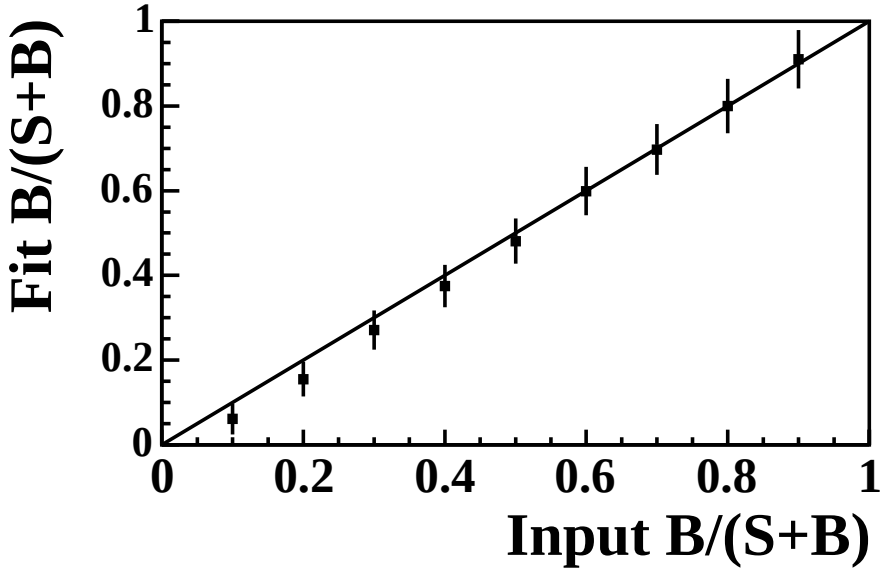

Figure 7.13: Consistency check for the *In-situ* fit method. A data sample was faked by adding up signal and background events in a given background fraction. The fit results are compared to different input background fractions, one would expect these to be equal, i.e. the points lying on the diagonal line, which is given in the limits of the error. The error is given by the fitting procedure.

to be done for each jet multiplicity, which can be well controlled. As the templates remain the same for all jet multiplicities, statistics in the templates is not a problem.

As done for the $\not\!\!E_T$ fit method, I also perform a consistency check for this method. From the signal and background templates, events are randomly selected with a variable background fraction. The fit is performed and the output background fraction from the fit compared to the input value. One can see in figure 7.13 that the input and the output values agree very well in the limit of the fit errors.

To compute the background contamination in the data sample obtained with a given cut on the network output, one has to integrate the weighted templates and can compute quantities like $S/(S+B)$, $S/\sqrt{B}$ or $S/\sqrt{S+B}$. With this method, a consistent comparison with the background fraction obtained after performing the standard CDF cuts is possible. The templates are modified in such a way that only events which pass the CDF cuts constitute the templates. The standard CDF cuts are also applied on the data. With the same fit method, the neural network output of the data is fitted, and the background and signal fraction comes directly out of the respective weights.
One can then compute for every jet multiplicity the relative difference of signal and background at a given neural network cut compared to the standard CDF cuts. The fits are presented in figure 7.14 and figure 7.15 for the fits with the CDF cuts. The number of background events versus the number of signal events in data is plotted

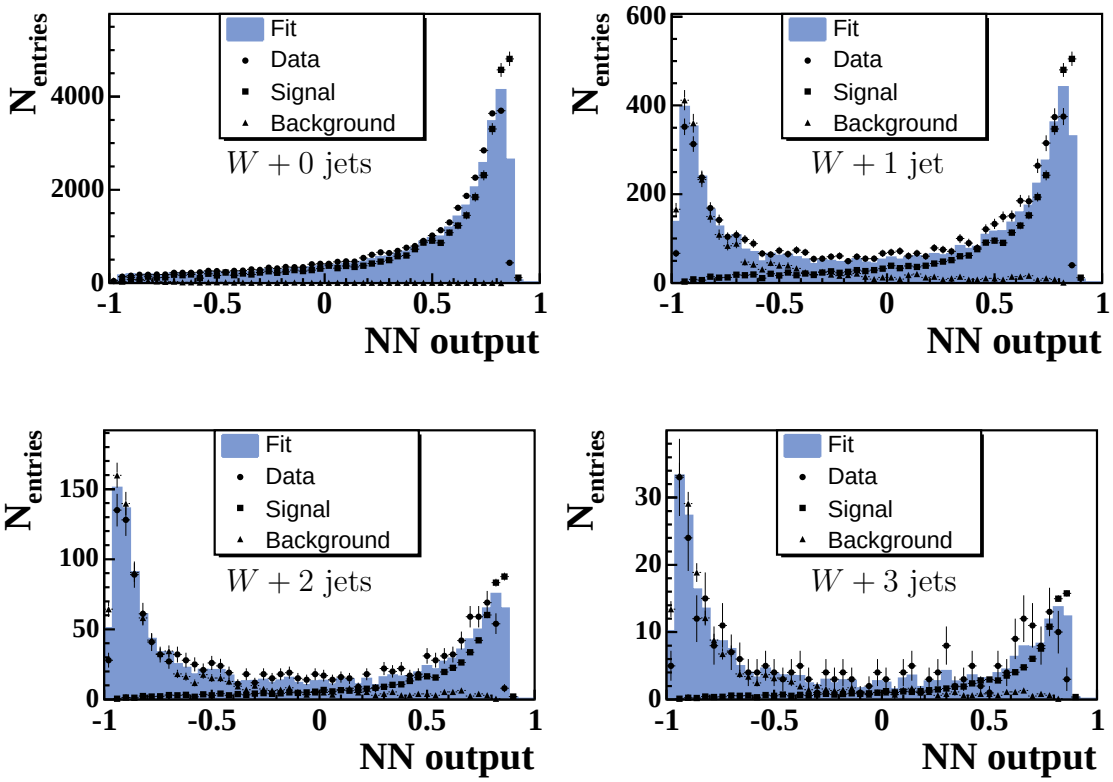

Figure 7.14: Background sample (triangles) and signal sample (squares) fit to data (dots). The fit result is shaded. From upper left to lower right plot: jet multiplicities from zero to three. Only a cut on $\not\!\!E_T > 20$ GeV has been performed.

in figure 7.16. The points in the lower left part result from harder cuts on the neural network output while the points in the upper right part result from softer cuts on the network output (closer to -1).

One can see in the $S/(S+B)$ in figure 7.17, $S/\sqrt{B}$ in figure 7.18, $S\sqrt{S+B}$ in figure 7.19 or the relative difference in figure 7.20 to the standard cuts distribution that the shapes are almost identical for each jet multiplicity. This shows that the neural network method is rather stable: naively, one would expect that the results differ with jet multiplicities getting higher: one of the important variables, the relative isolation, tends to slightly larger values as jets tend to spray energy in the electron region. As the jet multiplicity hardly influences the shape of the variables, one can choose a cut on the neural network output independent of the jet multiplicity. The cut value must be chosen appropriate to the analysis.
The quantities $S/(S+B)$, $S/\sqrt{B}$, $S/\sqrt{S+B}$ are of a large interest when one uses them to determine a cut for an analysis in which the selected sample is not subject to any further cuts. This would be the case for the measurement of the W mass e.g.. The search for electroweak top quark production follows a different strategy: further cuts are performed after the electron identification. It is therefore mandatory to use as many events as possible without selecting additional background. This is ensured

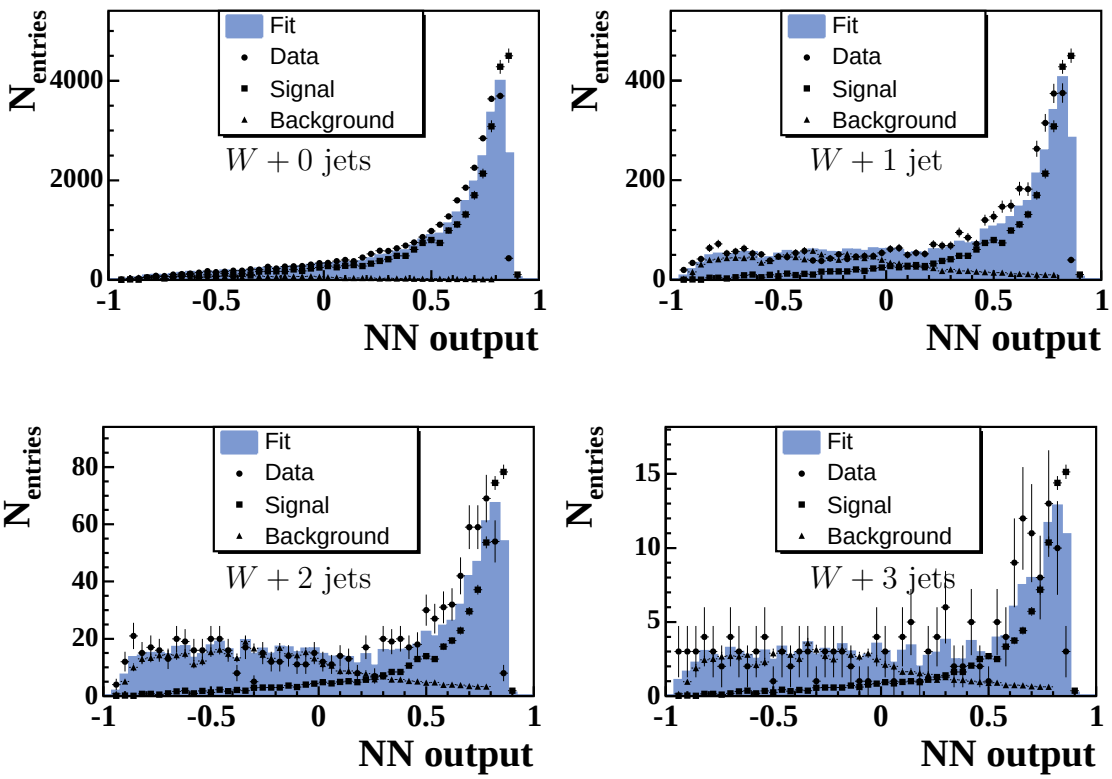

Figure 7.15: Background sample (triangles) and signal sample (squares) fit to data (dots). The fit result is shaded. From upper left to lower right plot: jet multiplicities from zero to three. A cut on $\not\!\!E_T > 20$ GeV and the CDF cuts have been performed.

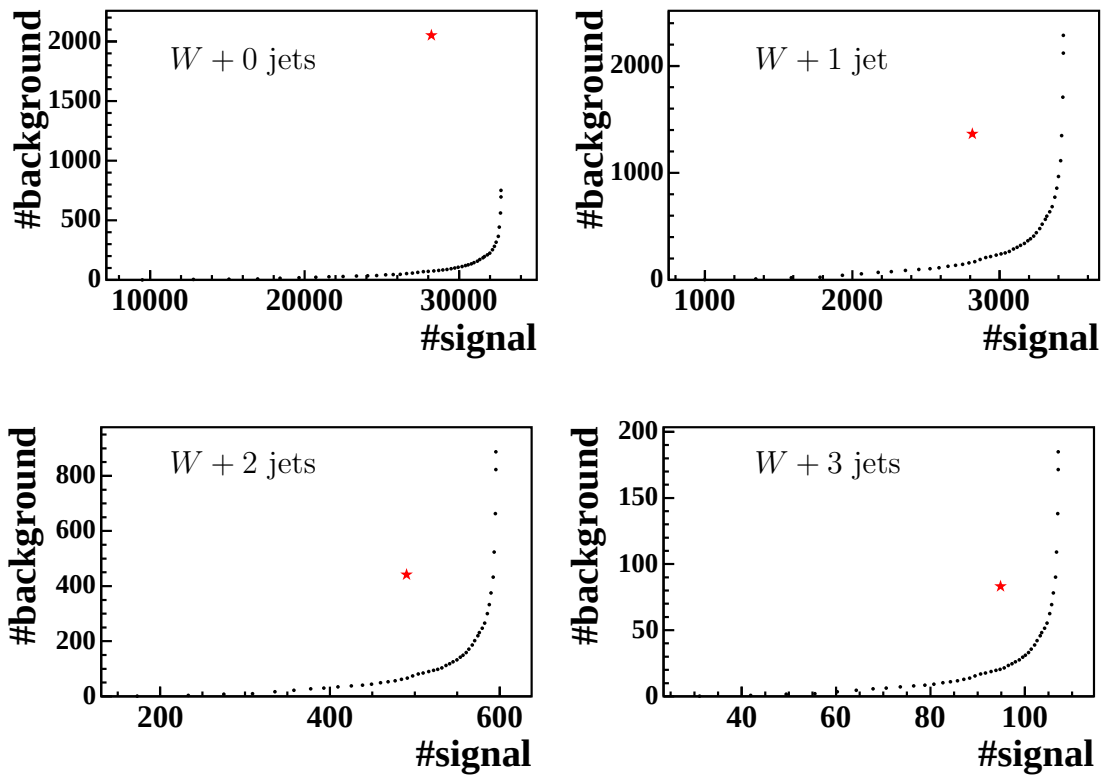

Figure 7.16: Number of background events versus the number of signal events for different cuts on the network output (dots) and the CDF cuts (star). From upper left to lower right plot: jet multiplicities from zero to three.

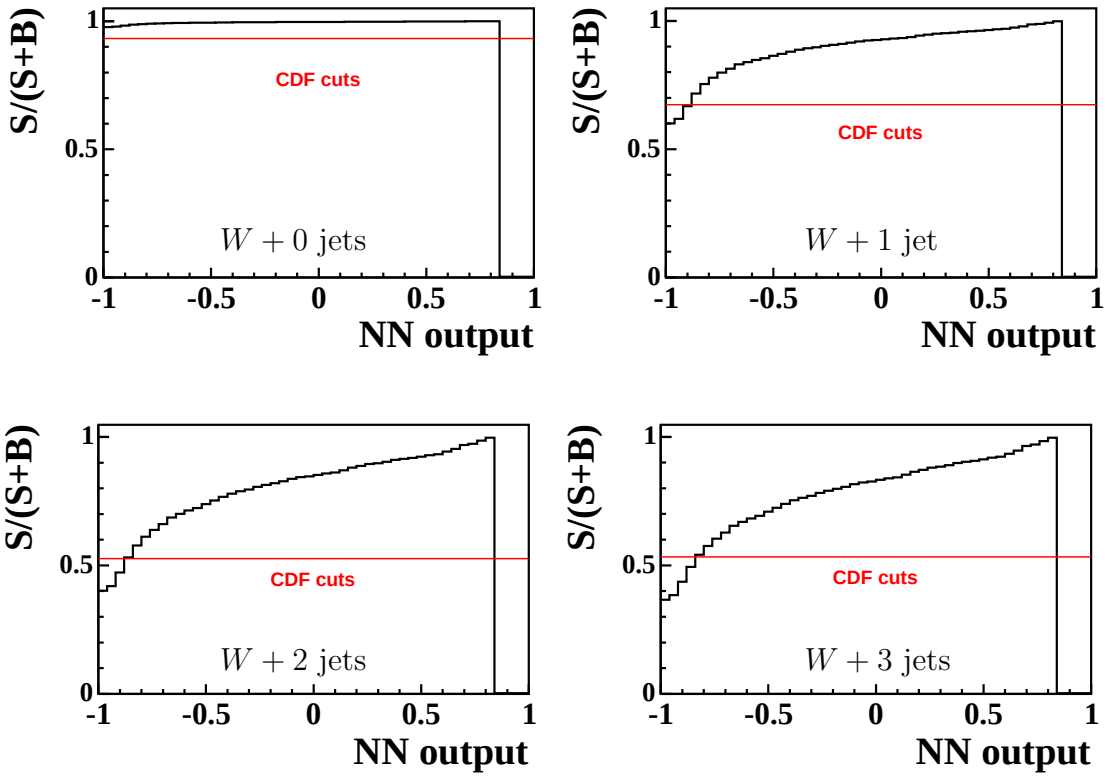

Figure 7.17: Purity $(S/(S+B))$ as a function of the NN cut. The red line is a comparison with the purity obtained with the standard cuts. From upper left to lower right plot: jet multiplicities from zero to three.

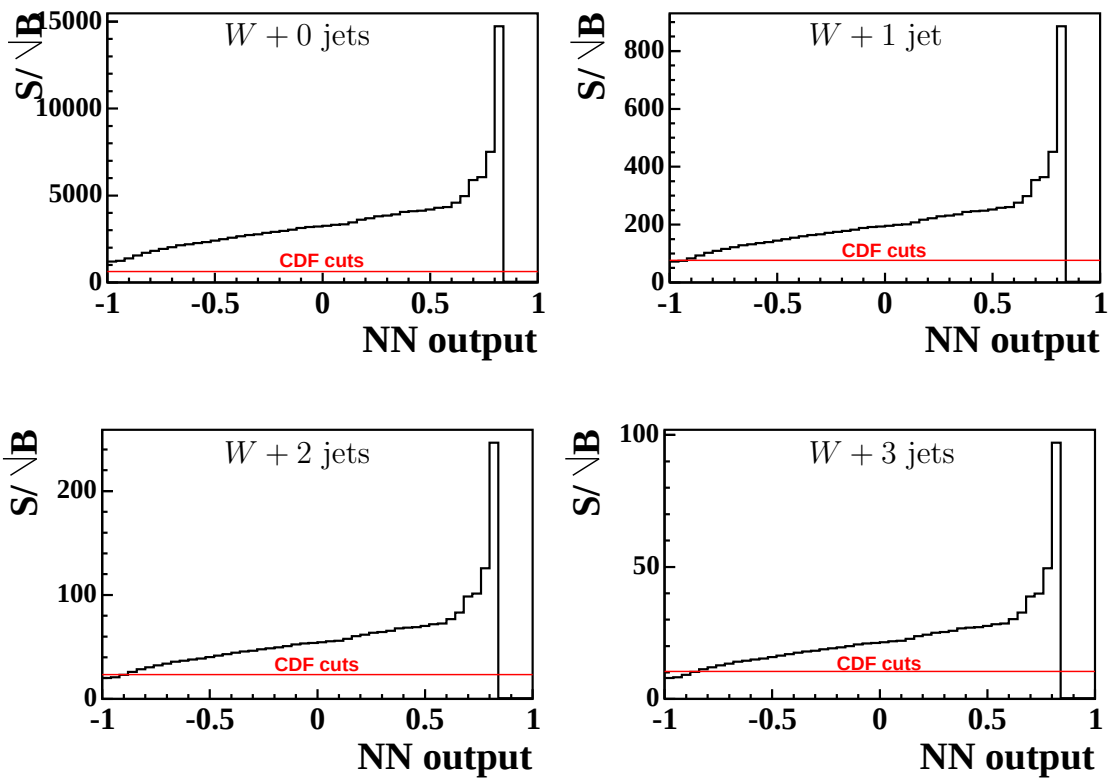

Figure 7.18: Sigma ($S/\sqrt{B}$) as a function of the NN cut. The red line is a comparison with the sigma obtained with the standard cuts. From upper left to lower right plot: jet multiplicities from zero to three.

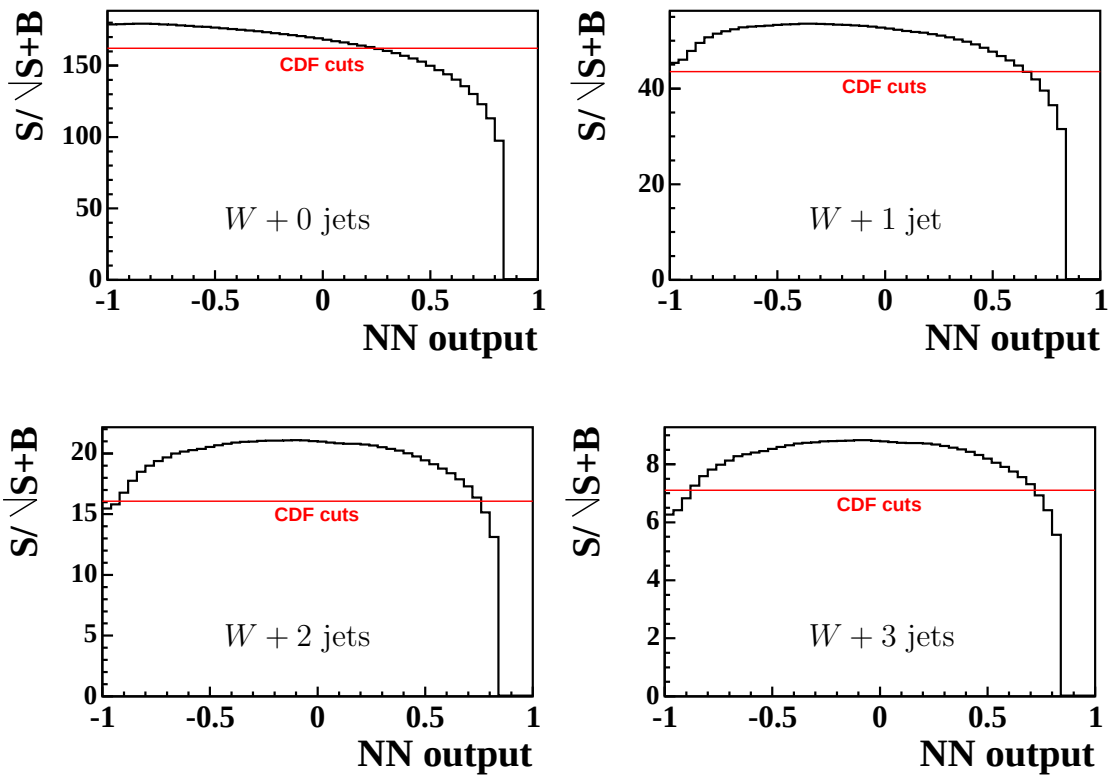

Figure 7.19: $S/\sqrt{S+B}$ as a function of the NN cut. This quantity reflects better than sigma that one wants to optimize also the number of signal events. The red line is a comparison with the standard cuts. From upper left to lower right plot: jet multiplicities from zero to three.

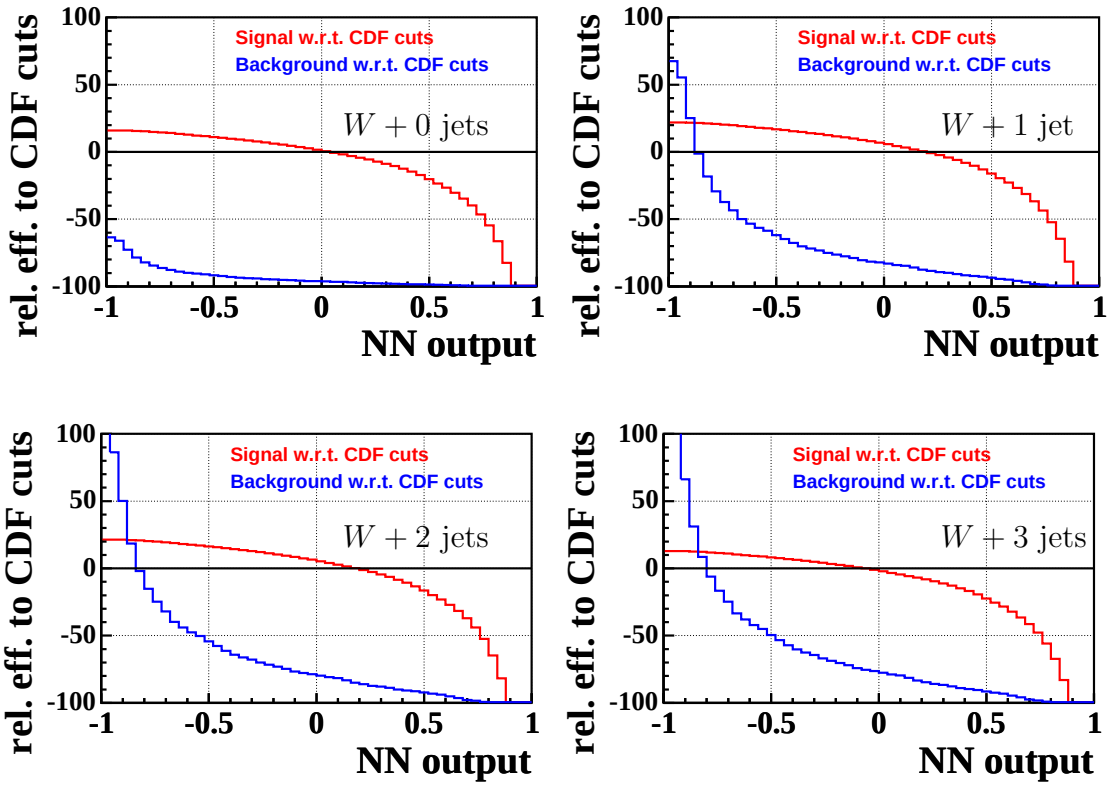

Figure 7.20: For each NN cut, the relative difference to the signal efficiency (red curve) and background efficiency (blue curve). From upper left to lower right plot: jet multiplicities from zero to three.

when making a cut around -0.6 and -0.5. A little increase in signal with a decrease in background events compares to the standard cuts is the result. Table 7.4 details several cut scenarios and compares them to the CDF standard scenario.

7.5 Summary of Different Background Estimation Methods

	jet 0	jet 1	jet 2	jet 3
uncorr. 4-sector (%)	3.58	13.5	27	21.9
MC-corr. 4-sector (%)	-0.36	9.1	20.4	17.4
Exact (%)	-0.5	8.9	19.8	17.1
$\not\!\!E_T$ fit: CDF (%)	3.86	39.17	57.68	42.45
$\not\!\!E_T$ fit: NN > -0.6 (%)	3.11	33.31	48.91	39.21
$\not\!\!E_T$ fit: NN > -0.5 (%)	3.43	35.72	51.51	40.52
In-Situ: CDF (%)	6.7	32.63	47.38	46.74
In-Situ: NN > -0.6 (%)	0.6	15.21	28.64	31.75
In-Situ: NN > -0.5 (%)	0.5	3.99	24.72	27.56

Table 7.4: Background estimation obtained with the different variants of the 4-sector method (lines 2-4). Background estimation for the CDF cuts and two possible network cut scenarios for the $\not\!\!E_T$ fit method (lines 5-7) and the *In-situ* fit method (lines 8-10).

Table 7.4 summarizes the results obtained with the different methods. One can see that the two fit methods agree reasonably well for their estimation of the background fraction in the different jet multiplicities when using the CDF cuts. The estimation is, however, significantly different from the estimation obtained with the 4-sector method. No final decision can be made at this point whether one method or another is the right description: further work must be done. It is of utmost importance to invest efforts in correcting the Monte Carlo description of the variables, and also to produce a Monte Carlo description of the QCD background processes.
The results also show that the use of a neural network can produce better results: it allows for better efficiency and for better purity.

7.6 Application to W+jets Analyses

7.6.1 Determination of W Boson Properties

The identification of electrons with a neural network can also be used for studies of the properties of the W boson such as the transverse mass. I will not present an in-depth analysis of W boson properties. This is out of the scope of this document. I will, however, show which improvements one can expect when using the neural network method for the electron identification.

The preselection I use is again different: no explicit cut on $\not\!\!E_T$ is required. As the trigger used is the `MET_PEM` trigger, events of lower $\not\!\!E_T$ are, however, less likely to appear in the sample. One could object that not to require a $\not\!\!E_T$ cut will mostly select QCD background. I will show that my identification method reduces this background better than the CDF sequential cuts. I do not distinguish between the jet multiplicities.

In order to have a comparison, I use the `wtop1i` W+jets Monte Carlo sample generated using Pythia. It is known that the Pythia Monte Carlo generator underestimates the recoil energy of the W boson. In his thesis, Hartmut Stadie [92] has corrected the $\not\!\!E_T$ distribution of the simulated events by scaling up the recoil energy vector by 5% and recomputing $\not\!\!E_T$. I do not apply this correction as it is only of importance for $W + 0$-jets. For all other jet multiplicities, the mismeasurement of $\not\!\!E_T$ due to mismeasured jets is more important than the contribution from the recoil energy. I have, however, applied an energy scale correction comparable to the one used in reference [85] to the electrons in the data sample.

Two quantities of importance are shown: the transverse mass of the W boson in figure 7.21 and its transverse momentum in figure 7.22. Quantities related to the longitudinal component of the W are experimentally not accessible, as the longitudinal part of the neutrino momentum cannot be measured at hadron colliders.

The transverse mass is defined as

$$M_T c = \sqrt{2\,(1 - \cos\Delta\varphi)\,P_{T,\text{electron}} P_{T,\text{neutrino}}}$$

$P_{T,\text{neutrino}}$ is the measured $\not\!\!E_T$ and $\Delta\varphi$ the difference of the angle between electron φ and the direction in φ of the missing transverse momentum.

Three cut scenarios are presented on both plots:

- The standard CDF selection for plug electrons. 47468 events are selected.
- The plug electron is selected using a cut on the neural network output > -0.4. With this cut, the signal height is approximately the same as for the CDF cuts. 44620 events remain after this selection. One can see that this selection has fewer events in the region $M_T \approx 40$ GeV/c^2, the region in which one expects

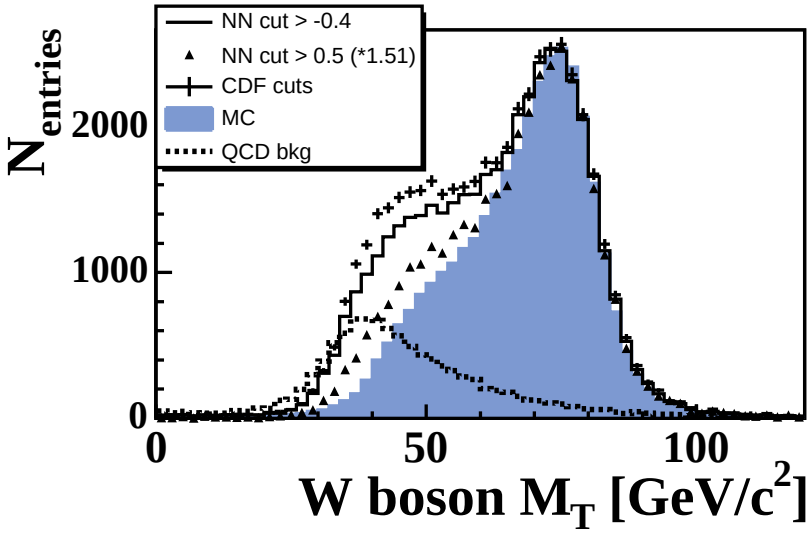

Figure 7.21: Transverse mass of the W boson for different cut scenarios. The CDF cuts are represented by crosses, a neural network cut > -0.4 is represented by a continuous line. A very hard cut on the network output > 0.5 is presented by triangles. This distribution is scaled by a factor 1.51 in order to be comparable with the other distributions. The distribution of the simulated events is shown as a shaded area. No cut on jet multiplicities is made. The shape of the QCD background distribution is indicated by a dotted line. The scale of its distribution is arbitrary.

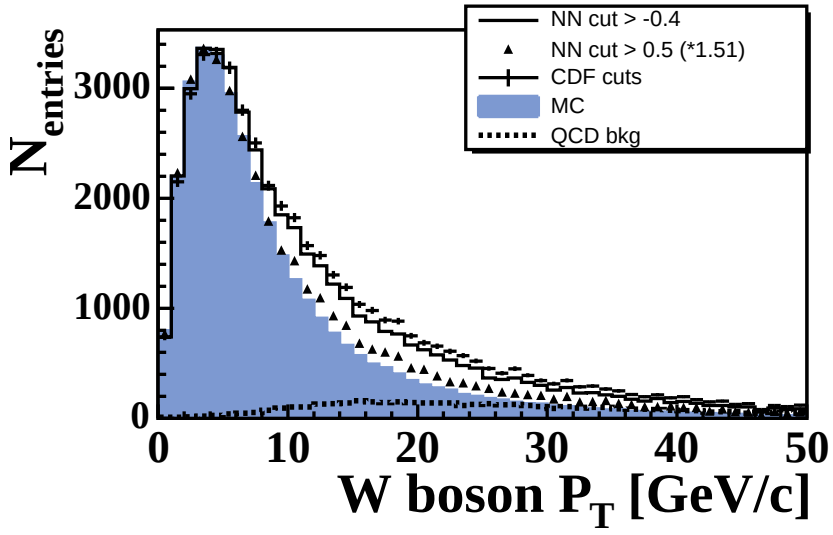

Figure 7.22: Transverse momentum of the W boson for different cut scenarios. The CDF cuts are represented by crosses, a neural network cut > -0.4 is represented by a continuous line. A very hard cut on the network output > 0.5 is presented by triangles. This distribution is scaled by a factor 1.51 in order to be comparable with the other distributions. The distribution of the simulated events is shown as a shaded area. No cut on jet multiplicities is made. The shape of the QCD background distribution is indicated by a dotted line. The scale of its distribution is arbitrary.

the background. On the P_T plot, one can see that the network distribution is below the CDF cuts distribution for larger values of P_T and its shape resembles more the shape of the distribution of the simulated events.

- The plug electron is selected using a cut on the neural network output > 0.5. This is a very hard cut which ensures high purity. As only 25180 events remain, the distribution is scaled by a factor 1.51 in order to better compare its shape with the other distributions. One can see that even fewer events are in the background region around $M_T \approx 40$ GeV/c^2, and the P_T distribution is very close to the one expected from simulated events.

In addition, the shape of the QCD background is indicated. This distribution was obtained by applying an inverted cut on the neural network output below -0.85. The height of the QCD background distribution is arbitrary and is not normalized to any signal distribution. One can see that this distribution peaks at $\not\!\!E_T$ values for which a discrepancy between simulated events and data events after, e.g. the standard cuts is seen. The excess in events can therefore be explained by QCD background events.

This qualitative study of W boson M_T and P_T shows the potential of my new methods. One can weaken the preselection cuts on $\not\!\!E_T$ and obtain, by appropriately choosing the cut on the neural network, a rather pure sample. The background template which I obtained by applying an inverted neural network cut on the data sample can qualitatively explain the difference between the Monte Carlo simulation and the data events. For an in-depth study of the W boson properties, my tool is very well suited. Of course, one would have to further investigate the electron energy scale. The recoil energy of the W boson would play an important role in the $W + 0$-jets events which would be utilized to measure a quantity like M_T. Other background processes to $W \to e\nu_e$ must also be considered, like $W \to \tau\nu_\tau$ and τ decaying into an electron.

7.6.2 Single-Top Projections

In the theory chapter in section 2.2.2, a simulation with the MadEvent Monte Carlo generator of the electroweak top quark production indicated an increase in the acceptance of 26.7% when using electrons in the forward region. This study estimates the number of W boson events with two or three jets. A requirement of $\not\!\!E_T > 20$ GeV is asked for events in the `ptop00` data sample. Electrons are selected using two neural network cut scenarios and the standard cut scenario. The event yield of the sample prior to the application of the b tagging algorithm is listed in table 7.5.

The second line in the table shows the event rates for events with an electron in the central region [94] obtained with comparable preselection cuts. The third and fourth lines show the event rates for events with an electron in the forward region selected using a cut on the network above -0.5 and -0.6 respectively. The last line shows the event rates for events with a forward electron selected using the standard CDF cuts. The second and third column indicate the number of events with additional two or three jets respectively. The fourth column is the sum of the

Selection	2-jets	3-jets	Sum	Normalized sum
Central electrons	3362	568	3930	100%
Forward: NN > -0.5	808	154	962	24.5%
Forward: NN > -0.6	866	163	1029	26.2%
Forward: CDF cuts	912	174	1086	27.6%

Table 7.5: Pretag event rates for $W+2$-jets, $W+3$-jets, the sum of these two and the normalization to the number of total central electron events. Different selection scenarios are applied to the electrons in the forward region. The numbers are not corrected for background content.

two previous columns. The percentages in the last column are the numbers in the fourth column normalized to the number of central electron events in the fourth column. The expectation from Monte Carlo simulation is 26.7% for the forward electron scenarios: this number is in good agreement with the numbers obtained from data. The event yield is not corrected for background events. The number of events obtained with the cuts on the neural network output is slightly lower than the number of events obtained with the standard CDF cuts. The standard cuts select more background events and fewer signal events than the neural network cuts. The decrease in background events is stronger than the increase in signal.
In order to obtain the final sample used in the single-top analysis, one would then apply the b tagging algorithm on the jets in the event. As shown in reference [94], this requirement selects 2.9% of the events in the pretag sample. The percentage is independent of the detector part the electron is measured with. Therefore, the acceptance gain will not change in the tagged sample.

Chapter 8

Conclusions

In this thesis, I have studied the identification of high energetic electrons measured with the forward detectors of the *Collider Detector at Fermilab, CDF*. These electrons are produced in proton-antiproton collisions at a center of mass energy of 1.96 TeV. The dataset used corresponds to an integrated luminosity of 320 pb^{-1}. This dataset contains all good runs with silicon tracking information from March 2002 until September 2004.

I have utilized the standard methods based on sequential cuts on calorimeter information to identify electrons. Furthermore, my thesis introduces a new identification method based on training a neural network with a signal and a background sample derived from data. In order to check the validity of the samples, both the signal and the background sample were compared to simulated events. The neural network identification method has the advantage that the purity and efficiency can easily be controlled. One can achieve 13% less background or 1.7% more signal events when using the neural network cut instead of the standard sequential cuts.

Using the standard cuts and different cuts on the output of the neural network, I studied W+jets events. Events with zero to three jets in addition to the W boson are considered. Especially W + 2-jets and W + 3-jets events are of importance for the search for electroweak top quark production. In addition to the standard CDF 4-sector method relying on events in the sidebands of the signal region, I presented two novel methods of estimating the QCD background content in a W+jets sample. The first method uses a fit to the distribution of a kinematic variable, the missing transverse energy $\not\!\!E_T$. The second method uses an *In-Situ* fit to the distribution of the output of the neural network. These two novel methods also work for the CDF cut-based identification of electrons. I have proven that on the W+jets sample, which is completely different from the training sample, the identification power of the neural network cut is better than the one of the CDF standard sequential cuts. For example, from all W + 2-jets events, 491 signal and 471 background events pass a cut on the network output above -0.5. The standard cuts result in 460 signal and 626 background events. This neural network cut selects 7% more signal and 25% less background than the standard cuts.

I have also proven that the two methods are self-consistent, reproducing the background content of a sample with known background fraction. When performing the standard CDF sequential cuts, the background fractions obtained with the new methods are rather identical. However, these background fractions significantly disagree with the background estimation obtained with the standard 4-sector method used in CDF. A clear understanding of the observed discrepancy will need detailed further studies.

This thesis opens the way for improvements in many analyses involving W bosons. For example, the measurement of the transverse mass of the W boson improves when using a neural network identification method. Most important for the physics program of the EKP is, however, the improvement to the search for electroweak top production: using electrons in the forward region of the CDF detector, the acceptance for electroweak top quark events in the electron plus jets channel will increase by 26%, as predicted by theory. The neural network developed in the context of this thesis enables a better background reduction and is ready to be used by the CDF single-top working group.

Bibliography

[1] Donald H. Perkins, "Introduction to High Energy Physics", Cambridge University Press (2000).

[2] David Griffiths, "Introduction to Elementary Particles", Wiley (1987).

[3] Francis Halzen, Alan D. Martin, "Quarks and Leptons: An Introductory Course in Modern Particle Physics", Wiley (1984).

[4] Wolfgang Wagner, "Top quark physics in hadron collisions", *Rep. Prog. Phys.* **68**:2409-2494 (2005).

[5] F. Abe *et al.*, *Phys. Rev. Lett.* **74**:2626 (1995).

[6] S. Abachi *et al.*, *Phys. Rev. Lett.* **74**:2632 (1995).

[7] **CDF** and **DØ** collaborations, the Tevatron Electroweak Working Group, "Combination of CDF and DØ Results on the Top-Quark Mass", `hep-ex/0507091` (2005).

[8] Dominic Hirschbühl, "Measurement of the charge asymmetry and the W boson helicity in top-antitop quark events with the CDF II experiment", PhD-Thesis, Institut für Experimentelle Kernphysik, Universität Karlsruhe, `IEKP-KA/2005-25` (2005).

[9] J.H. Kühn, "Theory of Top Quark Production and Decay", `hep-ph/9707321` (1997).

[10] M. Cacciari, S. Frixione, G. Ridolfi, M. Mangano and P. Nason, *JHEP* **404**:68 (2004).

[11] N. Kidonakis and R. Vogt, *Phys. Rev.* **D68**:114014 (2003).

[12] **CDF** collaboration, "Combination of CDF top pair production cross section measurements", CDF note 7794, to be published as conference note, 2005.

[13] S. Eidelman *et al.* (Particle Data Group), "2004 Review of Particle Physics" *Phys. Lett.* **B592**:1 (2004) and 2005 partial update for the 2006 edition available on the PDG WWW pages: `http://pdg.lbl.gov/`

[14] B. W. Harris, E. Laenen, L. Phaf, Z. Sullivan and S. Weinzierl, "The fully differential single-top-quark cross section in next-to-leading order QCD", *Phys. Rev.* **D66**:054024 (2000).

[15] Z. Sullivan, "Understanding single-top-quark production and jets at hadron colliders", *Phys. Rev.* **D70**:114012 (2004).

[16] **CDF** Collaboration, D. Acosta *et al.*, *Phys. Rev.* **D65**, 091102 (2002); **D69**:052003 (2004).

[17] **DØ** Collaboration, V. Abazov *et al.*, *Phys. Lett.* **B517**, 282 (2001); *Phys. Rev.* **D63**:031101 (2000).

[18] **CDF** Collaboration, D. Acosta *et al.*, "Search for electroweak single-top-quark production in $p\bar{p}$ collisions at $\sqrt{s}$ =1.96 TeV", *Phys. Rev.* **D71**:012005 (2005).

[19] **D0** Collaboration, V. Abazov, *et al.*, "Search for single-top quark production in $p\bar{p}$ collisions at $\sqrt{s}$ =1.96 TeV", *Phys. Lett.* **B626**:55 (2005).

[20] **CDF** single-top group, sensitivity projections (web page, status 02/10/06): `http://www-cdf.fnal.gov/physics/new/top/public/singletop/projections/Projections.html`

[21] S. Budd *et al.*, "Validation of the matched MadEvent Single-Top Signal Sample with ZTOP NLO Calculations", CDF note 7701.

[22] **CDF** collaboration, T. Affolder *et al.*, "Measurement of the two jet differential cross-section in proton anti-proton collisions at $\sqrt{s}$ = 1800 GeV" *Phys. Rev.* **D64**:012001 (2001).

[23] **CDF** Collaboration, T. Affolder *et al.*, "Measurement of the Top Quark Mass with the Collider Detector at Fermilab", *Phys. Rev.* **D63**:032003 (2001).

[24] CDF 's official Luminosity Web Page (status 02/10/06): `http://www-cdf.fnal.gov/~konigsb/lum_official_page.html`

[25] Run II Luminosity Upgrade Project Plan v2.0 (January 30, 2004) web page (status 02/10/06): `http://www-bdnew.fnal.gov/doereview04/RunII_Upgrade_Plan_v2.0.pdf`

[26] Up-to-date and historical information about the Tevatron performance (status 02/10/06): `http://www-bdnew.fnal.gov/pbar/AEMPlots/`

[27] **CDF-II** Collaboration, R. Blair *et al.*, "The CDF-II detector: Technical design report", FERMILAB-PUB-96-390-E.

[28] **CDF**-II Collaboration, F. Abe et. al., "Proposal for Enhancement of the CDF II Detector: An Inner Silicon Layer and A Time of Flight Detector", FERMILAB-PROPOSAL-909.

[29] D. Bortoletto *et al.*, *Nucl. Instr. and Meth.* **A386**:87 (1997).

[30] A. Sill *et al.*, *Nucl. Instr. and Meth.* **A447**:1 (2000).

[31] T. Affolder *et al.*, *Nucl. Instr. and Meth.* **A526**:249 (2004).

[32] L. Balka *et al.*, *Nucl. Instr. and Meth.* **A267**:272 (1988).

[33] S. Bertolucci *et al.*, *Nucl. Instr. and Meth.* **A267**:301 (1988).

[34] F. Abe *et al.*, *Nucl. Instr. and Meth.* **A271**:387 (1988).

[35] G. Apollinari *et al.*, "CDF end plug calorimeter upgrade project", Proceedings of the Fourth International Conference on Calorimetry in High Energy Physics, La Biodola, Italy (1993).

[36] A. Artikov *et al.*, *Nucl. Instr. and Meth.* **A538**:532 (2005).

[37] M. Albrow, "CDF Run II Trigger Table and Dataset plan", CDF note 4718.

[38] B. Ashmanskas *et al.*, *Nucl. Inst. and Meth.* **A518**:532 (2004).

[39] G. Gomez-Ceballos *et al.*, *Nucl. Instr. and Meth.* **A518**:522 (2004).

[40] W. Wagner, H. Stadie, T. Arisawa, K. Ikado, K. Maeshima, H. Wenzel and G. Veramendi, FERMILAB-CONF-02-269-E.

[41] R. Brun and F. Rademakers, ROOT system homepage (status 02/10/06): `http://root.cern.ch/`

[42] M. S. Neubauer, "Computing for Run II at CDF", *Nucl. Instrum. Meth.* **A502**:386 (2003).

[43] 1003.1 **IEEE** Standard for Information Technology, "Portable Operating System Interface (POSIX)".

[44] **GNU** Compiler Collection, web page (status 02/10/06): `http://gcc.gnu.org/`

[45] P. Schemitz, "Eine Hard- und Software-Umgebung fr Physik-Analysen bei CDF II", PhD thesis (in german), Institut für Experimentelle Kernphysik, Universität Karlsruhe, `IEKP-KA/2003-1` (2003).

[46] D. A. Patterson, G. Gibson and R. H. Katz, "A case for redundant arrays of inexpensive disks (RAID).", Proceedings of the ACM SIGMOD International Conference on Management of Data (1988).

[47] Network File System (NFS). See RFC1094 and RFC1813 of the *Internet Engineering Task Force* (`http://www.ietf.org/rfc/`).

[48] InfiniBand Trade Association. InfiniBand Architecture Specification, Release 1.0, October 24 (2000).

[49] Sequential data Access via Meta-data (SAM) (status 02/10/06): http://d0db.fnal.gov/sam/

[50] Michael K. Johnson, "Whitepaper: Red Hat's New Journaling File System: ext3",http://www.redhat.com/support/wpapers/redhat/ext3/index.html

[51] XFS filesystem homepage (status 02/10/06): http://oss.sgi.com/projects/xfs/

[52] Y. Kemp, Fragmentation script (status 02/10/06) http://www-ekp.physik.uni-karlsruhe.de/~kemp/fragment/

[53] CVS home page (status 02/10/06): http://www.nongnu.org/cvs/

[54] XEN home page (status 02/10/06): http://www.cl.cam.ac.uk/Research/SRG/netos/xen/

[55] VMware product page (status 02/10/06): http://www.vmware.com/

[56] dCache home page (status 02/10/06): http://www.dcache.org/

[57] sudo home page (status 02/10/06): http://www.courtesan.com/sudo/

[58] Lightweight Directory Access Protocol (LDAP). See RFC2251 of the *Internet Engineering Task Force* (http://www.ietf.org/rfc/).

[59] Opteron test benchmarks, web page (status 02/10/06): http://www-ekp.physik.uni-karlsruhe.de/~x86-64/

[60] Network Address Translator (NAT). See RFC1631 of the *Internet Engineering Task Force* (http://www.ietf.org/rfc/).

[61] AC++ Framework home page (status 02/10/06): http://www-cdf.fnal.gov/upgrades/computing/projects/framework/framework.html

[62] H. Wenzel, "Tracking in the SVX", CDF note 1790.

[63] A. Mukherjee, "CTC and VTX Tracking", CDF note 5490.

[64] P. Azzi, G. Busetto, P. Gatti, and A. Ribon,"Histogram Tracking in the COT", CDF note 5562.

[65] K. Bloom and W.-M. Yao, "Outside-In"Silicon Tracking at CDF", CDF note 5991.

[66] S. Menzemer, "Spurrekonstruktion im Silizium-Vertexdetektor des CDF-II-Experiments", PhD thesis, Institut fur Experimentelle Kernphysik, Universitat Karlsruhe, IEKP-KA/2003-4 (2003).

[67] F. Bedeschi et. al., "Primary Vertex Finding Package", CDF Note 1789.

[68] H. Stadie et. al., "The Beam Position in Run II", CDF Note 6327.

[69] Run II Alignment group (status 02/10/06):
`http://www-cdf.fnal.gov/internal/upgrades/align/alignment.html`

[70] G. C. Blazey et. al., "Run II jet physics"in Physics at Run II: QCD and Weak Boson Physics Workshop (1999) `hep-ex/000501`.

[71] A. Bhatti *et al.*, "Determination of the Jet Energy Scale at the Collider Detector at Fermilab", FERMILAB-PUB-05-470 (2005) `hep-ex/0510047`.

[72] **CDF** Collaboration, D. Acosta *et al.*, "Measurement of the $t\bar{t}$ production cross section in $p\bar{p}$ collisions at $\sqrt{s}$ =1.96 TeV using lepton + jets events with secondary vertex b-tagging", *Phys. Rev.* **D71**:072005 (2005).

[73] R. Brun and F. Carminati, "Cern programming library, long writeup" **W5013**.

[74] G. Grindhammer, M. Rudowicz, and S. Peters, "The fast simulation of electromagnetic and hadronic showers", *Nucl. Instrum. Meth.* **A290**:469 (1990).

[75] E. Gerchtein and M. Paulini, "Cdf detector simulation framework and performance" *ECONF* C0303241 (2003) TUMT005, `physics/0306031`.

[76] TopNtuple and TopMods documentation, CDF notes 5947, 6737 and 7267.

[77] M. Feindt, "A Neural Bayesian Estimator for Conditional Probability Densities", (2004) `physics/0402093`.

[78] M. Feindt, S. Richter, W. Wagner, "A Neural Network b Tagger for Single-Top Analyses", CDF note 7816.

[79] C. Issever, B. Wagner, A. Robson, C. Mills, "Plug Electron Baseline Cuts as defined in Summer 2003 and their Efficiencies", CDF Note 6789.

[80] D. Glenzinski, C. Hill, J. Incandela, N. Lockyer, P. Merkel, C. Mills, P. Savard, J. Thom, P. Wittich, A. Yagil, "Dilepton Analysis Tools in 5.3.X TopFind", CDF note 7034.

[81] M. Albrow for the CDF collaboration, "The CDF Plug Upgrade electromagnetic calorimeter: test beam results", *Nucl. Instrum. Meth.* **A480**:524 (2002).

[82] T. Nelson, R. Snider, D. Stuart, "Forward Electron Tracking with the PhoenixMods Package", CDF note 6278.

[83] J. Lewis, D. Saltberg and M. Shochet for the Trigger and Datasets Working Group, "CDF Run-II Trigger Table and Datasets Plan", CDF note 4718.

[84] B. Knuetson, M. Shochet for TDSWG, "Datasets and Raw Datastreams for Run II", CDF note 5565.

[85] A. Robson, G. Manca, P. Renton, G. Veramendi, Y.-K. Kim, "A Measurement of $\sigma \cdot \mathrm{Br}(Z^0 \to e^+e^-)$ using Run 2 Central and Plug Electrons in 72 pb^{-1}", CDF note 6642.

[86] private communication with Thorsten Scheidle about stripping `bpel0d` down to `ptop00`.

[87] A. Attal, J. Hauser, M. Lindgren, B. Mohr, "Investigation and Reduction of Cross Talk in the Run 2 Endplug Calorimeter PES and PPR Detectors", CDF note 6812.

[88] T. Sjöstrand, P. Edén, C. Friberg, L. Lönnblad, G. Miu, S. Mrenna and E. Norrbin, *Computer Phys. Commun.* **135**:238 (2001) (LU TP 00-30, `hep-ph/0010017`) .

[89] M. L. Mangano, M. Moretti, F. Piccinini, R. Pittau, A. Polosa, "ALPGEN, a generator for hard multiparton processes in hadronic collisions", *JHEP* **0307**:001 (2003) `hep-ph/0206293`.

[90] G. Corcella *et. al.*, "HERWIG 6.5", *JHEP* **0101**:010 (2001), `hep-ph/0011363`, `hep-ph/0210213`.

[91] A. Foland *et. al.*, "Preliminary Method 2 Backgrounds for Top Pair Production in Lepton Plus Jets with SecVtx Tags", CDF note 7486.

[92] Hartmut Stadie, "Results on the production and detection of W bosons with the Collider Detector at Fermilab in proton antiproton collisions at a center-of-mass energy of 1.96 TeV", FERMILAB-THESIS-2003-33.

[93] R. J. Barlow and C. Beeston, "Fitting using finite Monte Carlo samples", *Comput. Phys. Commun.* **77** (1993) 219-228.

[94] Wolfgang Wagner, "Data based background estimate for the single-top analysis", Talk given at the CDF Top Properties Meeting, 10/14/2005.

Acknowledgments

I wish to sincerely thank my supervisor Prof. Dr. Thomas Müller and my co-supervisor Prof. Dr. Günter Quast. I had the chance to work on many interesting subjects under their guidance and with their support.

I must thank Dr. Wolfgang Wagner for giving me the opportunity to work closely with him in the CDF single-top working group. He always supported my work and gave me a multitude of valuable advices. I owe a debt of gratitude to Dr. Hartmut Stadie, who helped me setting up this analysis and a lot of technicalities before his departure to DESY.

I wish to thank all the people that build the CDF detector and keep it running. Without a detector taking good data, any analysis is unimaginable.

I am also indebted to the people that carefully read and commented on this manuscript, especially Dr. Wolfgang Wagner, Michael Schroff, Dr. Anja Vest and Jan Lück.

I would like to thank all those that helped keeping the computing cluster running.

I thank the members of the Institut für Experimentelle Kernphysik and especially my room mates for the nice working atmosphere. I will not forget the many interesting discussions about physics and other important things in live.

I had a great time with the cabaret ensemble *Die Kratzbürsten*. I enjoyed playing the piano with them. I have to thank Frank Zappa for his great music which gave me inspiration and incitation.

Support of this work was provided in form of a graduate scholarship by the government of the Grand-Duchy of Luxembourg.

Last but not least I thank my parents for their love and support.