universität wien

DIPLOMARBEIT

Titel der Diplomarbeit

Überprüfung und Kalibrierung der Vertexrekonstruktion mit ersten CMS Daten

angestrebter akademischer Grad

Magister/Magistra der Naturwissenschaften (Mag. rer.nat.)

Verfasserin / Verfasser:	Christian Thomay
Matrikel-Nummer:	0103355
Studienrichtung (lt. Studienblatt):	A 411 Physik
Betreuerin / Betreuer:	Univ.Doz. DI Dr. Rudolf Frühwirth

Wien, am 31.1.2011

Contents

Chapter 1

Introduction

1.1 Overview

The purpose of this thesis is to give a report on first-hand experience with analysis data from the CMS detector at the LHC machine of CERN. It will give an overview of vertex fitting in the CMS detector and propose a new data-driven measure of the efficiency of the vertexing algorithms currently in use. The LHC machine and the various detector subsystems of the CMS will be explained (chapters 2, 3, and 4), and an overview of the analysis algorithms (chapters 5 and 6) and the software framework (chapter 7) of CMS will be given. Finally, chapter 8 will describe in detail a data-driven method to obtain a measure of vertex reconstruction efficiency.

1.2 Vertex Fitting in CMS

The identification of points of common origin (*vertices*) in a set of reconstructed particle tracks is an important tool to practically all analysis scenarios in CMS. The correct identification of the beam collision point (referred to as the *primary vertex*) is an essential starting point for all further analysis of a collision event, and the identification of points of decay (*secondary vertices*) of produced particles is a crucial ingredient in many analysis scenarios, both in the Standard Model and and in physics beyond. Of particular interest are the decays which originate from hadrons containing bottom quarks; these feature in many of the processes CMS was designed to explore. Figure 1.1 shows some of the Feynman graphs which involve b decays. In all of those processes, bottom hadrons are produced, and due to their relative longevity and high mass, can be precisely identified with both vertex- and lepton-based methods.

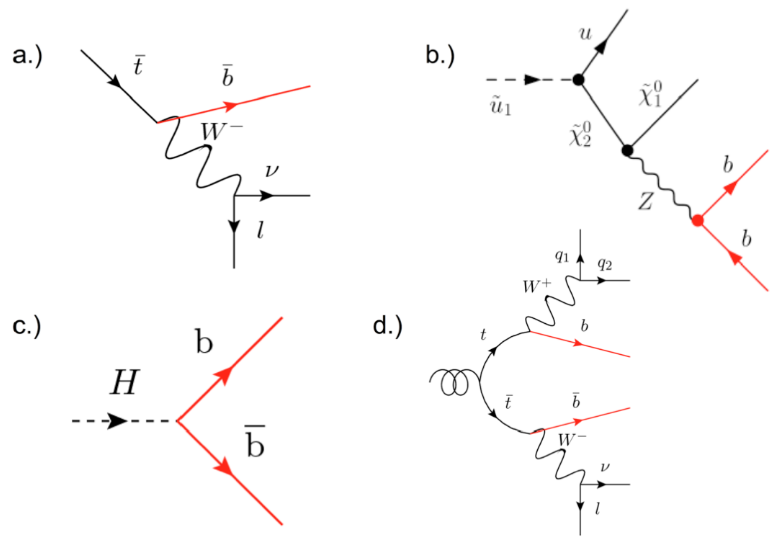


Figure 1.1: Feynman graphs of some interesting b decays. *a.*) Top decay, *b.*) SUSY neutralino to Z decay, *c.*) Higgs decay, *d.*) Top decay (SUSY background).

Chapter 2

LHC

2.1 The Machine

The Large Hadron Collider (LHC) is a two-ring superconducting hadron accelerator and collider installed in the existing tunnel that was constructed between 1984 and 1989 for the CERN Large Electron Positron Collider (LEP) machine. The 27 km tunnel lies between 45 m and 170 m below both Swiss and French soil in the area near Geneva. It contains two beam pipes and accelerates protons or heavy ions (such as lead) in opposite directions with four collision points, around which the main experiments are located (see Figure 2.1). 1232 superconducting dipole magnets along the ring keep the particle beams on their circular path, and 392 quadrupole magnets ensure the focusing of the beams. The proton-proton collisions occur currently at a center of mass energy of 7 TeV; this is planned to be increased to 14 TeV in the next years.
The number of events per second generated in the LHC collisions is given by

$$N_{\text{event}} = L\sigma_{\text{event}} \tag{2.1}$$

where σ_{event} is the cross section for the event under study and L the machine luminosity, given for a Gaussian beam distribution by

$$L = \frac{N_b^2 n_b f_{\text{rev}} \gamma_r}{4\pi\epsilon_n \beta^*} F \tag{2.2}$$

where N_b is the number of particles per bunch, n_b the number of bunches per beam, f_{rev} the revolution frequency, γ_r the relativistic gamma factor, ϵ_n the normalized transverse beam emittance, β^* the betatron function at the collision point, and F the geometric luminosity reduction factor due to the crossing angle.
Of the four main LHC experiments, ATLAS (*A Toroidal LHC ApparatuS*) and CMS (*Compact Muon Solenoid*) are intended for high luminosity, both aiming for a peak luminosity of $L = 10^{34}$ cm^{-2}s^{-1} for proton operation. The other main proton experiment, LHCb (*LHC*

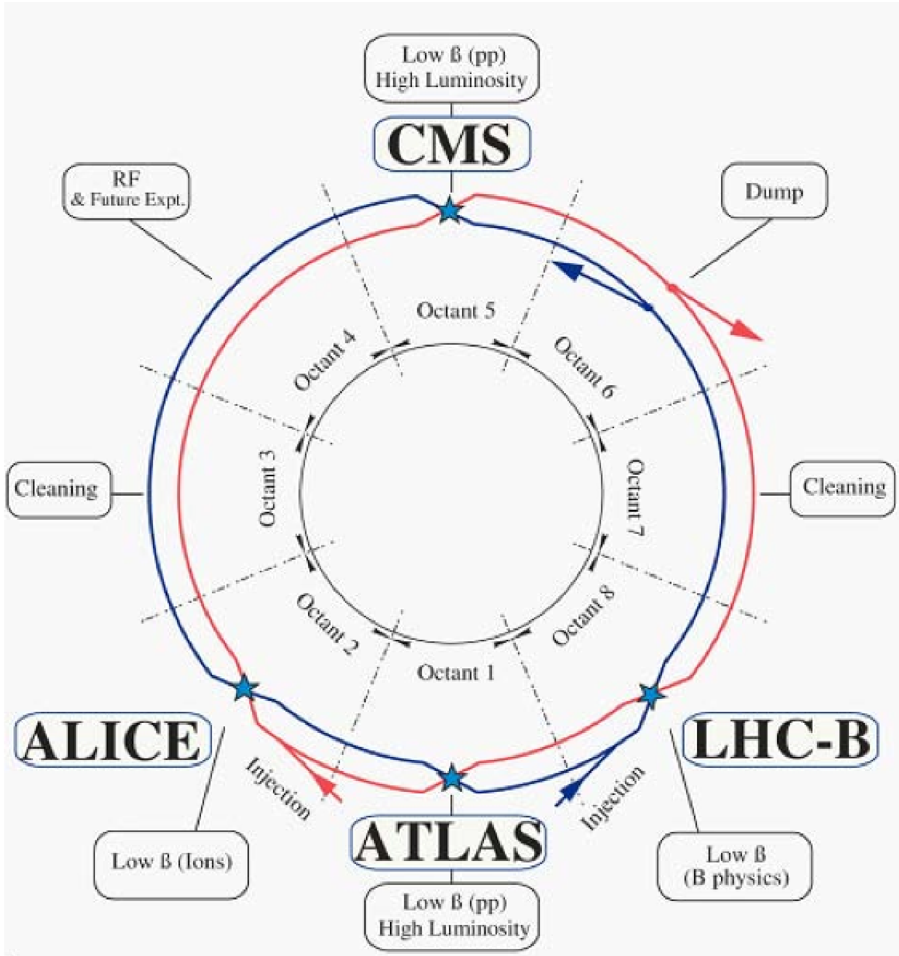


Figure 2.1: Schematic layout of the LHC ring and main experiments.[1]

beauty experiment, focusing on b physics) is a low-luminosity experiment, aiming for a peak of $L = 10^{32}$ cm^{-2}s^{-1}. Apart from the proton collisions, LHC will also be operated with ion beams, with the experiment ALICE (*A Large Ion Collider Experiment*, peak $L = 10^{27}$ cm^{-2}s^{-1}) in place to observe the lead-lead collisions. The LHC is supplied with protons from the injector chain Linac2 – Proton Synchrotron Booster (PSB) – Proton Synchrotron (PS) – Super Proton Synchrotron (SPS). This way, the previous generations of accelerators at CERN are being re-used as pre-accelerators in the injection chain. Figure 2.2 shows the injection chain and the energies to which the particles are accelerated in each step; both proton and lead ion energies are noted.

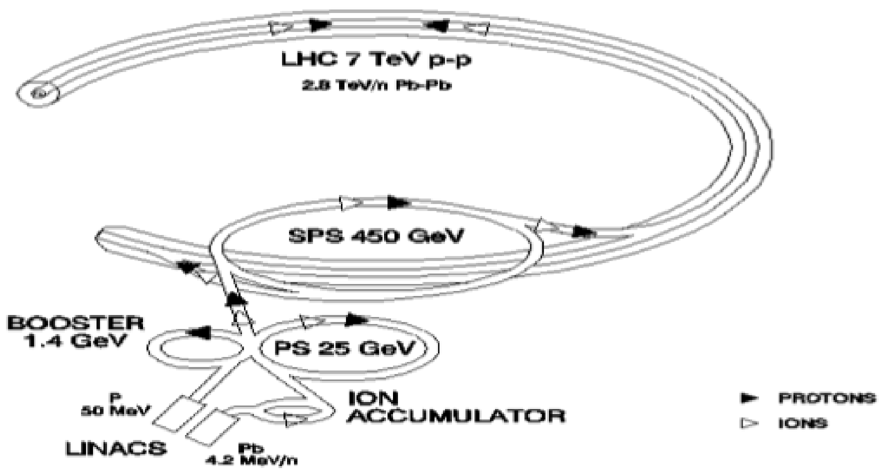


Figure 2.2: Schematic view of the LHC injector complex.[1]

2.2 Physics at the LHC

2.2.1 Search for the Standard Model Higgs Boson

The Higgs mechanism is a cornerstone of the Standard Model (SM) and its supersymmetric extensions. Due to spontaneous symmetry breaking, the electroweak gauge bosons W and Z as well as the fermions acquire masses through the interaction with the Higgs field. In the Standard Model one weak isospin Higgs doublet is introduced and leads to the existence of one elementary Higgs particle. The Higgs couplings to the electroweak gauge bosons and all fermions grow with their masses. The only unknown parameter of the Higgs boson itself is the value of its mass m_H. The search for the Higgs boson is a crucial endeavour for establishing the standard formulation of the electroweak theory. [2]
Depending on the Higgs mass m_H, different decay channels can be exploited for discovery; Figure 2.3 shows the ratio of different Higgs decays depending on the mass. Figure 2.4 shows the Higgs production cross sections in picobarn (1 pb = 10^{-36} cm^{-2}) as a function of the Higgs mass.
As of end of 2010, the Standard Model Higgs mass has been excluded by the LEP up to a mass of 114.4 GeV [3]. Experiments at the Fermilab Tevatron have excluded the region between 162-166 GeV [4]. (All exclusions to 95% confidence level.) Above 185 GeV, indirect measures suggest a discovery would be less likely; still, searches in that mass region will be performed.
Consequently, one can extract from Figure 2.3 that the currently favored mass regions decay to a large degree into pairs of $b\bar{b}$. Other important channels include $H \rightarrow \gamma\gamma$ and the "gold-plated" channel $H \rightarrow ZZ \rightarrow 4\mu$.

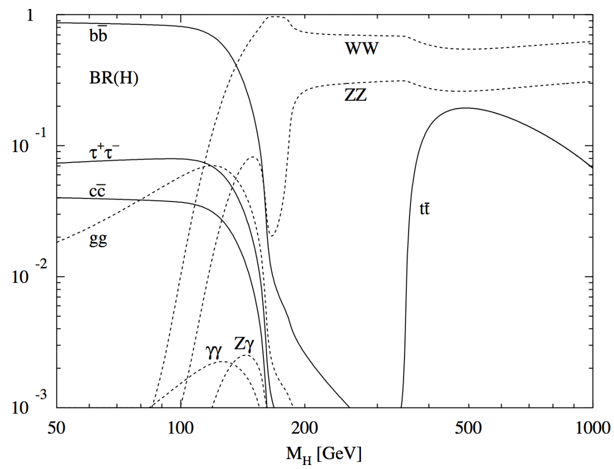


Figure 2.3: Decay branching ratios.[2]

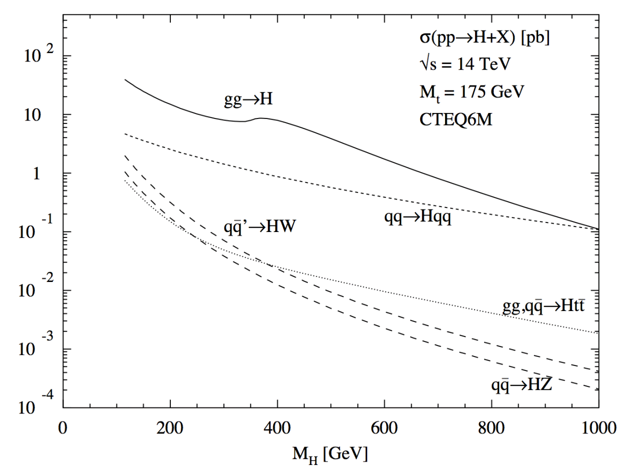


Figure 2.4: Higgs production cross sections.[2]

2.2.2 SUSY Searches

Supersymmetry (SUSY) is one of the best-motivated candidates for physics beyond the Standard Model. Low-energy SUSY is well-motivated since it stabilizes the electroweak scale; it provides a quantitatively accurate unification of gauge couplings, and is also a promising candidate for the cold dark matter theory. Moreover, it is consistent with electroweak precision data. Since the mechanism of SUSY breaking is unknown, supersymmetric extensions of the Standard Model contain a large number of unknown parameters, 105 alone in the Minimal Supersymmetric Standard Model (MSSM). Specific assumptions on the SUSY-breaking mechanism, in particular about the unification of parameters at the grand-unification (GUT) scale, considerably reduce the number of free parameters. For instance, in the minimal super gravity model (mSUGRA) one can reduce the parameters to only four (and one sign). Experiments at the LHC will not only have to discover SUSY, but also to determine precisely the underlying SUSY-breaking scenario with as few theoretical prejudices as possible.
In general, if SUSY particles exist, they will be produced at the LHC; if one assumes conservation of R parity (which essentially forbids the decay of SUSY particles into nothing but non-SUSY particles), the lightest SUSY particle is stable, but hardly interacts with matter. Therefore, the hermeticity of the detectors and their ability to identify missing transverse energy is crucial to the detection of supersymmetric particles.
More on the topic of SUSY searches at LHC can be found at [6]. For a detailed introduction to supersymmetry, see [7].

2.2.3 CP Violation

CP violation is the violation of the conservation of the combined symmetries of charge conjugation C and parity P. The core question behind this is: Why does the universe consist of matter and not antimatter? We can observe that less than 0.01% of the matter content of the universe is antimatter; but if the distribution of matter and antimatter was symmetric at some point of the universe's genesis, there must be a mechanism by which matter became predominant.
The LHC-b experiment is designed to perform high-precision CP violation measurements in the B-meson system, such as measuring the CP violating phase in the $B_s \longrightarrow J/\psi\ \phi$ decay. The tagging of neutral B mesons, and to find their flavour at production, is essential for many CP asymmetry measurements.
An overview of flavour physics and CP violation can be found in [8].

2.2.4 Heavy Ion Physics

The collision of heavy ions like lead will open up the possibility to explore a new state of matter. The collisions will compress and heat the ion nuclei, so that their individual protons and neutrons overlap. The high energy and particle density creates a local volume where, for a short time, a relatively large number of unbounded quarks and gluons can exist. This form of matter, where the partons are no longer confined, is called quark-gluon plasma. It is thought that this plasma has existed ten millionths of a second after the Big Bang, at a time before confinement and before the baryogenesis started.
In addition to simulating the very early universe, more insight in the physics of high-density stars can be gained. For more information, see [9].

2.2.5 Search for Extra Dimensions

One of the more spectacular possibilities of the LHC is that it could produce noticeable amounts of microscopic black holes. Normally, black hole formation in particle colliders would not be a relevant process unless the collision energy reaches the Planck scale; since $M_{\text{Planck}} \approx 10^{19}$ GeV, this would not be possible with a man-made collider in any foreseeable future. The Planck scale might, however, be corrected downwards to an observable domain by additional spatial dimensions, which are predicted by theories such as Kaluza-Klein [10] or string theory [11]. In that case, the transplanckian domain could be experimentally probed and could lead to the formation of black holes.
Figure 2.5 shows the amount of energy available plotted against the smallest distance one can probe with the given energy. Figure 2.6 shows the same, with extra dimensions taken into account. If the collision energy is much larger than the Planck energy, black holes form and only what lies outside the Schwarzschild radius (which depends directly on the Planck scale) can be probed. The formation of black holes, as summarized by the so-called Hoop Conjecture [12], would occur simply by the compactifying of energy on a small scale.

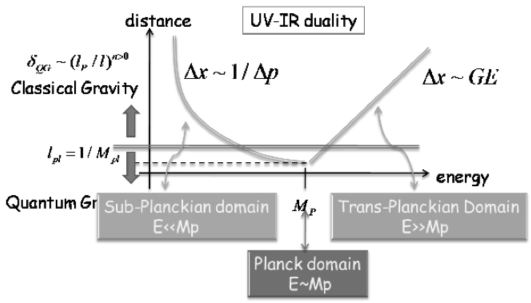


Figure 2.5: Energy vs. distance.[13]

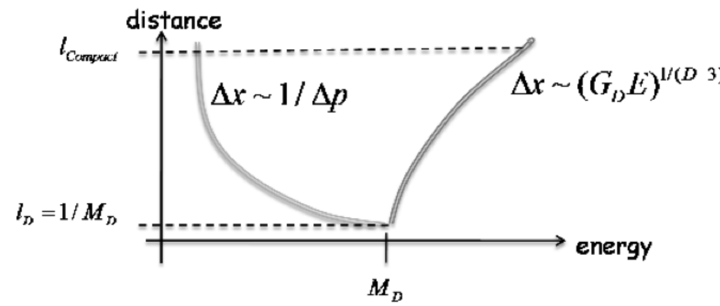


Figure 2.6: The same figure with extra dimensions.[13]

Black hole formation at the LHC would lead to decay chains involving the fundamental particles such as leptons, gauge bosons, jets etc. Thereby one could measure the mass and Hawking temperature of the black holes and extrapolate to the number of extra dimensions. More details on the formation of black holes in collider experiments can be found in [13].

Chapter 3

The CMS Experiment

3.1 Overview

The Compact Muon Solenoid (CMS) is designed as a multipurpose experiment with one of the prime motivations being the study of the Higgs mechanism. The whole expected mass range of the Higgs from 114 GeV to 1 TeV is covered by the experiment. Many experimental signatures are possible, involving high transverse energy muons, electrons, photons and jets[1]. In order to cleanly detect these signatures, their identification and precise measurement over a large energy scale and at high luminosities is essential.
The detector requirements for CMS to meet the goals of the LHC physics program can be summarized as follows:

- Good muon identification and momentum resolution over a wide range of momenta and angles, good dimuon mass resolution ($\approx 1\%$ at 100 GeV), and the ability to determine unambiguously the charge of muons with $p < 1$ TeV;
- Good charged-particle momentum resolution and reconstruction efficiency in the inner tracker. Efficient triggering and offline tagging of τ's and b-jets, requiring pixel detectors close to the interaction region;
- Good electromagnetic energy resolution, good diphoton and dielectron mass resolution ($\approx 1\%$ at 100 GeV), wide geometric coverage, π^0 rejection, and efficient photon and lepton isolation at high luminosities;
- Good missing-transverse-energy and dijet-mass resolution, requiring hadron calorimeters with a large hermetic geometric coverage and with fine lateral segmentation.

The design of CMS, as described in this chapter, meets these requirements. The main distinguishing features of CMS are a high-field solenoid, a full-silicon-based inner tracking

[1] A jet is a narrow shower of decay particles originating from the hadronization of a coloured particle.

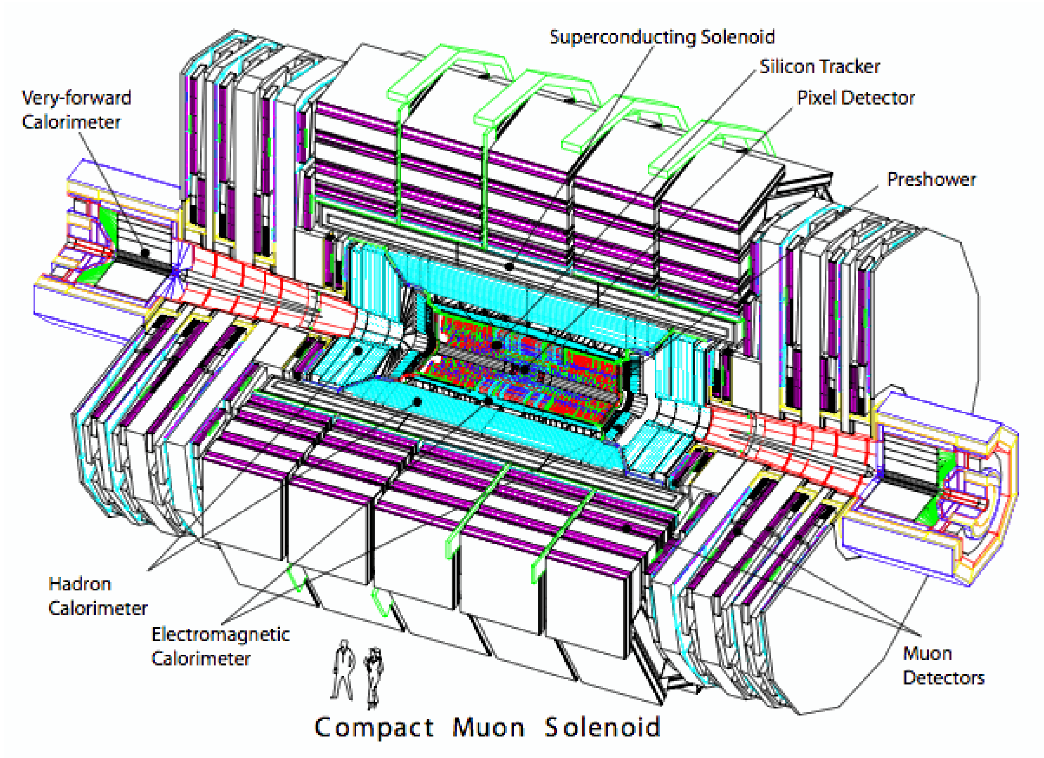


Figure 3.1: Perspective view of the CMS detector.[5]

system, and a homogeneous scintillating-crystals-based electromagnetic calorimeter. Full details on the systems of CMS can be found in [5].

3.2 Detector Layout

The overall dimensions of the CMS detector, shown in Figure 3.1, are a length of 21.6 m, a diameter of 14.6 m and a total weight of 12500 tons.

The detector is constructed in onion-like layers around the interaction point of the proton beams. In the center, there is the tracking system, consisting of 10 layers of silicon microstrip detectors and 3 layers of silicon pixel detectors in the central part. Around it are the calorimeter systems, consisting of the electromagnetic calorimeter (ECAL) and the hadronic calorimeter (HCAL); all of these are embedded in a 4-Tesla superconducting solenoid. In the return yoke necessitated by the enormous magnetic field, 4 muon stations are integrated, each consisting of several layers of aluminium drift tubes (DT), cathode strip chambers (CSC) and resistive plate chambers (RPC).

	electrons	photons	pions	neutrons	muons
Tracker	×		×		×
ECAL	×	×	×		×
HCAL			×	×	×
Muon System					×

Table 3.1: Types of particles which can be detected with the different systems of CMS.

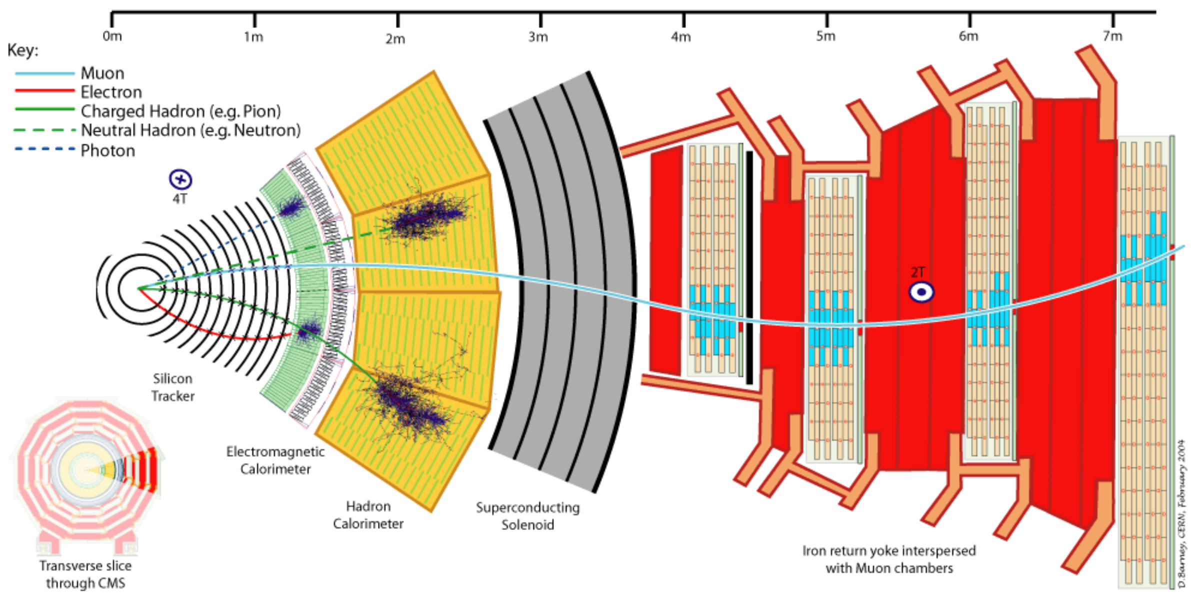


Figure 3.2: Particles passing through the CMS detector systems.[14]

Table 3.1 shows which detectors are sensitive to which type of particles, and Figure 3.2 illustrates the particles' passing through the various systems of the detector.

3.3 The Superconducting Magnet

One of the most important aspects in measuring charged particles' momenta is the magnetic field in which the particles' trajectories are bent. Therefore, a high magnetic field (4 Tesla) is necessary to obtain good momentum resolution both in the tracker and in the muon systems.
To achieve this, a superconducting coil is placed around the calorimeter and tracker systems. The coil is 13 m in length with an inner diameter of 5.9 m and consists of refrigerated

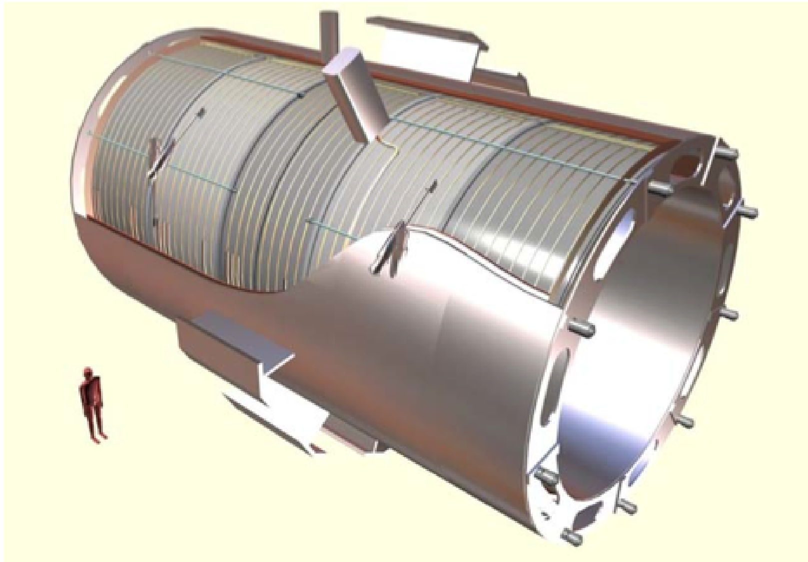


Figure 3.3: Artistic view of the CMS solenoid.[5]

superconducting niobium-titanium filaments embedded in a solid copper matrix (illustrated in Fig. 3.3). Around the coil, in order to close the field lines, is a 10 000-ton iron yoke. The magnet system's total weight is about 12 000 tons, making it the largest superconducting magnet ever built.

3.4 The Tracker System

The inner tracking system of CMS is designed to provide a precise and efficient measurement of the trajectories of charged particles originating from the LHC collisions, as well as a precise reconstruction of secondary vertices.

The solenoid provides a homogeneous magnetic field of 4 Tesla over the whole volume of the tracker; this way, the momentum and charge of particles can be measured by analyzing the trajectories (curved by the Lorentz force from the magnetic field) of the charged particles moving through the tracker.

Figure 3.4 shows a schematic overview of the tracker system. The angular component is given in pseudorapidity η, $\eta = -\ln(\tan(\frac{\theta}{2}))$. The Tracker Inner Barrel and Disks (TIB/TID) are composed of 4 barrel layers, supplemented by 3 disks at the end. The TIB/TID is surrounded by the Tracker Outer Barrel (TOB), consisting of 6 barrel layers of micro-strip sensors. Beyond it are the Tracker Endcaps (TEC+ or TEC- depending on their position on the z axis), composed of 9 disks with up to 7 rings of micro-strip detectors.

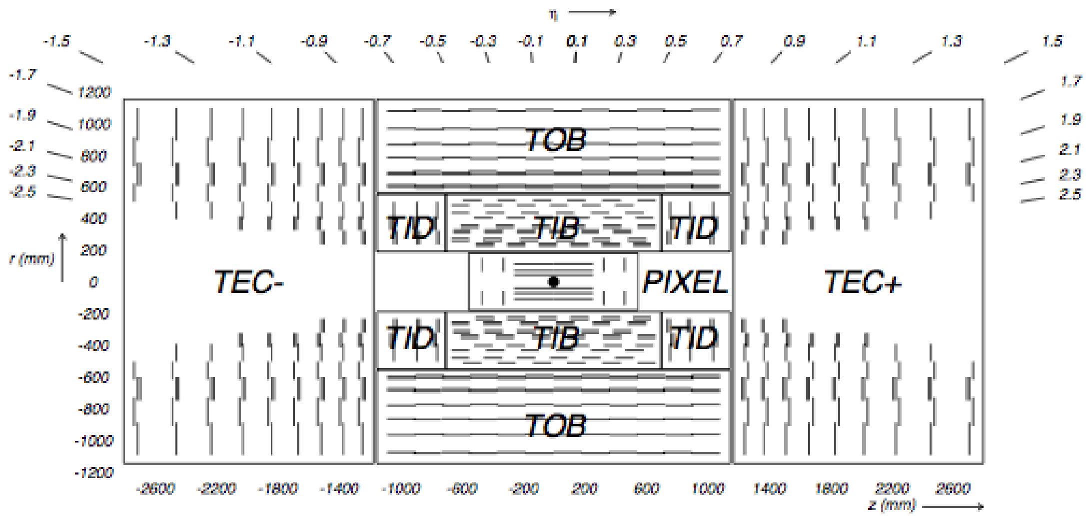


Figure 3.4: Schematic view of the CMS inner tracker.[5]

3.4.1 Pixel Detectors

The pixel detector, located in the innermost region around the interaction point, is built out of small silicon pixel cells on pixel detector modules, each module consisting of a 250 µm thin, segmented sensor plate (Figure 3.5). Each pixel cell, 100 x 150 μm^2 in size, consists of high-dose n-implants introduced in a high-resistance n-substrate, with the pn-junction placed on the back side of the sensor. The pixel cells are bump-bonded to highly integrated readout chips, as shown in Figure 3.6. The spatial resolution is in the range of 15-20 µm.

3.4.2 Strip Tracker

The silicon strip tracker, located around the pixel detector, consists of single-sided p-on-n type silicon micro-strip sensors. When passing through the depletion zone (Figure 3.7), charged particles cause opposite-sign pairs to drift to the electrodes and induce a signal; with the different signals in the strips the position of the passing particles can be reconstructed.

3.5 Calorimetry Systems

3.5.1 The Electromagnetic Calorimeter

The electromagnetic calorimeter (ECAL) is designed to precisely measure the energies of electrons and photons; as such, one of the driving criteria in the design was the capability to detect the decay of the Higgs boson into two photons.

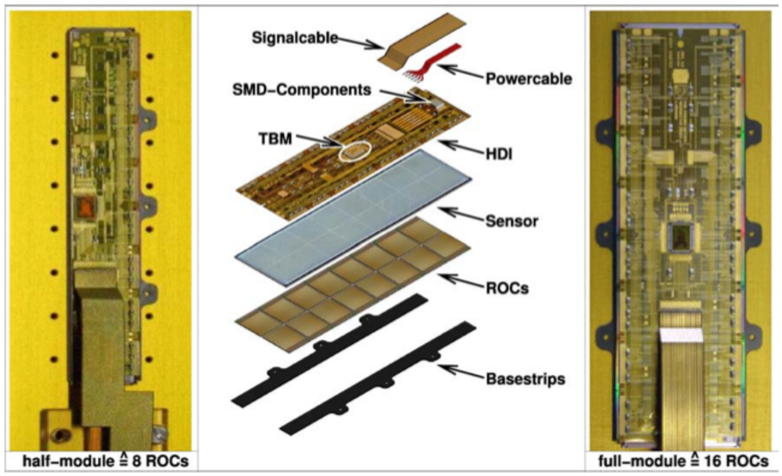


Figure 3.5: Exploded view (middle) of a barrel pixel detector full module (right).[5]

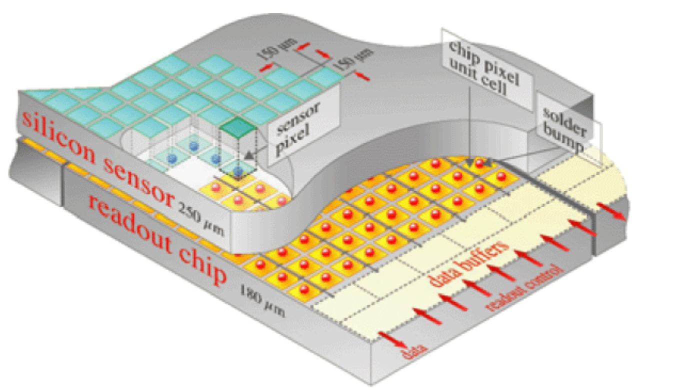


Figure 3.6: Schematic view of a pixel detector element.[15]

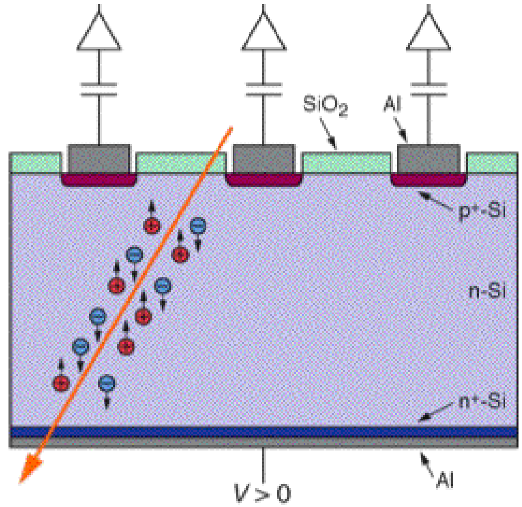


Figure 3.7: Schematic view of a particle passing through a silicon strip detector element.[16]

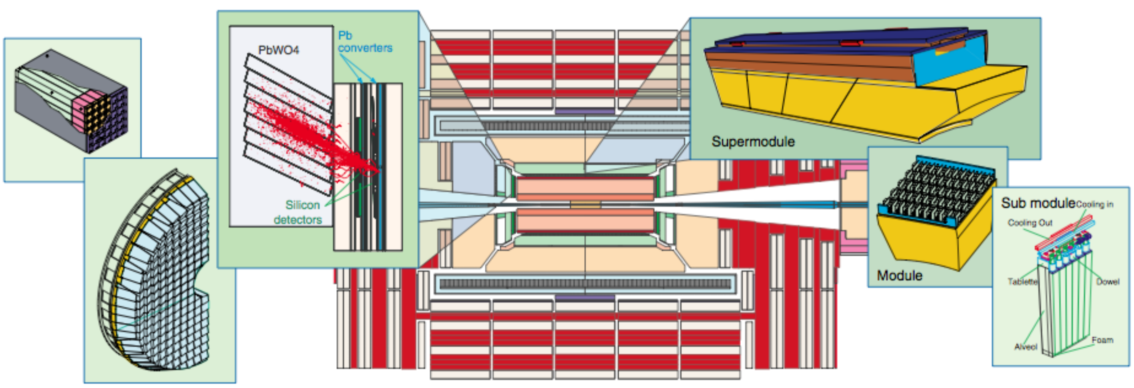


Figure 3.8: Overview of the CMS electromagnetic calorimeter.[20]

The ECAL (Fig. 3.8) is a hermetic homogeneous calorimeter made of 61 200 lead tungstate ($PbWO_4$) crystals mounted in the central barrel part and 7 324 crystals in each of the two endcaps. The characteristics of the lead tungstate crystals make them an appropriate choice for operation at the LHC: the high density (8.28 g/cm^3), short radiation length (0.89 cm) and small Molière radius (2.2 cm) result in a fine granularity, compactness, and radiation resistance. The scintillation light emitted by the crystals is amplified by avalanche photodiodes and vacuum phototriodes in the barrel and the endcaps respectively.

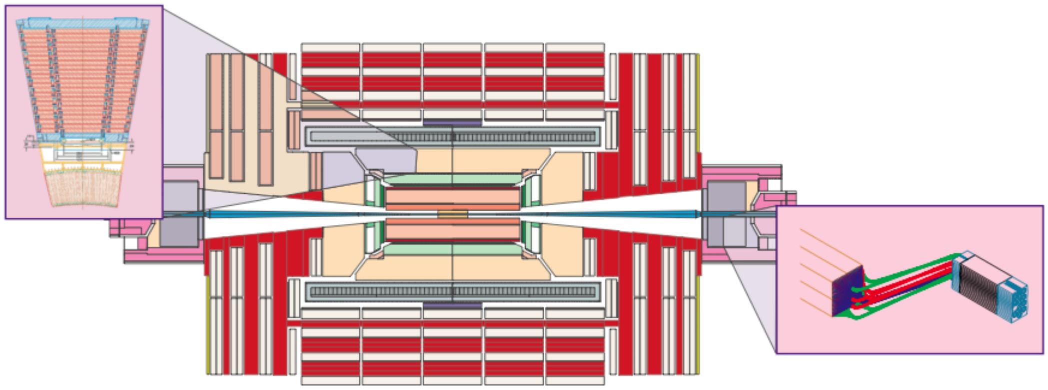


Figure 3.9: Overview of the CMS hadron calorimeter. (Zoomed in left: central hadron calorimeter. Zoomed in right: very-forward calorimeter.)[21]

3.5.2 The Hadron Calorimeter

The hadron calorimeter (HCAL) is designed to measure the energy and position of strongly interacting hadrons such as protons, neutrons, pions and kaons. Hadronic showers start to develop later and have larger longitudinal and lateral dimensions than electromagnetic ones; therefore, the HCAL has to be thicker than the ECAL in order to absorb the same amount of energy.

The HCAL (Figure 3.9) sits behind the tracker and the ECAL. The hadron barrel (HB) is placed between the ECAL and the inner part of the magnetic coil; this constrains the total amount of material which can be used to absorb hadronic showers. Therefore, an outer hadron calorimeter (HO) is placed outside the solenoid to complement the barrel calorimeter. At low angles to the z axis, the forward hadron calorimeter (HF) is placed at 11.2 m from the interaction point (Figure 3.10).
The HCAL is made of plastic scintillator tiles inserted between copper absorber plates. The HF uses quartz fibers (for the high radiation hardness vital in the forward region) embedded in a copper absorber matrix.

3.6 The Muon System

In general, the only known particles left from the collision outside of the calorimeters and the magnet's return yoke are the virtually non-interacting neutrinos, and muons. Since a large amount of physical signals the LHC is designed to explore involves muons, precise detection and reconstruction of muons is essential. Also, muon detection is an integral element of triggering.
To this end, four stations of muon chambers are placed in the saturated iron yoke. The

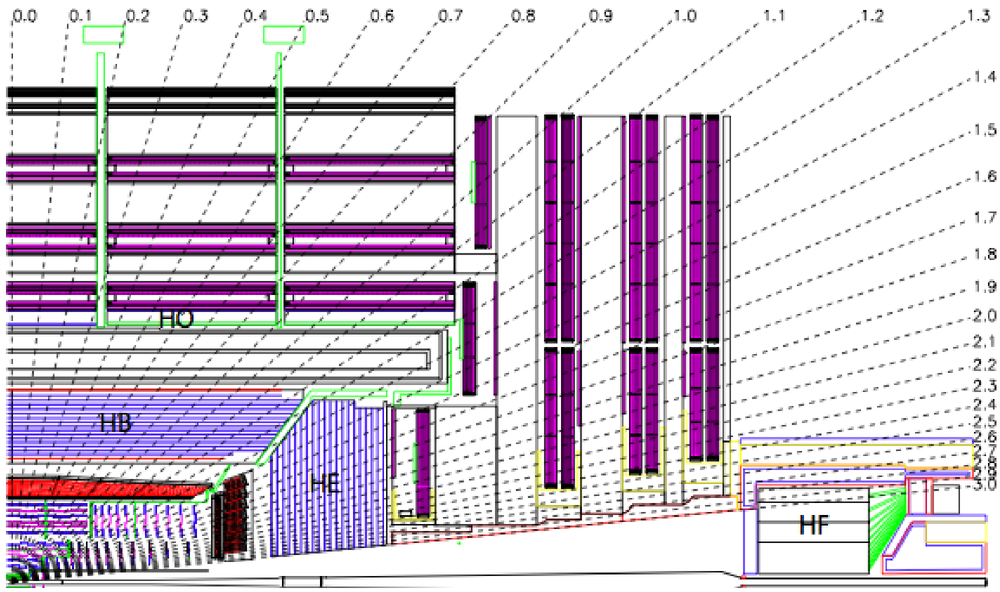


Figure 3.10: Schematic view of the hadron calorimeter subsystems.[5]

muon systems consist of drift tubes (DT) in the barrel region, resistive plate chambers (RPC) overlapping in barrel and forward regions, and cathode strip chambers (CSC) solely in the forward region.

3.6.1 Drift Tubes

The drift tube system is used in the barrel region; it consists of 250 drift tubes, 42 mm in width, arranged in four layers. Since the magnetic field in this region is low and practically fully contained by the magnet yoke, a gas-based drift approach to muon detection can be realized, using a gas mixture of 85 % Ar + 15 % CO_2.
Muons passing through the drift tubes set free electrons from the gas mixture, causing them to move along the field lines to the wire where they induce a signal (Fig. 3.11). From the precisely measured drift time the coordinates of the particle at the time of its passing can be determined.

3.6.2 Resistive Plate Chambers

Resistive parallel plate chambers (RPC) are used in both the barrel and the forward regions of the muon system. They are gaseous parallel-plate detectors that combine adequate spatial resolution with a time resolution of ≈ 1 ns (Fig. 3.12). Since the average time between

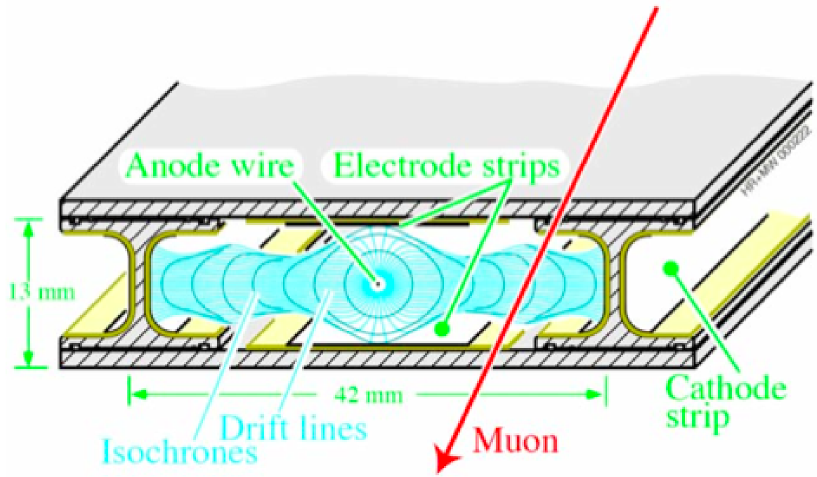


Figure 3.11: Schematic view of a drift tube, with field lines.[5]

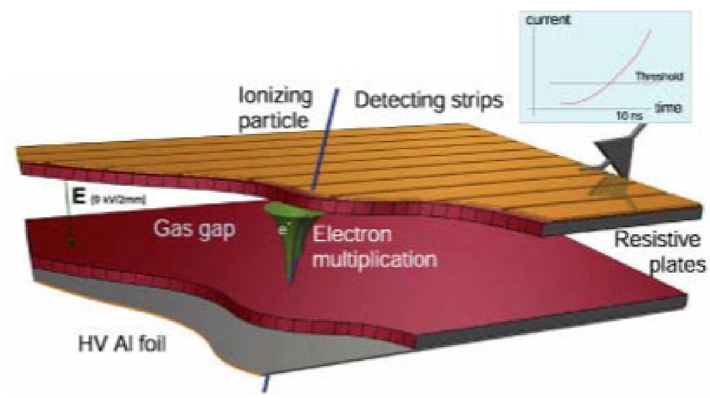


Figure 3.12: Schematic view of an RPC. The graph sketches signal time evolution.[22]

two LHC bunch crossings is 25 ns, the RPCs can associate muon signals unambiguously to the appropriate bunch, making them a logical choice for muon-based triggering.

3.6.3 Cathode Strip Chambers

Cathode strip chambers (CSC) are used exclusively in the forward region of the muon system. CSCs are multiwire proportional chambers with one cathode plane segmented into strips running across wires (Fig. 3.13). As opposed to detector systems like the gas-based drift tubes, they can operate at high rates in large and non-uniform magnetic fields and

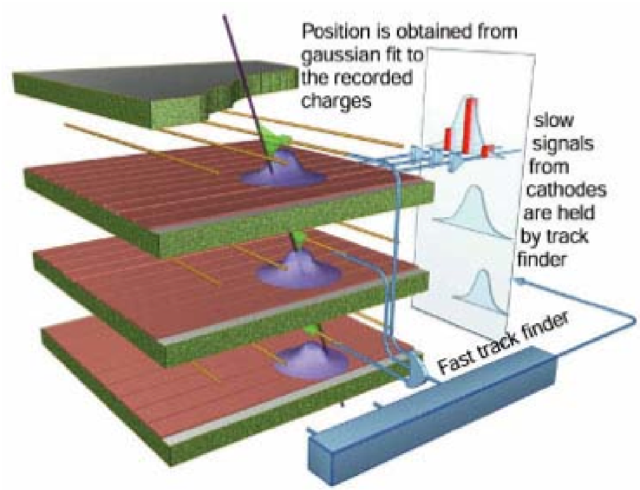


Figure 3.13: Schematic view of a CSC. The graph sketches signal detection.[22]

are not dependent on precise gas, temperature, or pressure control, which makes them a sensible choice in the high-field and high-background region of the muon endcaps.

Since CSCs utilize both strips and wires, they can resolve radial and angular coordinates at the same time, and due to the layering of the CSC modules, the third spatial coordinate can be obtained as well. Their high-precision time and space measurement makes them also well-suited for triggering.

3.7 Data Acquisition

3.7.1 Overview

The key component in all data analysis at the LHC is the *event*, essentially a digital snapshot from the detector containing all measured and reconstructed information of one bunch crossing. Each event starts out as raw data, a collection of readout signals from the detector subsystems. This raw data comes at a frequency of about 40 MHz, corresponding to the LHC bunch crossing rate of 25 ns. The CMS detector has about 15 million readout channels, each between 1 and 14 bits, which in total amounts to a rate of about 600 TB/sec. Reading out and retaining such a vast amount of data is beyond current technical possibility; therefore, a selection needs to be made regarding which events to keep, and which to discard.

This selection is called *triggering* and is arranged into multiple levels, ranging from simple

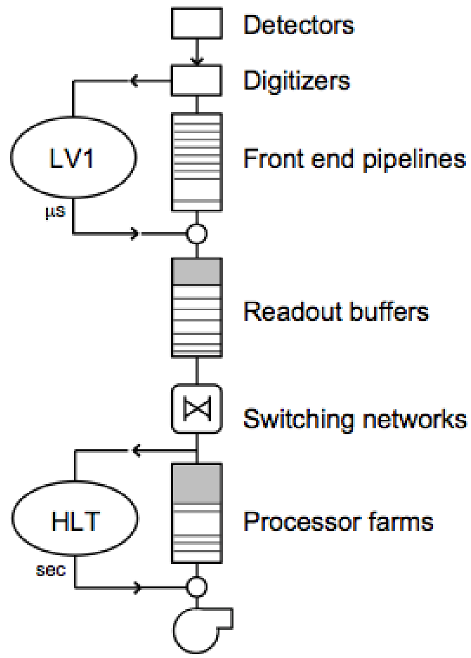


Figure 3.14: Schematic overview of the trigger system.[24]

hardware-based triggers to more sophisticated software-based triggers. Each level filters out a certain fraction of the readout data, until the amount of events is sufficiently small to be retained in its entirety. Since with this reduction of the data volume comes a choice of events, triggering has to take various physical observables into account in order to maximize the amount of "interesting" events that are kept.

Figure 3.14 shows an overview; LV1 is the Level 1 Trigger, and HLT stands for High Level Trigger.
More information on triggering at CMS can be found in [5]. A full discussion of the trigger systems can be found in [23] and [24].

3.7.2 L1 Trigger

The L1 Trigger is the first step in reducing the amount of data. It is a solely hardware-based triggering system, designed to reduce the incoming average data rate to a maximum of 100 kHz. It selects data with interesting signatures by processing information from the calorimeters and the muon systems.
The L1 Trigger has local, regional and global components. At the bottom end, the Local Triggers are based on energy deposits in calorimeter trigger towers and track segments or hit

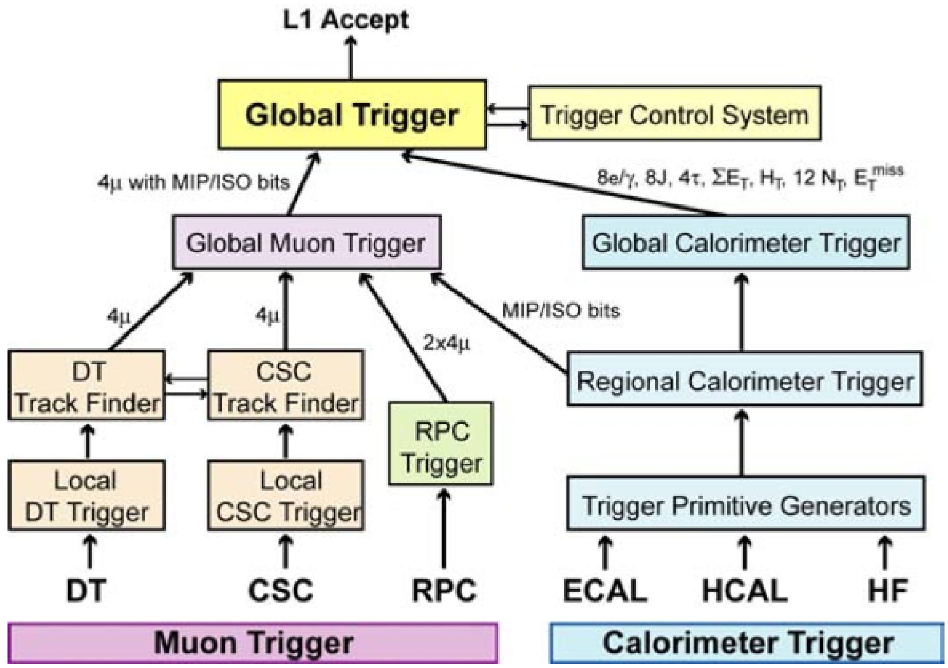


Figure 3.15: Architecture of the L1 Trigger.[5]

patterns in muon chambers, respectively. Regional Triggers combine their information and use pattern logic to determine ranked and sorted trigger objects such as electron or muon candidates in limited spatial regions. The Global Calorimeter and Global Muon Triggers determine the highest-rank calorimeter and muon objects across the entire experiment and transfer them to the Global Trigger, the top entity of the Level-1 hierarchy (see Fig. 3.15).

3.7.3 High Level Trigger

The second main step is the High Level Trigger (HLT). The HLT has access to the complete read-out data and can run reduced versions of the full reconstruction software on about 1000 processors in a filter farm. This way, the L1 accept rate of about 100 kHz is reduced to a final output rate of 100 Hz.

The HLT makes selections based on reconstructed particles (photons, electrons, muons) or more complex objects such as jets (also in combinations of multiple jets or jets plus missing energy). The simpler versions of b-tagging (the impact parameter-based track algorithms and the muon-based algorithms, see next chapter) can also be used to select jets originating from bottom quarks.

Chapter 4

Tracking Efficiency

4.1 Overview

Since tracks are the main ingredient to the methods described in chapter 5 and onward, this section will illustrate the efficiency of the tracker systems. Two data-driven methods to obtain tracking efficiencies will be explained: *track embedding* [17] and *muon tag-and-probe* [17]. A method to determine the resolution of an important derived value, the *IP resolution* [18], is also explained. Since jets from bottom hadrons are of special interest in the further chapters, the *b jet tracking efficiency* [19] (based on simulation data) will be illustrated as well.

4.2 Track Embedding

This method measures tracking efficiency by embedding simulated tracks in data events. It simulates an isolated charged particle, reconstructs the track using the hits generated in the silicon tracker, and then embeds these hits in a data event to determine if a track matching the original is successfully reconstructed. The total efficiency to reconstruct a track is split into two values: the *acceptance* and the *reconstruction efficiency*. The acceptance is defined as the probability that a charged particle produces a sufficient number of hits in the tracker to be reconstructed by the track-finding algorithm, while the efficiency is defined as the probability that these hits are used to reconstruct a track with parameters representative of those of the original particle. Figures 4.1 and 4.2 show these values for muon and pion tracks, respectively, embedded in data events and, as a cross-check, in simulated events.

4.3 Muon tag-and-probe

The tag-and-probe method is a standard technique for measuring lepton reconstruction efficiencies in data. It is based on the reconstruction of a well-known mass dilepton reso-

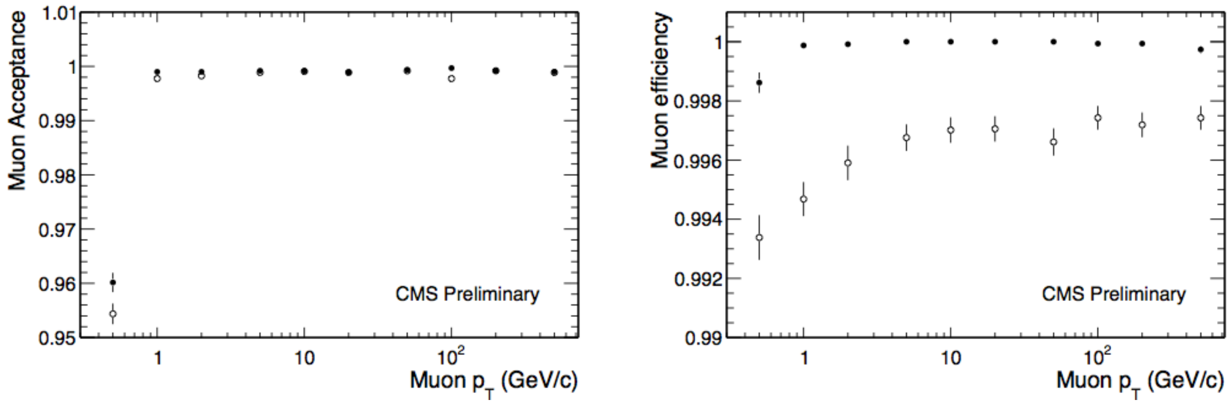


Figure 4.1: Muon tracking acceptance and efficiency, embedded in data (open circles) and simulated events (full circles). [17]

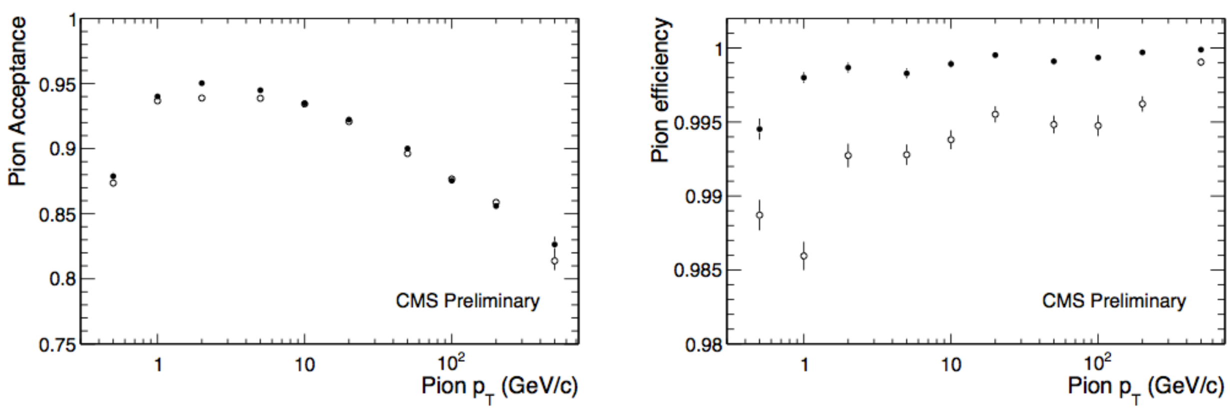


Figure 4.2: Pion tracking acceptance and efficiency.[17]

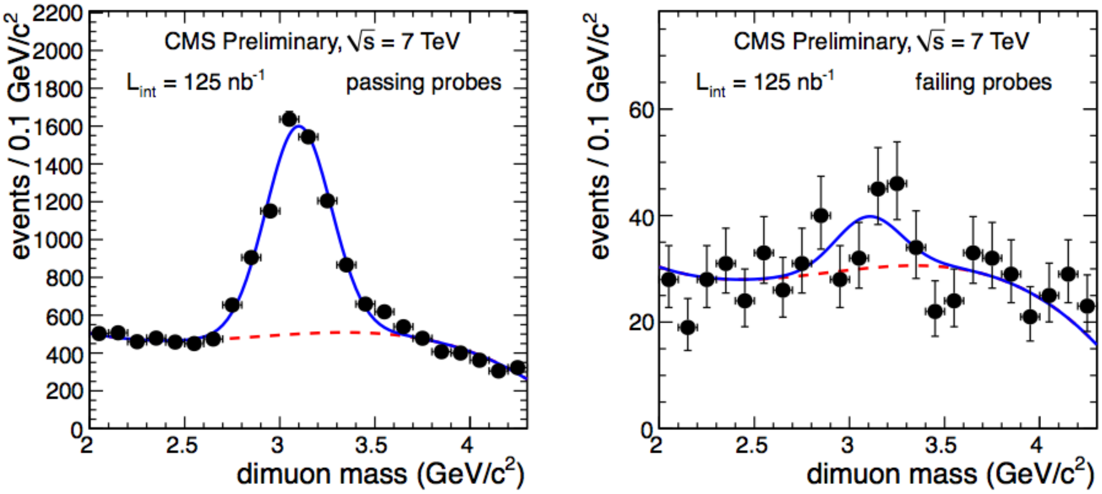


Figure 4.3: Dimuon invariant mass for tag muons paired with passing (left) and failing (right) probes. [17]

Region	Data Eff. (%)	Sim Eff. (%)	Data/Sim
$0.0 \leq \mid \eta \mid \leq 1.1$	$100.0^{+0.0}_{-0.3}$	$100.0^{+0.0}_{-0.1}$	$1.000^{+0.001}_{-0.003}$
$1.1 \leq \mid \eta \mid \leq 1.6$	$99.2^{+0.8}_{-1.0}$	$99.8^{+0.1}_{-0.1}$	$0.994^{+0.009}_{-0.010}$
$1.6 \leq \mid \eta \mid \leq 2.1$	$97.6^{+0.9}_{-1.0}$	$99.3^{+0.1}_{-0.1}$	$0.983^{+0.009}_{-0.010}$
$2.1 \leq \mid \eta \mid \leq 2.4$	$98.5^{+1.5}_{-1.6}$	$97.6^{+0.2}_{-0.2}$	$1.010^{+0.015}_{-0.016}$
Combined	$98.8^{+0.5}_{-0.5}$	$99.2^{+0.1}_{-0.1}$	$0.996^{+0.005}_{-0.005}$

Table 4.1: Measured tracking efficiency values from tag and probe on data and simulation. [17]

nance (such as the Z or J/ψ), where one lepton is chosen under strict quality criteria (the *tag*), and then compared to a set of other leptons meeting more loose quality criteria (the *probes*), requiring that the invariant dilepton mass is compatible with the chosen resonance. The efficiency is then determined by comparing the number of passing probes to the total number of probes. Figure 4.3 shows the the invariant mass distribution around the J/ψ resonance. Table 4.1 summarizes the tracking efficiency in data and simulated events in different η ranges.

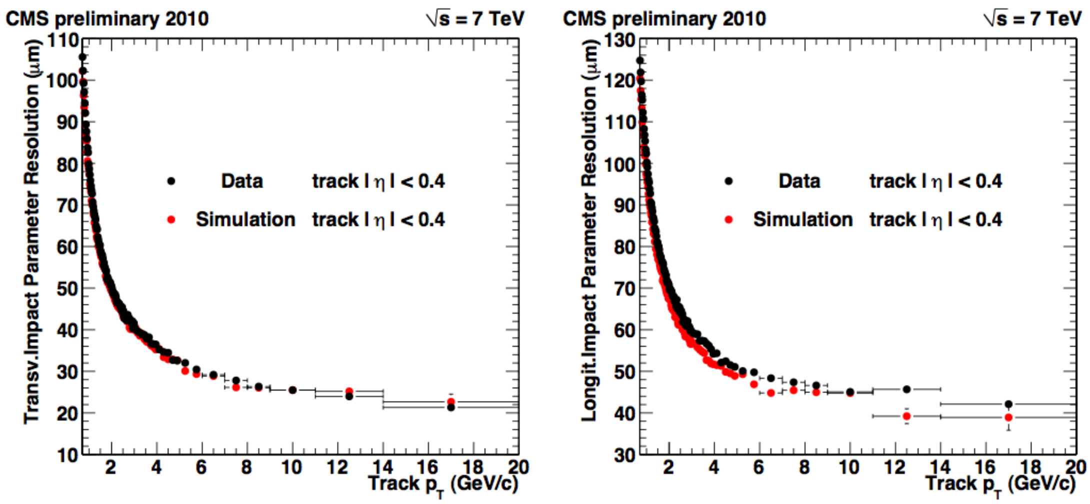


Figure 4.4: Measured resolution of the transversal (left) and longitudinal (right) track impact parameter.[18]

4.4 IP resolution

This method aims to obtain from data the resolution of a value that will play an important role in the algorithms in the later chapters: the track *impact parameter* (IP), the distance between the track and the (primary) vertex at the point of closest approach.

The method chooses a high-quality track and determines the primary vertex position using all other tracks in the event (this way, the position of the track can be compared to the unbiased vertex position). If the error on the vertex position is not too large, the transversal and longitudinal IP values of the track are evaluated in respect to the vertex. The IP value is then histogrammed in bins of track p_{t} and fit with two Gaussian distributions, one for the track IP uncertainty and one for the primary vertex reconstruction uncertainty. The width of the second Gaussian is estimated from simulation; the width of the first is defined as the IP *resolution* and extracted from the fit. The resulting IP resolution can be seen in Figure 4.4.

4.5 B jet tracking efficiency

The previous sections have dealt with the tracking efficiency of isolated particles and have demonstrated very high tracking efficiencies. These correspond to the *algorithmic efficiency*, which means the efficiency of correctly reconstructing tracks that meet certain quality criteria. Another relevant value to consider is the *global efficiency*, which means the reconstruction efficiency for *all* tracks. Due to geometric effects and a higher density

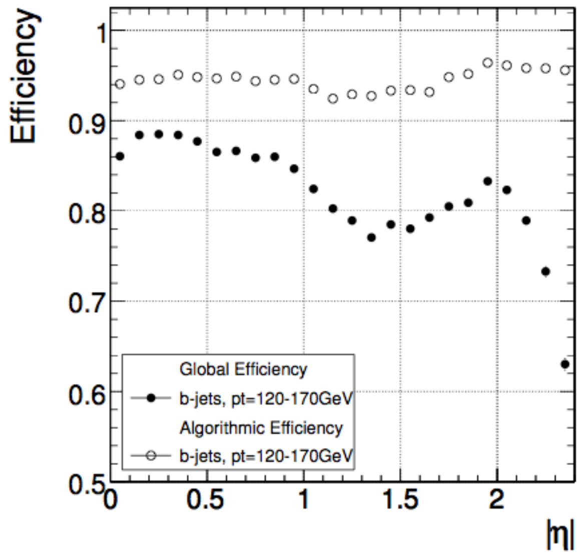


Figure 4.5: B jet algorithmic and global efficiency vs. pseudorapidity. [19]

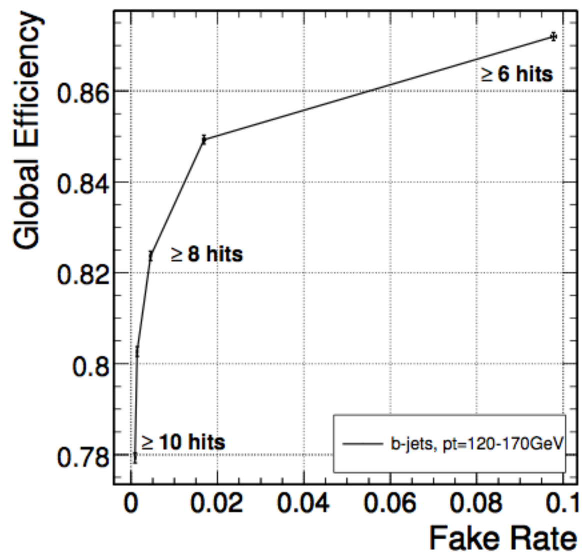


Figure 4.6: Global efficiency vs. fake rate, depending on hits in tracker.[19]

of tracks, it can be seen in Figures 4.5 and 4.6 for b jets with 120 GeV $< p_{\mathrm{t}} <$ 170 GeV that the global efficiency in jets is lower than the efficiency of isolated tracks.

Chapter 5

B Tagging

5.1 Introduction

Many analysis scenarios at CMS, both within the Standard Model and in the search for new physics, rely on the identification of jets (narrow showers of particles which originate from hadronization processes) coming from bottom quarks.
Due to the B hadrons' relative longevity ($c\tau \approx 480$ µm), large mass ($m_{\mathrm{b}} \approx 4.2$ GeV), the hard fragmentation, and the relatively high fraction of semileptonic decays, b jets can be distinguished from jets originating from gluons, light quarks and charm quarks with various algorithms using the different detector subsystems of CMS.
Full details on the algorithms described in the following sections can be found in [27] and [28].

5.2 Observables

The most powerful single-track observable is the *impact parameter* (distance between track and vertex at the point of closest approach, Fig. 5.1). For B hadrons the IP is Lorentz-invariant and can be calculated either in the transverse plane or in 3D; the good z resolution of the CMS pixel system allows 3D reconstruction to be used. Due to the fact that the uncertainty can be of the same order of magnitude as the IP, a better observable for b-tagging is the *impact parameter significance* s, defined as $s = \frac{IP}{\sigma_{IP}}$.
The IP is "life time signed". The IP sign is obtained from the sign of the scalar product of the IP segment with the jet direction. A "sign flip" can happen due to differences between the reconstructed jet axis and the true B hadron flight direction. For decays without a sizeable lifetime, the IP is expected to be symmetric with respect to zero; for B hadrons decaying weakly, it is mostly positive.

Another ingredient to identify b jets is the identification of of vertices, i.e. the finding of points of common origin in a set of tracks. Vertices can be categorized as *primary*

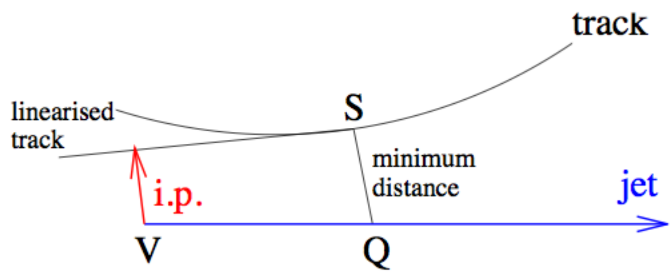


Figure 5.1: Schematic representation of a track's impact parameter.[27]

(the point of collision) and *secondary* (the point of decay of a particle). By finding such secondary vertices, decays of long-lived particles such as K_s and B mesons can be identified. Finally, due to the high amount of semi-leptonic b decays, reconstructed leptons (electrons in the tracker and ECAL, or muons in the tracker and muon systems) are also used to identify b jets.

5.3 Algorithms

All algorithms in the b-tagging framework of CMS produce a single numerical value, the *discriminator*. With this value, jets can be selected according to the priority for either efficiency or purity. A loose cut on the discriminator means that more jets will be tagged as b, but as a consequence the mistag rate of accepting non-b jets will be higher as well. On the other hand, a tighter cut means a higher probability that each jet will in fact be a b jet, but fewer jets will be tagged overall. Generally, for the cut three working points are used: loose, medium, and tight, which correspond to a light jet tag rate of about 10%, 1% and 0.1%, respectively. In addition, some of the algorithms come in two variations, called *high efficiency* and *high purity*, the choice of which again results in either more tagged jets or a lower mistag rate.
The discriminator can be a simple physical quantity like a distance significance or a more complicated variable like the output of a likelihood ratio or neural network.

5.3.1 Track Counting

The track counting algorithm (TC) considers a set of tracks, ordered by decreasing IP significance. The set of tracks is determined by association to the jet in question: the tracks have to be within a cone in $\Delta\eta\Delta\phi$ of a certain radius around the jet axis.
The algorithm identifies a jet as a b jet if there are at least N tracks each with an IP significance exceeding a value S. The algorithm returns the significance of the Nth track as its discriminator, which is the lowest IP significance exceeding S due to the ordering of

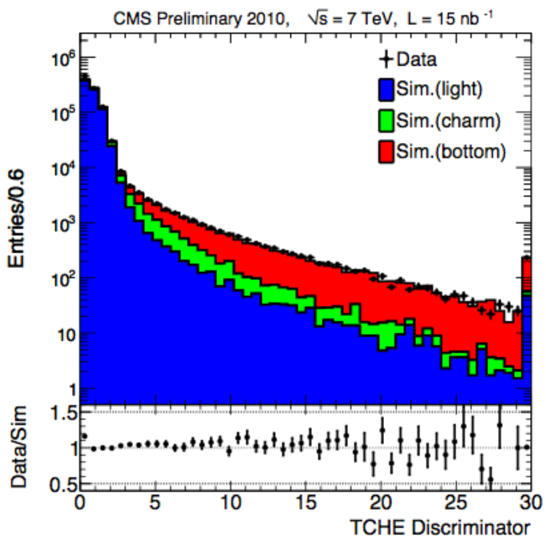


Figure 5.2: TC HighEff Discriminator.[28]

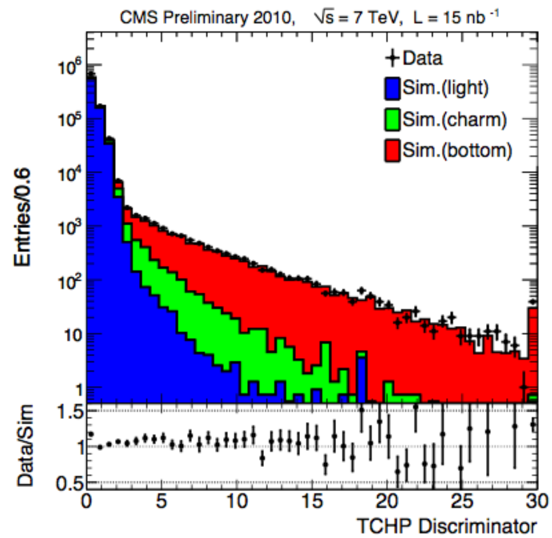


Figure 5.3: TC HighPur Discriminator.[28]

the tracks.
The algorithm has two major parameters (N and S). For high efficiency the second track can be used (Fig. 5.2); for higher purity the third track is a better choice (Fig. 5.3). This is due to the fact that there are on average 5 charged tracks coming from a B hadron decay; combined with a track reconstruction efficiency of about 80% (as seen in section 3.3.3), this means an average of 4 reconstructed tracks per B decay. In these and many following figures, the Data/Sim section below indicates the ratio between the respective value on real data and on Monte Carlo simulation.
The advantage of the track counting algorithm is its simplicity and robustness; on the other hand, it does not use the entirety of the information available in the tracker.

5.3.2 Jet Probability

The jet probability algorithms are a natural extension of the track counting algorithms. The idea is to combine the information coming from all selected tracks used in the TC algorithm. Two discriminators are provided; the first, labelled *jet probability* (JP), is strictly related to the combined probability that all the the tracks in the jet come from the primary vertex, defined as

$$P_{\text{jet}} = \Pi \cdot \sum_{j=0}^{N-1} \frac{(-\ln \Pi)^j}{j!}, \tag{5.1}$$

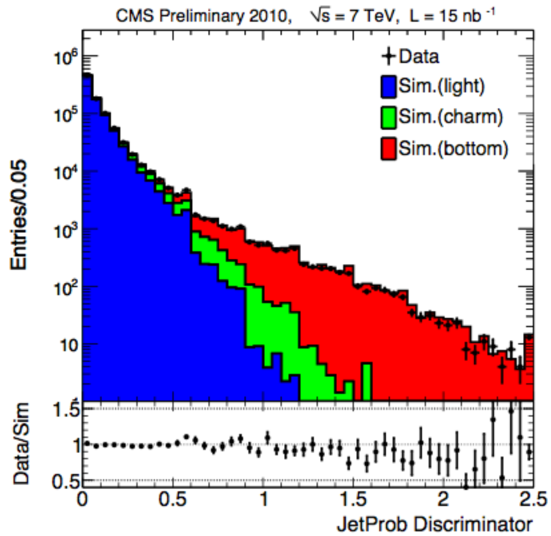


Figure 5.4: JP Discriminator.[28]

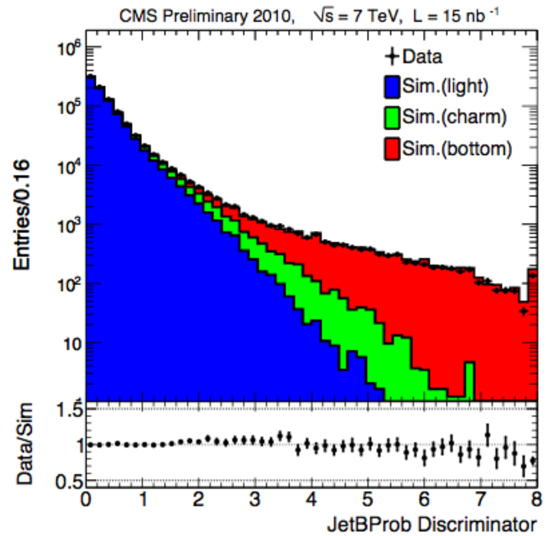


Figure 5.5: JBP Discriminator.[28]

where

$$\Pi = \prod_{i=1}^{N} P_{\mathrm{tr}}(i) \tag{5.2}$$

and the $P_{\mathrm{tr}}(i)$ are probabilities indicating how likely it is that a given track i comes from the primary vertex. These are obtained by integration of the resolution function R in regard to the significance s_i of track i,

$$P_{\mathrm{tr}}(s_i) = \mathrm{sgn}(s_i) \cdot \int_{|s_i|}^{\infty} R(x)dx. \tag{5.3}$$

The resolution function, because it is not easily modelled, is obtained by numerical integration over histogrammed negative IP significances, since those are the ones most likely to come from the primary vertex. Figure 5.4 shows the resulting discriminator.

A variant of the algorithm, labelled *jet B probability* (JBP) (Figure 5.5), estimates how likely it is that the four most displaced tracks are compatible with the primary vertex.

In comparison to the track counting algorithm, the jet probability algorithms are more complicated, but use the full information of the tracking system, which translates into a higher purity at comparable efficiency.

More details on track IP-based b-tagging can be found in [29] and [30].

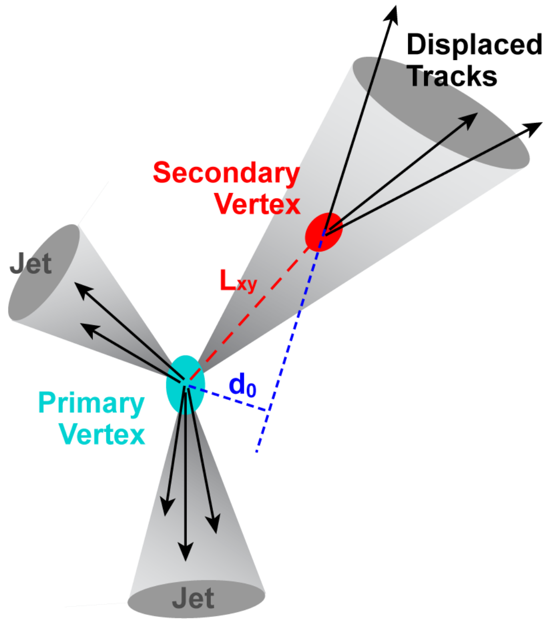


Figure 5.6: Schematic representation of a B decay event.[25]

5.3.3 Simple Secondary Vertex

The simple secondary vertex algorithm (SSV) identifies a jet as a b jet based on a reconstructed secondary vertex (illustrated in Fig. 5.6). The algorithm returns the significance of the 3D flight distance of the vertex as discriminator; if no vertex is found, the algorithm returns no discriminator, limiting its maximum efficiency to the probability of finding a vertex in a B hadron decay (around 60-70%).
Figure 5.7 shows the high efficiency discriminator (at least 2 tracks associated to the vertex), and Figure 5.8 shows the high purity discriminator (at least 3 tracks).
The advantage of the SSV algorithm is that it uses topological information (the position of the tracks in relation to each other), which yields a higher purity, at the cost of efficiency in comparison to the simpler algorithms.

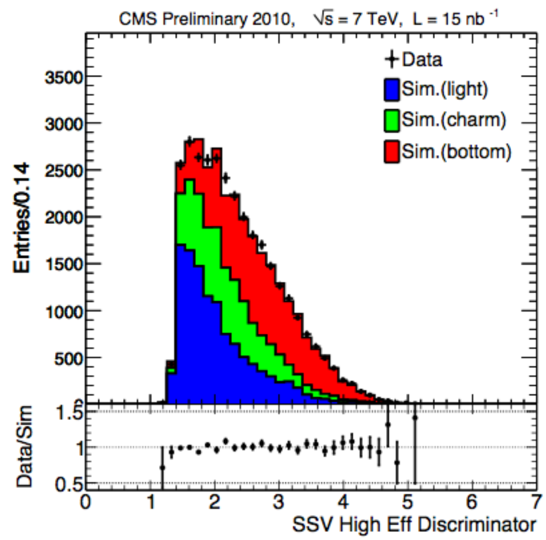


Figure 5.7: Simple Secondary Vertex High Eff Discriminator.[28]

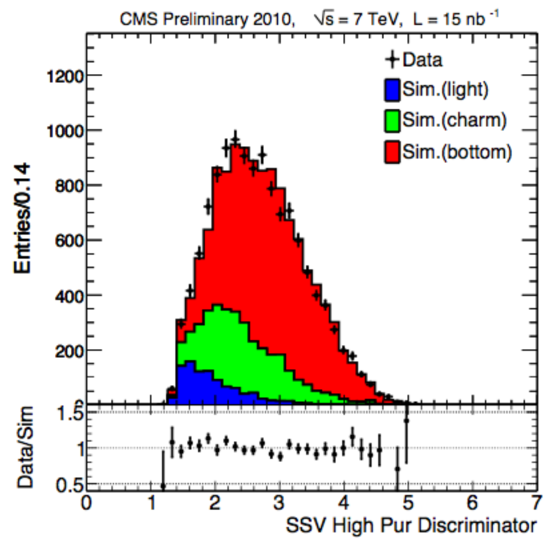


Figure 5.8: Simple Secondary Vertex High Pur Discriminator.[28]

5.3.4 Combined Secondary Vertex

A more complex approach involves the use of secondary vertices, together with other lifetime information like the IP significance or decay lengths. By using these additional variables, the combined secondary vertex algorithm provides discrimination even when no secondary vertices are found, so the maximum possible b-tagging efficiency is not limited by the secondary vertex reconstruction efficiency. In many cases, tracks with an IP significance > 2 can be combined, allowing for the computation of a subset of secondary vertex-based quantities even without an actual vertex fit. When even this is not possible, a "no vertex" category reverts simply to track-based variables similarly to the jet probability algorithm.

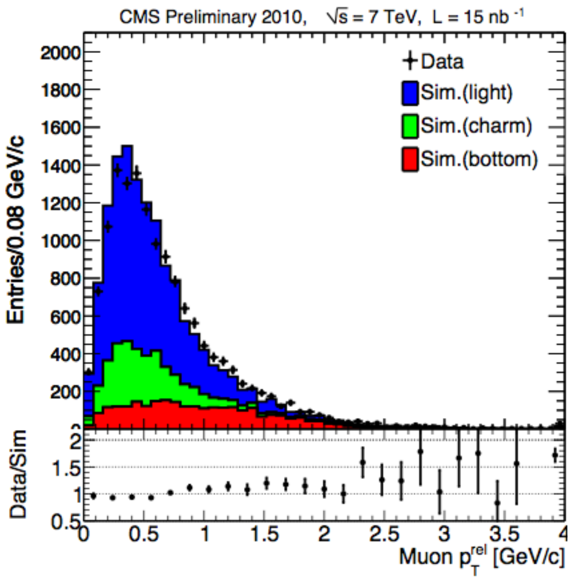


Figure 5.9: Muon p_{t} Discriminator.[28]

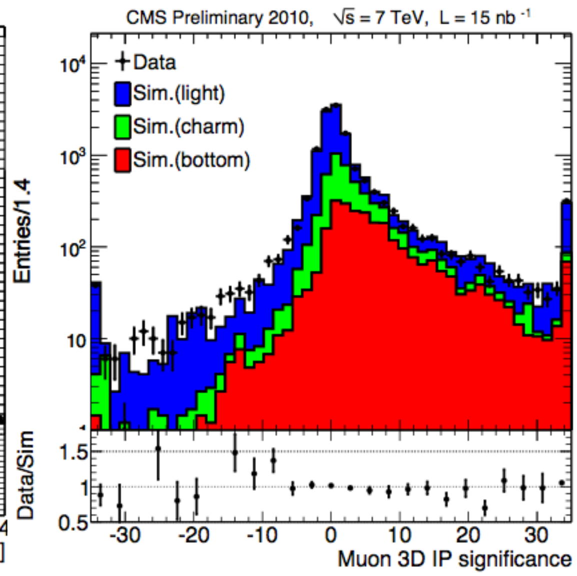


Figure 5.10: Muon IP Discriminator.[28]

5.3.5 Lepton-Based

The presence of a muon close to the jet is already a hint of a weak decay of a B hadron. This can be complemented with additional quantities in order to build a discriminator. In the *soft muon by* $p_{\mathrm{t_{rel}}}$ algorithm the p_{t} of the muon (Fig. 5.9) with respect to the jet axis is used; harder cuts yield higher purities. In the *soft muon by IP significance* algorithm the IP significance of the muon (Fig. 5.10) is used instead, but only when found to be positive. In all cases, when more than one muon is reconstructed, the one with the highest discriminator value is used.

Chapter 6

Vertex Fitting

6.1 Overview and Classical Approach

As described in the last chapter, an important ingredient in the identification of b decays is the finding of secondary vertices. These need to be properly reconstructed and, since the track multiplicity in a b secondary vertex is generally an order of magnitude smaller than in the primary collision vertex, correctly separated from the primary vertex.
In general, one is interested in the position v of the common vertex of a set of tracks, and the momentum vectors p_i of the associated tracks with their respective covariance matrices V_i. These values are obtained from estimated track parameters $\tilde{q}_i$ at some reference surface (see Fig. 6.1).

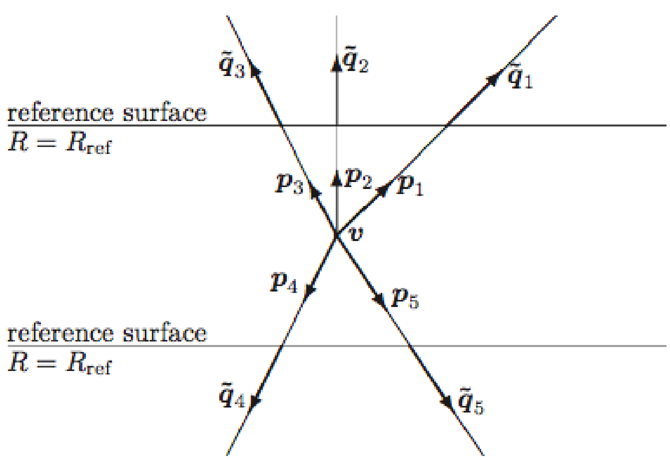


Figure 6.1: Vertex and track parameters for a vertex with 5 tracks. Reference surface is a cylinder with radius R_{ref}. [31]

The classical approach to this problem is with a least-squares method (LSM). This involves the minimization of an objective function F_{obj},

$$F_{\text{obj}}(v, p_1, \cdots, p_n) = \sum_{i=1}^{n} e_i^T G_i e_i, \tag{6.1}$$

$$e_i = \tilde{q}_i - q_i, \qquad V_i = G_i^{-1}. \tag{6.2}$$

Since this is a non-linear problem, effective calculation usually involves approximations of some kind, e.g. in the Gauss-Newton method. Using this method, the system can be reduced to a linear regression problem.
An effective method to estimate a linear regression problem (effective in the sense that it requires an order of magnitude fewer operations to estimate the desired values than a general regression problem) is an algorithm based on the Kalman filter and smoother.

6.2 The Kalman Filter

The Kalman Filter is a recursive estimator of the state vector of a (discrete) linear dynamic system. The evolution of such a system is described iteratively by defining at each step k the *system equation* (representing the state of the system at k as a function of the system at step $k-1$) and the *measurement equation* (representing what is actually measured).

Generally, the system equation of a state vector (such as, in our case, the position of the vertex v) is described by a linear transformation F plus a random disturbance (the process noise w),

$$v_k = F_{k-1} v_{k-1} + w_{k-1}. \tag{6.3}$$

In the case of vertex fitting, this is very simple, as there is no process noise:

$$v_k = v_{k-1} = v. \tag{6.4}$$

The general form of a measurement equation is a linear function of the state vector plus the measurement uncertainty,

$$p_k = H_k v_k + \epsilon_k. \tag{6.5}$$

The Kalman filter updates at each step the system's state vector by considering the previous state vector and the current measurement. This allows for the determination of the value of the system vector at a given step k, which can be spatial (as in track fitting), temporal (as in radar tracking), or a dimensionless integer, as in the fitting of vertices. In our case, the filter updates the position of a vertex with $k-1$ tracks to a vertex with k tracks until all supplied tracks have been considered.

Additionally, using all the measurements taken up to a step k, one can also estimate the state vector at a step $k' < k$ in a process called *smoothing*. In our case, this does not actually contribute to the quality of the vertex fit (since a vertex with k tracks already contains the maximum amount of information), but it can be used for *track refitting*, which means the improvement of the reconstructed track parameters with the additional information that the tracks come from a shared point of origin.

The Kalman filter has many advantages: estimates are unbiased, consistent, and the filter minimizes the mean square estimation error; if both process noise and measurement uncertainty are Gaussian random variables, the Kalman filter is the optimal filter. In other cases, it is simply the optimal linear filter. The least-squares method shares these properties; the advantages of the Kalman filter lie in its efficiency and flexibility.
A more detailed overview of Kalman filters in vertexing and track fitting can be found in [32] and [33].

6.3 Adaptive Vertex Fitting

A drawback of the Kalman filter, however, is its lack of robustness: outlying measurements have a large influence on the estimate, which can lead to serious distortions. By an improvement of the algorithm, vertex fitters can be made robust, i.e. insensitive to outlying or mismeasured tracks; one such improved algorithm is the Adaptive Vertex Fitter (AVF) [34].

The AVF is an iterative weighted Kalman filter. Each track is given a weight which is a function of the reduced distance between the tracks and the smoothed vertex position. If χ_i^2 denotes the square of the standardized residual, then the weight w_i equals

$$w_i(\chi_i^2) = \frac{\exp(-\chi_i^2/2T)}{\exp(-\chi_i^2/2T) + \exp(-\sigma_{cut}^2/2T)}. \tag{6.6}$$

A cutoff parameter σ_{cut}^2 has been introduced which is defined as the standardized residual for which $w_i = 0.5$, and a "temperature" parameter T, which defines "softness" of the weight function. This weight can be interpreted as a track-to-vertex assignment probability. Instead of minimizing the least sum of squares, the algorithm now minimizes the weighted least sum of squares. In order to avoid falling prematurely into local minima, a deterministic annealing schedule is introduced: in each iteration step the temperature parameter T is lowered. Typically, a “quasi-geometric” annealing schedule that converges towards 1 is employed:

$$T_i = 1 + r \cdot (T_{i-1} - 1). \tag{6.7}$$

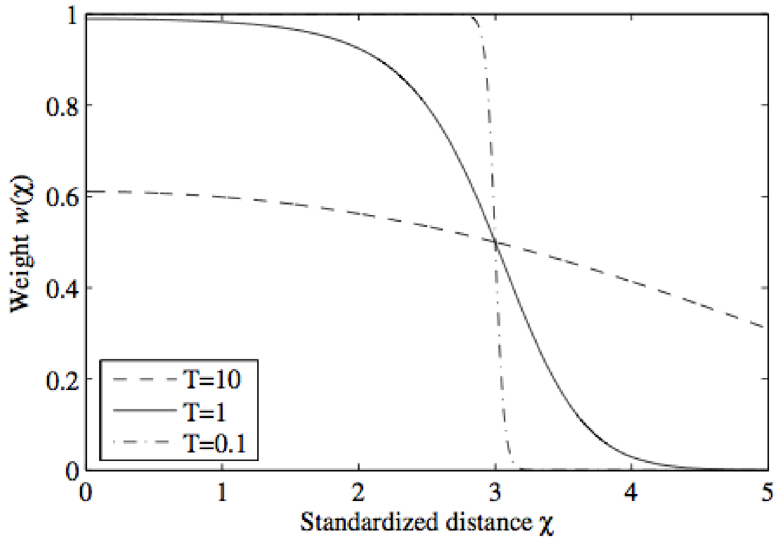


Figure 6.2: Weights of the AVF for different values of T. σ_{cut} has been chosen to be 3.0.[34]

Here, T_i refers to the temperature parameter T at iteration i, and r denotes the annealing ratio. For convergence, $0 < r < 1$ is needed. Figure 6.2 shows the weight as function of χ_i^2 and T.

Chapter 7

CMS Software Framework

7.1 Overview

The CMS Software Framework (CMSSW) is the main framework in which data at CMS is processed and analyses are run. The CMS Event Data Model (EDM) is based on the concept of an event; as mentioned in the section on triggering, an event is the collection of all data pertaining to a single bunch crossing. From the CMS software point of view, an event is a C++ object container for all readout and reconstructed data. These containers are accessed by plug-in modules which contain all the algorithms for event reconstruction and analysis. A single executable, cmsRun, executes these modules; only the required modules for a given task need to be loaded at the beginning of a job. The modules to be used are steered via Python job configurations.

The framework modules generally belong to three main types:

- *EDAnalyzers*
  Access objects from the event and run analysis algorithms. EDAnalyzers can produce output files, but do not manipulate the event itself.

- *EDProducers*
  Access the event and produce new objects in the event; the new content is retained and can be accessed by other framework modules.

- *EDFilters*
  Filter the event flow based on certain criteria; return boolean values to keep or discard individual events.

7.2 An Example – SSV B Tag Event Flow

To illustrate the way event analysis is handled in CMSSW, the workflow of the Simple Secondary Vertex B Tag will be explained here. This is also of interest since it's the underlying workflow that will be modified in chapter 8.

The first thing one needs to consider (apart from the data set one wishes to run on) is triggering. Usually one chooses a high level trigger path – which corresponds to a certain choice in triggering – in an EDFilter (for instance, the HLT_Jet30U trigger path requires at least one jet with p_{t} exceeding 30 GeV). Also, certain cleaning steps might be necessary, depending on the quality and properties of the data set in question.

An EDM event contains all reconstructed physical objects; of interest to us are tracks and jets, since we want to identify jets as b jets based on vertex reconstruction. Figure 7.1 shows the EDProducer steps that lead from tracks and jets to the final discriminator result. Each producer module adds a new collection of objects to the event, which is then used as input to the next module. The modules used, and their basic properties, are:

- *JetTrackAssociationProducer*
  Creates an association map between tracks and jets. By default tracks have to be within a cone of radius 0.5 in $\Delta\eta\Delta\phi$ around the jet axis to be considered.
- *TrackIPProducer*
  Produces track impact parameter information in regard to each jet.
- *SecondaryVertexProducer*
  Performs the vertex fit (using the AVF) separately for the set of tracks from each jet.
- *JetTagProducer*
  Creates an association map between the jets and the discriminator value obtained from the vertex fit.

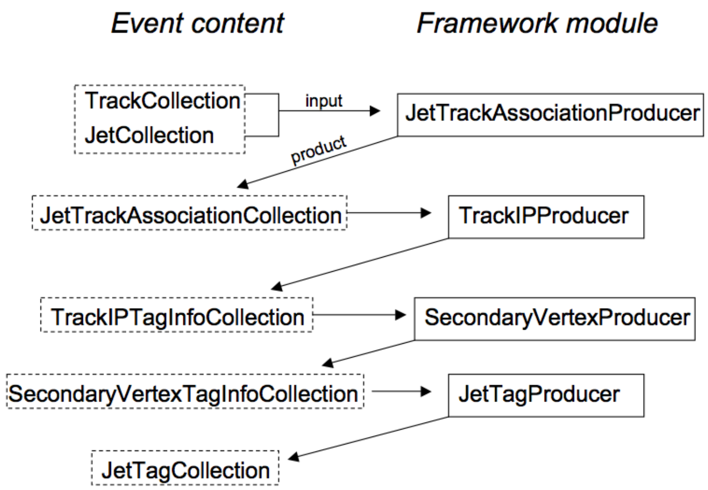


Figure 7.1: Schematic of the SSV b-tag workflow.

7.2.1 Jet & Track Selection

AntiKt5[1] particle flow[2] jets are used, chosen according to the following criteria:

- transversal momentum $p_\mathrm{t} > 30$ GeV;
- pseudorapidity $|\eta| < 2.4$.

The tracks associated to the jet are filtered according to the following criteria:

- at least 8 total reconstructed hits in the tracker;
- at least 2 reconstructed hits in the pixel detector;
- $p_\mathrm{t} > 1$ GeV;
- angle (in $\Delta\eta\Delta\phi$) between jet and track momentum $\Delta R < 0.3$;
- track quality 'highPurity'[3].

[1] Anti-k_t is a jet clustering reconstruction algorithm; details on this method can be found in [35].
[2] Particle flow is an event reconstruction method that aims at identifying particles in the detector by combining information from the various detector subsystems, see [36].
[3] The flag 'highPurity' refers to a collection of quality criteria from track reconstruction; a complete definition of all aspects can be found in Appendix A of [37].

7.2.2 Vertex Reconstruction

Each set of tracks from the considered jets is fitted with the Adaptive Vertex Fitter. The resulting vertices are filtered kinematically to exclude primary vertices:

- Transverse distance to primary vertex: 100 µm $< d_{\mathrm{t}} <$ 2.5 cm.
- 3D distance to primary vertex: $d_{\mathrm{3D}} <$ 2 cm.
- Transverse significance (i.e., transverse distance to primary vertex divided distance uncertainty): $\frac{d_{\mathrm{t}}}{\sigma_{d_{\mathrm{t}}}} > 3$.
- Track multiplicity ≥ 2.
- No more than 65% of the tracks shared with the primary vertex.
- Angle between vertex and jet axis: $\Delta R < 0.5$.
- Mass: $m_{\mathrm{SV}} <$ 6.5 GeV.
- Vertices with two oppositely charged tracks with an invariant mass within ± 50 MeV around the K_S^0 mass are removed.

7.2.3 Results

Since we are interested in the performance of a b-tagging algorithm, we run the sequence on a B-enriched Monte Carlo simulation dataset:

- InclusiveBB_Pt30/Summer10-START36_V9_S09-v1.

The discriminator D of the SSV algorithm is defined as

$$D = \ln(1 + s_{\mathrm{3D}}), \tag{7.1}$$

where s_{3D} is the 3-dimensional flight distance significance of the reconstructed vertex ($s_{\mathrm{3D}} = \frac{d_{\mathrm{3D}}}{\sigma_d}$). Figures 7.2 and 7.3 show the high efficiency and high purity discriminator respectively. These plots are analogous to the bottom quark part in figures 5.7 and 5.8.

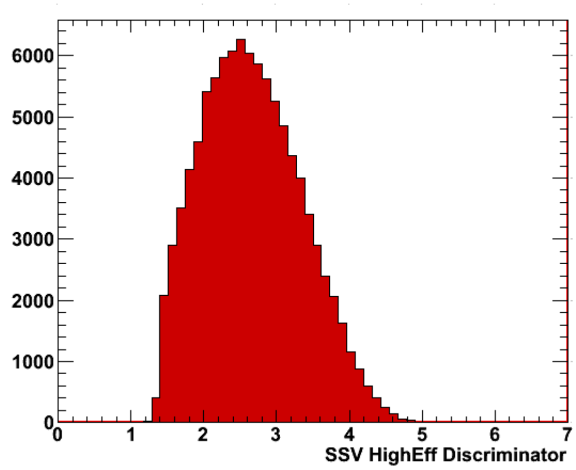


Figure 7.2: High Efficiency Discriminator.

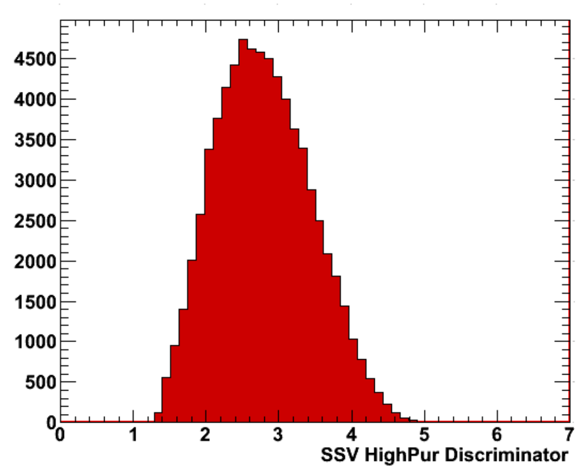


Figure 7.3: High Purity Discriminator.

Chapter 8

Calibration of Secondary Vertex Efficiency

8.1 Introduction

In order to measure the performance of the described b-tagging algorithms, their efficiency needs to be calibrated, which is done by comparing simulation data with the results of the algorithm running on such data. From the amount of B hadron decays generated by the simulation and the amount of b jets found by the algorithms their efficiency can be determined.

Monte Carlo simulation data is generated in two steps: simulating the physics (usually done with PYTHIA [38]), and simulating the detector response (done with GEANT4 [39]). The conclusiveness of efficiencies on Monte Carlo, however, is limited, since it has to model detector effects, such as misalignment and detector inefficiencies.

For the estimation of secondary vertex reconstruction efficiency, a different approach is proposed: modifying a light-quark jet on real data so that it kinematically resembles a b jet, by moving a certain amount of reconstructed tracks along the jet axis. In regard to this pseudo-b decay vertex, efficiencies can be obtained in a wholly data-driven way, since nothing is used but reconstructed data to estimate the efficiency; effects from detector misalignment, track quality and track resolution are automatically included this way.

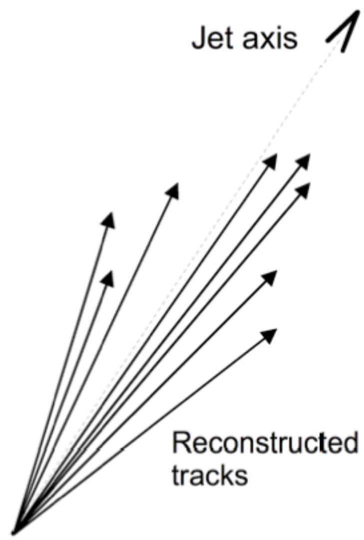


Figure 8.1: Light quark jet.

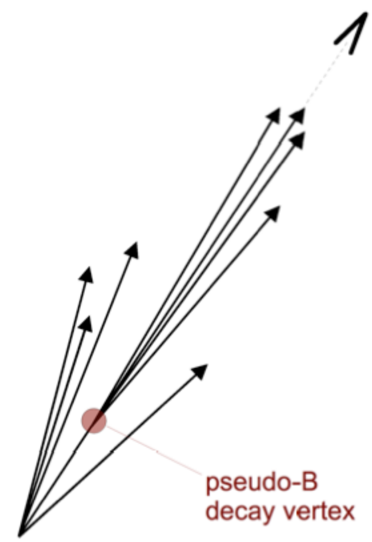


Figure 8.2: Pseudo-b jet.

8.2 Algorithm Description

8.2.1 Overview and Aim

The algorithm generates the pseudo-secondary vertex by choosing tracks from a high-energetic reconstructed jet and moving them along the jet axis. Then, on the set of tracks of this modified jet a vertex fit is performed.
Using this method, the impact of track reconstruction, track resolution, detector effects, and the power of the vertex fitting algorithm can be measured directly on data. However, the method offers no insight into the dynamics of b decays since it is a strictly kinematic modelling of secondary vertices. Therefore, the effects of different b decay chains or the fragmentation of b quarks are not considered in regard to the reconstruction efficiency.

8.2.2 CMSSW Workflow

The algorithm modifies the secondary vertex workflow illustrated in section 7.2 by adding an EDProducer module (TrackMover) at the beginning which chooses and moves the tracks in the light quark jet. Then, the framework sequence is followed, until an EDAnalyzer module (VertexAnalyzer) creates the efficiency and kinematic output plots (see Fig. 8.3).

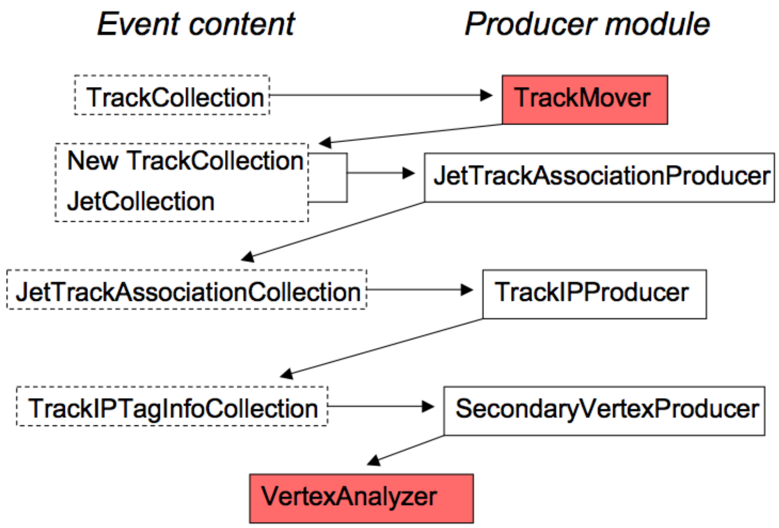


Figure 8.3: Flow diagram of the algorithm. The red modules indicate the modifications of the standard SSV b-tag sequence.

8.2.3 Number of Tracks & Flight Distance

Two quantities, the number of tracks to be moved n and the flight distance f, are modelled according to to distribution functions obtained from generator b decays (Figures 8.4 and 8.5). The algorithm chooses these values at random, then moves the n tracks with the highest p_t that meet the quality criteria (as in section 7.2.1) by f along the jet axis (more precisely, along the direction of the sum of the tracks' momenta). If not enough tracks are available, the event is discarded and no move is performed.

8.2.4 Data Sets and Selection Criteria

The following data sets have been used:

- Monte Carlo sample: QCD_Pt30/Summer10-START36_V9_S09-v1
  Event selection: L1 Technical Trigger Bits (40 || 41) to ensure beam crossing.

- Data sample: JetMETTau/Run2010A-PromptReco-v4
  Event selection: HLT_Jet30U trigger, good quality run selection from 132440 to 140401.

Tracks are moved in the highest-p_t jet above 30 GeV. The resulting vertices are filtered according to the same criteria as in section 7.2.2 to exclude primary vertices.

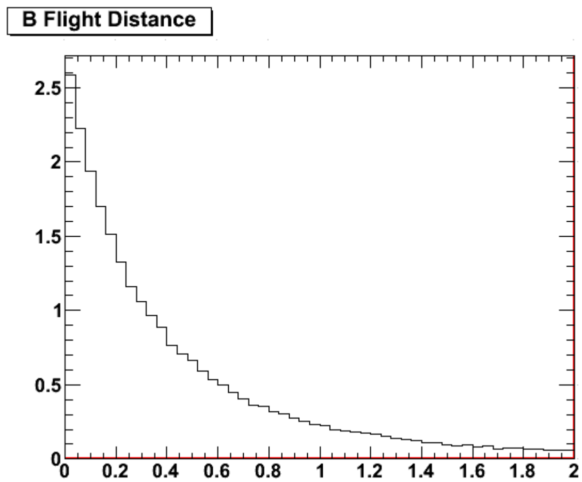


Figure 8.4: Flight distance [cm].

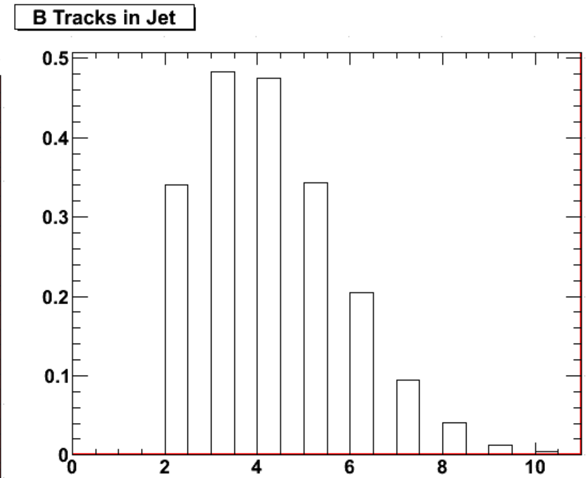


Figure 8.5: Number of reconstructed tracks in b decays.

8.2.5 Results

The reconstruction efficiency ε is defined as the number of events where a secondary vertex meeting the selection criteria has been found in the jet divided by the number of events where a pseudo-vertex has been created.
Fig. 8.6 shows the reconstruction efficiency versus the flight distance for pseudo-vertices created on Monte Carlo and on real data.
In addition, kinematic variables have been studied:

- Reconstructed secondary vertex mass (Fig. 8.7);
- Number of tracks associated to the vertex (Fig. 8.8);
- Transversal momentum of the jet (Fig. 8.9);
- Fraction of reconstructed SV energy / total jet energy (Fig. 8.10);
- Angle between reconstructed SV direction and jet direction (Fig. 8.11).

Note that all plots have been re-weighted so that the jet p_{t} distributions match; consequently, there is no difference between the curves in the jet p_{t} plot (Fig. 8.9).

8.2.6 Secondary Vertex Resolution

Using a method similar to the one described in section 4.4 (IP resolution), the resolution of secondary vertices can be determined in a data-driven way. The full set of tracks from

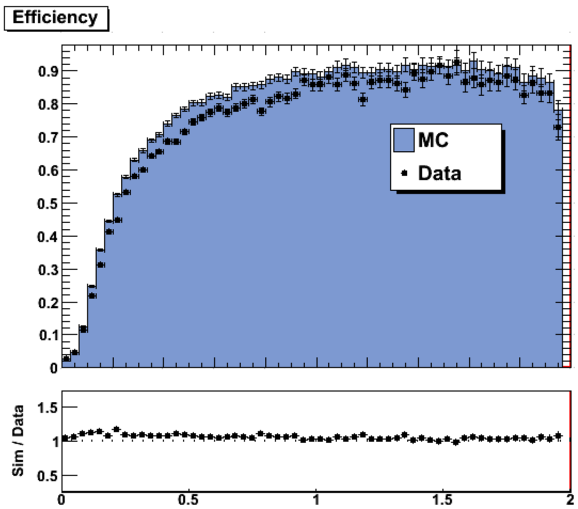


Figure 8.6: Reconstruction efficiency vs. flight distance [cm].

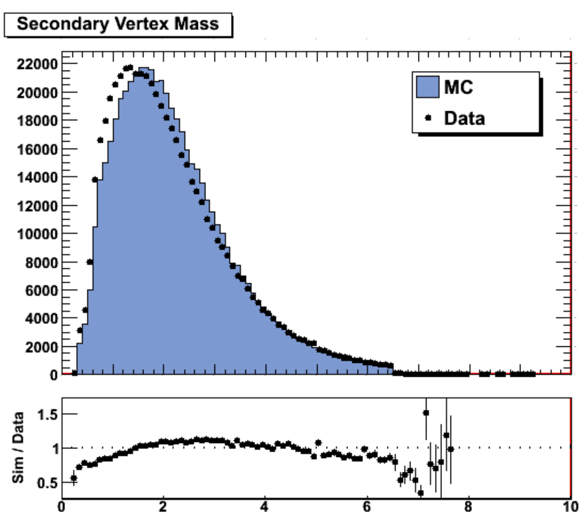


Figure 8.7: Sec. vertex mass [GeV].

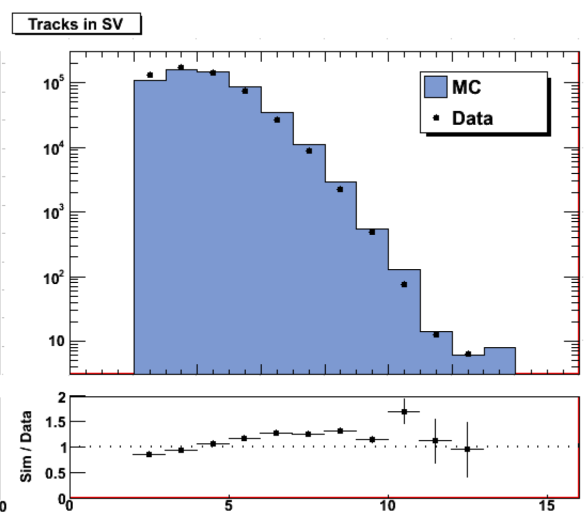


Figure 8.8: Tracks in secondary vertex.

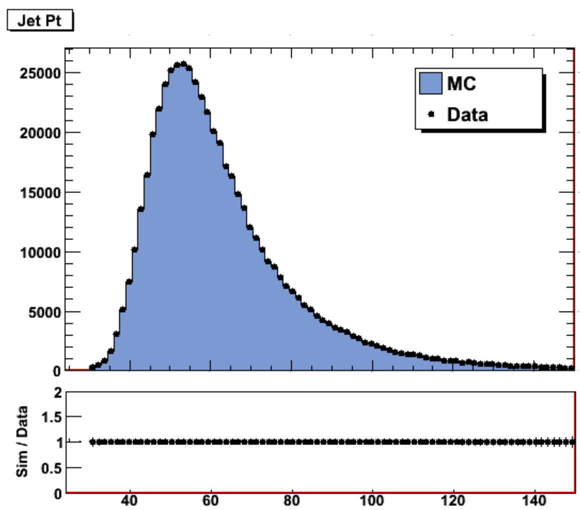


Figure 8.9: Jet p_t [GeV].

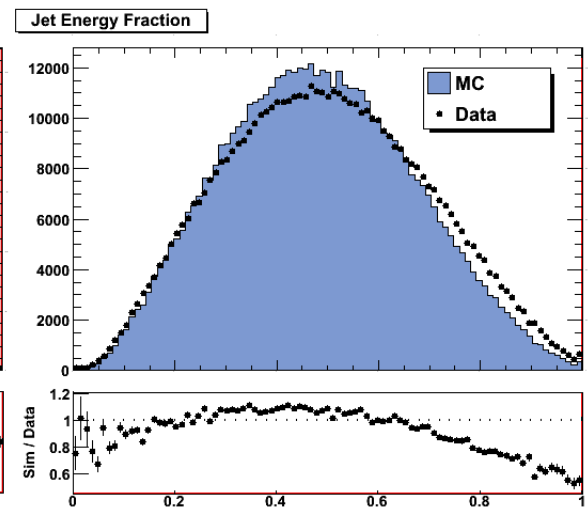


Figure 8.10: SV energy divided by jet energy.

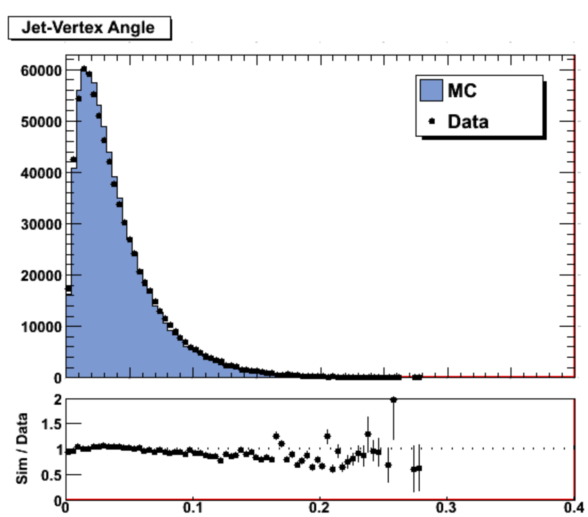


Figure 8.11: Angle bw. jet and vertex axis [rad].

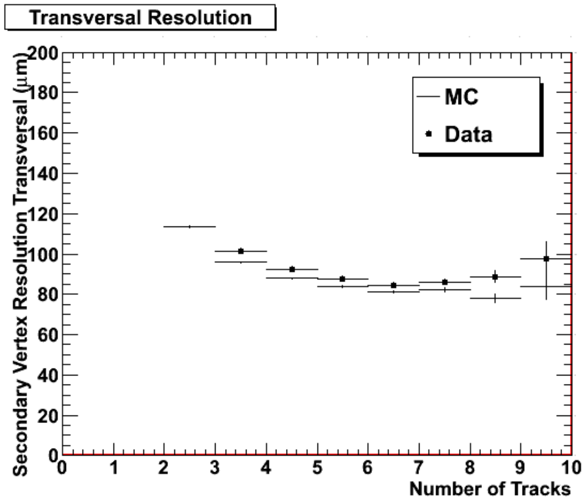


Figure 8.12: Transversal resolution.

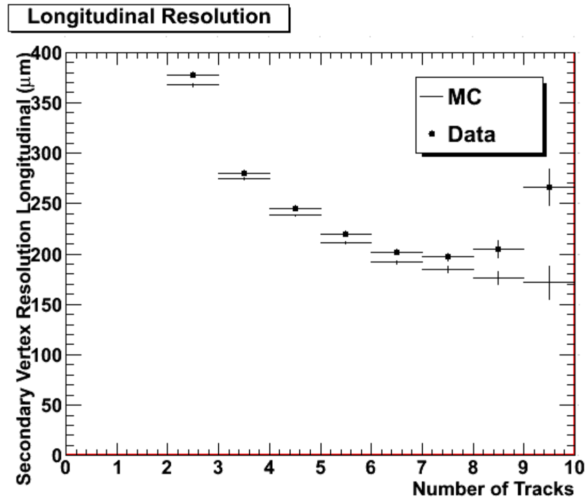


Figure 8.13: Longitudinal resolution.

the event, except the tracks that have been moved to generate the pseudo-vertex, are fitted again with the AVF. This yields an unbiased primary vertex position, which is then shifted by the same vector as the pseudo-vertex. We take this value as the simulated secondary vertex position $v_{\rm sim}$, and obtain resolution and pull distributions in the following ways:

- *Resolution:* We take the residual of the reconstructed secondary vertex position, defined as $v_{\rm rec} - v_{\rm sim}$, and histogram it for different numbers of tracks in the reconstructed vertex. In each track bin, a Gaussian fit is performed, and the width of the fit is taken as the resolution.

- *Pull:* The same method is used for the pulls, which are defined as $\frac{v_{\rm rec}-v_{\rm sim}}{\sqrt{\sigma v_{\rm rec}^2+\sigma v_{\rm sim}^2}}$. For consistency, the fit in each track bin should be centered around zero with a width of 1.

Figures 8.12 and 8.13 show (in regard to the jet axis) the transversal and longitudinal resolution distributions, respectively, and figures 8.14, 8.15 the pull distributions. The resolution, as is to be expected, is larger than for primary vertices due to about one order of magnitude fewer tracks in secondary vertices. For a secondary vertex within a jet, it is also to be expected that the uncertainty in the direction of the jet is larger than in the direction orthogonal to it; this is consistent with the observed difference between transversal and longitudinal resolution.

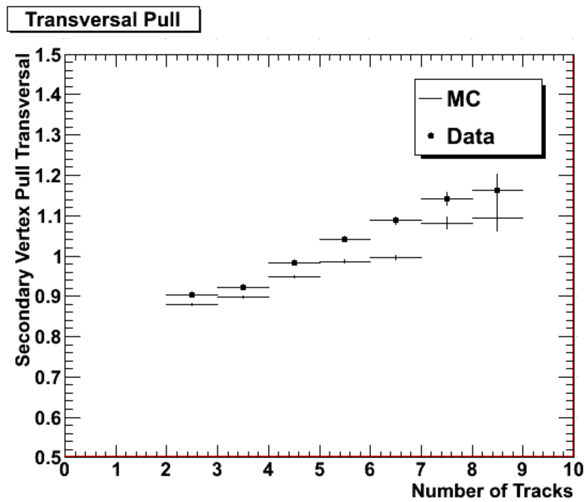


Figure 8.14: Transversal pull.

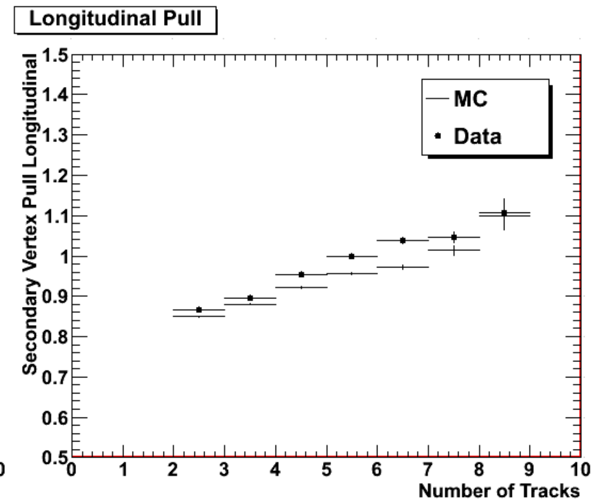


Figure 8.15: Longitudinal pull.

8.3 Comparison to Monte Carlo b decays

Since the algorithm basically models b decays on light quark jets, the results have to be compared to results obtained on Monte Carlo b decays. To be able to compare the efficiency vs. flight distance plots, the "real" flight distance of the MC b decays needs to be correctly identified from the MC generator information.
The following data set has been used for the comparison:

- InclusiveBB_Pt30/Summer10-START36_V9_S09-v1.

Some issues need to be addressed in the difference between the pseudo-vertices and the MC b decays:

- The algorithm generates only secondary vertices; no further decays are modelled. In real b decays, tertiary vertices, such as from c decays, might occur, and since there is no sure-fire way of associating a reconstructed vertex to a simulated decay, this could lead to an underestimate of the flight distance of the reconstructed vertex.

- Gluon splitting (processes such as $g \to b\bar{b}$) can lead to two b vertices being in the same reconstructed jet. The flight distance of one simulated b decay might then be associated to a reconstructed vertex from the other b decay, again resulting in a potential distortion.

The gluon splitting cases could be excluded with a veto, but since they occur in ca. 23% of all MC b events, a better solution for both problems is to take the largest flight distance of all generator particles in the b decay chain(s) as the relevant flight distance; this is sensible

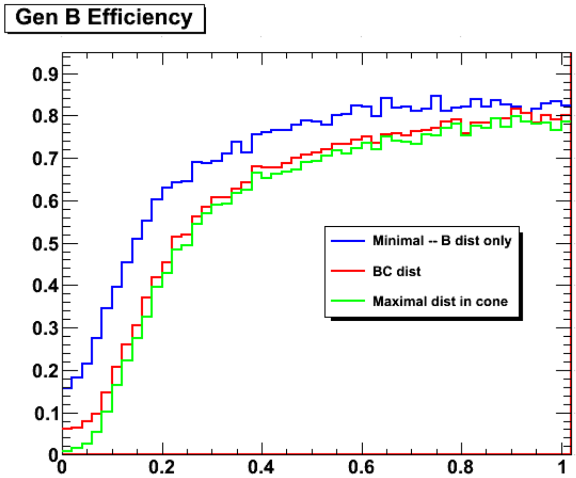


Figure 8.16: Efficiency vs. dist [cm].

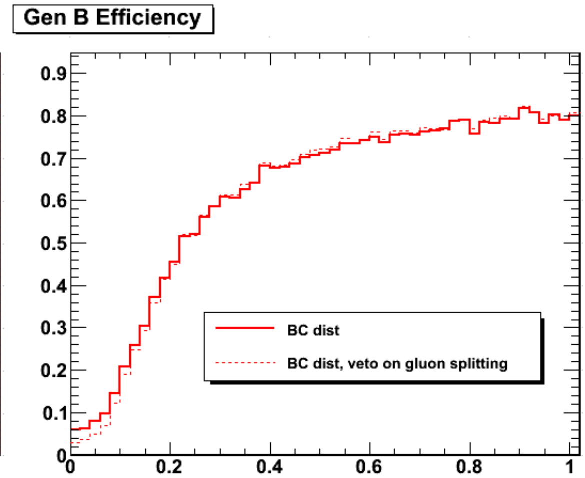


Figure 8.17: Efficiency w. veto on gluon splitting.

since we are only interested in the reconstruction efficiency of the secondary vertex, not the exact physical process it comes from.

Figure 8.17 shows the effect a veto on the gluon splitting cases would have. This veto, however, is not used in the subsequent plots. In Figure 8.16, the reconstruction efficiency is shown depending on three values for the flight distance: the minimal distance only between the interaction point and the b decay; the distance between interaction point and the decay from a charm particle coming from the b decay (if no charm decay is present, this is simply the minimal distance); and the maximal distance between interaction point and all particles originating from the b decay that are still in the jet cone.

8.3.1 Results

Fig. 8.18 shows the pseudo-vertex reconstruction efficiency on Monte Carlo and on data, compared to the secondary vertex reconstruction efficiency on MC b decays; the maximal generator particle decay distance in the jet cone has been taken as the b flight distance. Figures 8.19 through 8.23 show the kinematic variables as defined above.

8.4 Tag Efficiency

The results of this method can be used to calibrate the systematic uncertainty on the reconstruction efficiency for actual b decays. In order to do this, we define a *tagging*

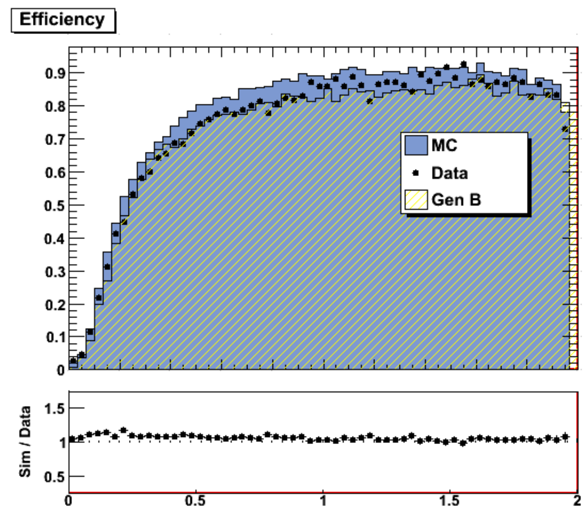


Figure 8.18: Reconstruction efficiency vs. flight distance [cm], compared between pseudo-vertices (on MC and data) and simulated b decays.

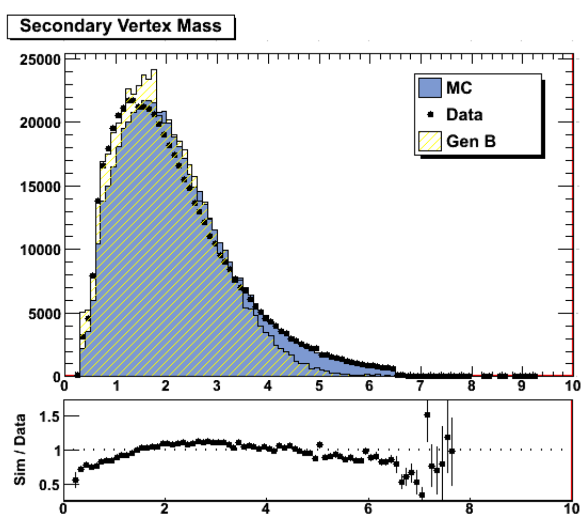


Figure 8.19: Sec. vertex mass [GeV].

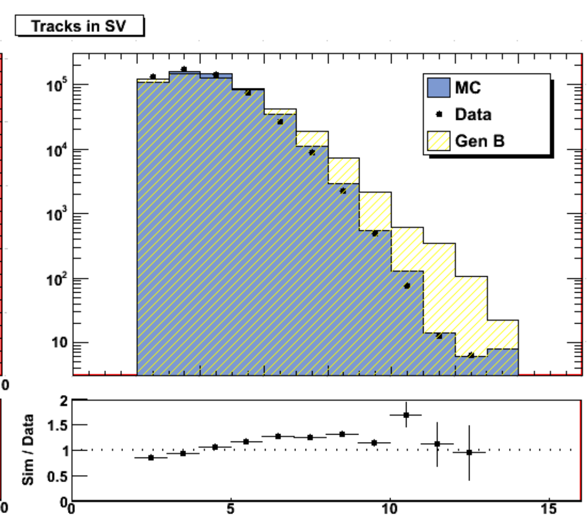


Figure 8.20: Tracks in secondary vertex.

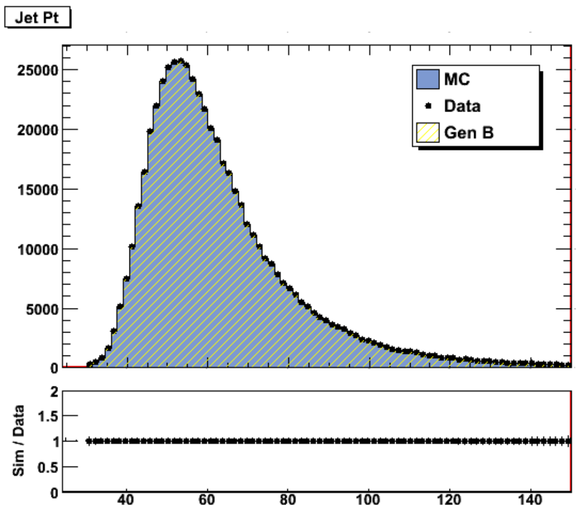


Figure 8.21: Jet p_t [GeV].

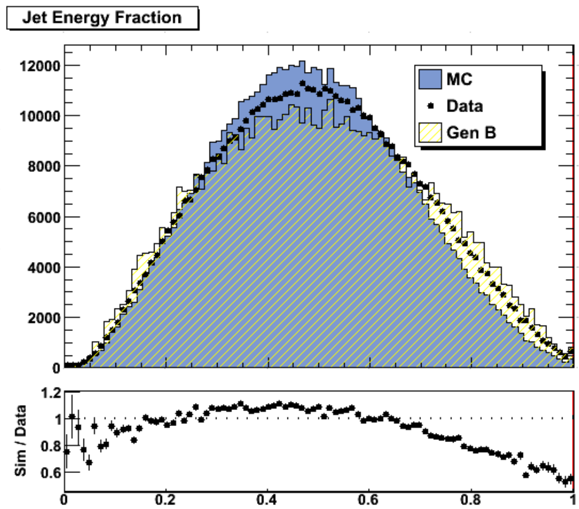


Figure 8.22: SV energy divided by jet energy.

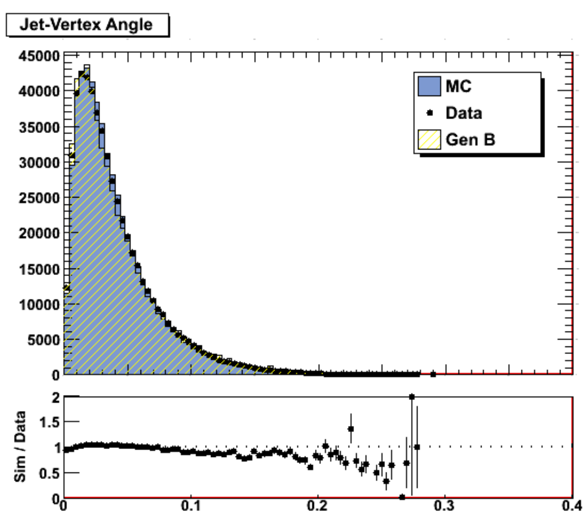


Figure 8.23: Angle bw. jet and vertex axis [rad].

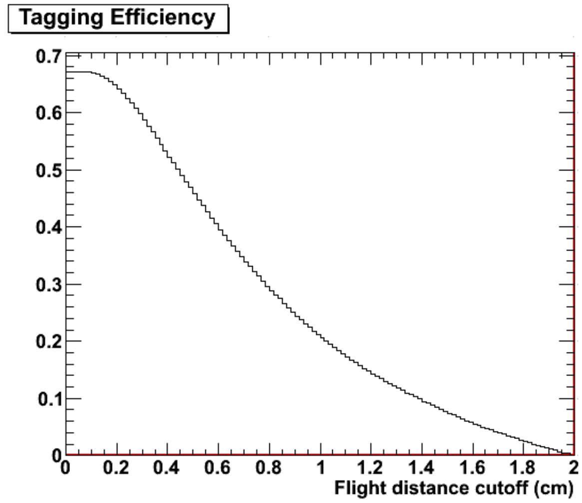


Figure 8.24: Tag efficiency vs. flight distance cut [cm].

efficiency $\varepsilon_{\mathrm{tag}}$: the efficiency vs. flight distance histogram on MC b decays is multiplied with the reconstructed b flight distance histogram, then the resulting histogram $H(f)$ is integrated starting from a flight distance f_i:

$$\varepsilon_{\mathrm{tag}}(f_i) = \int_{f_i}^{\infty} H(f) df. \tag{8.1}$$

Numerically integrating for all values f_i yields a tag efficiency vs. flight distance function (Fig. 8.24). This tells us how many b decays we can expect to find for a given flight distance cut-off. For example, if we take into account only b decays with a flight distance of 4 mm or more, we get 50 % of all b decays.

So far, the pseudo-vertices of the method have not come into play: now, we interpret the distance in efficiency $\Delta\varepsilon$ for pseudo-vertices on Monte Carlo and data in each flight distance bin f_i,

$$\Delta\varepsilon(f_i) = \varepsilon_{\mathrm{MC}}(f_i) - \varepsilon_{\mathrm{data}}(f_i), \tag{8.2}$$

as the systematic uncertainty. This way, we obtain a measure of systematic uncertainty on the tag efficiency for b decays, as seen in Fig. 8.25.

8.5 Kinematic Cross Checks

The dependence of the results on three values is investigated: jet pseudorapidity η, jet transversal momentum p_{t}, and the number of moved tracks.

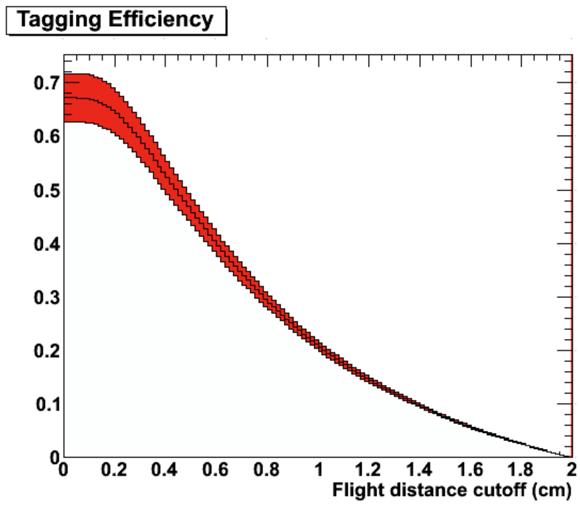


Figure 8.25: Tag efficiency vs. flight distance cut [cm]. The red region around the curve denotes the systematic uncertainty.

8.5.1 Jet η

The relevance of this value is due to the different detector structure in the barrel region ($0 < |\eta| < 1.4$) and the forward region ($|\eta| > 1.4$). Efficiency is expected to be higher in the barrel and lower in the forward region because the track parameter resolution is better in the transversal plane than in the longitudinal plane, and due to the fact that particles with higher η have to pass through more material.
To investigate this effect, the efficiency and kinematic plots have been split in regard to the barrel and forward region. The upper bound for the forward region has been chosen as $|\eta| < 2.17$; this is so all the considered jets are still completely contained in the calorimeter systems, based on the standard jet cone size of 0.5.
Figure 8.26 shows clearly the expected difference in efficiency depending on the detector region. The kinematic plots can be found in Appendix A.

8.5.2 Jet p_{t}

The dependence of the results on the transversal momentum of the jet is investigated; the plots can be found in Appendix B. There is a systematic effect of fewer reconstructed tracks in low-p_{t} jets on data; this results in a lower vertex reconstruction efficiency in these regions. For higher jet p_{t}, pseudo-vertices on Monte Carlo and on data are in good agreement.

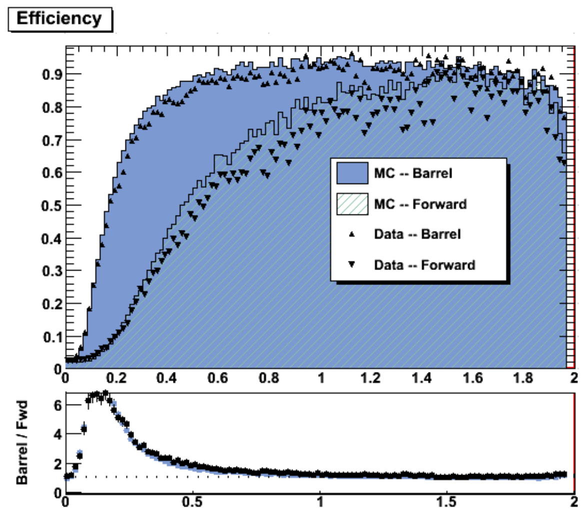


Figure 8.26: Reconstruction efficiency vs. flight distance in barrel and forward region [cm].

8.5.3 Displaced Tracks

The dependence of the results on the number of displaced tracks is investigated. Since this variable reflects the quality of the pseudo-vertex modeling, the results are compared to Monte Carlo b decay vertices, depending on the "real" number of tracks in the b decay from the generator information: this number is defined as the number of reconstructed tracks which are matched to generator particles coming from the b decay and which pass the necessary quality criteria to be considered by the vertex fitting.

The results have been split into plots of 2, 3, 4, 5, and ≥ 6 tracks; the plots can be found in Appendix C. The increase in efficiency with the number of tracks is evident; more tracks equal better vertex reconstruction. The mass plots are in best agreement between 3 and 4 tracks; since we use reconstructed tracks from light-quark jets (which correspond mostly to kaons and pions) to create the pseudo-vertices, it is to be expected that the invariant mass of the pseudo-vertices will be larger for ≥ 4 tracks, since we cannot choose which type of particle (and consequently, which mass) we use for the pseudo-vertex. However, the good agreement of the efficiency plots shows that this modeling effect is ultimately negligible for the reconstruction efficiency we are interested in.

8.6 Conclusion

The previous sections have shown how the pseudo-vertex method can be used to obtain efficiency and resolution values for secondary vertices. Figures 8.19 through 8.23 demonstrate that the kinematic variables of the pseudo-vertices created by the method, both on Monte Carlo and real data, are consistent with the kinematic values of simulated b decays. Figure 8.18 shows a good agreement both between pseudo-vertices on Monte Carlo and data, and between pseudo-vertices and simulated b decay vertices. Consequently, a large part of the uncertainty on secondary vertex-based b-tagging on data can be described well by the method; this allows us to calibrate the tagging efficiency, as seen in Fig. 8.25, with the results obtained with the method.

B-tagging is a crucial ingredient to a large amount of analysis scenarios in CMS. Despite its relative simplicity, the Simple Secondary Vertex algorithm has proven to be a powerful tool in the b-tag framework.

The first physics results using b-tagging on data from the 2010 runs have been published, and the importance of the algorithms in the next run will increase even further. Consequently, a thorough understanding of the b-tagging framework on recent data and a precise calibration of the algorithms are an important basis of future physics results.

Appendix A

Cross Check Plots: Jet η

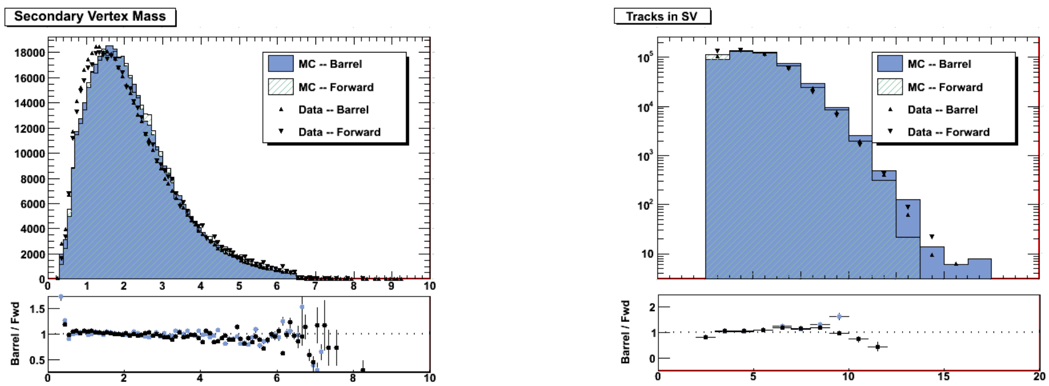


Figure A.1: Left: vertex mass [GeV], right: tracks in vertex.

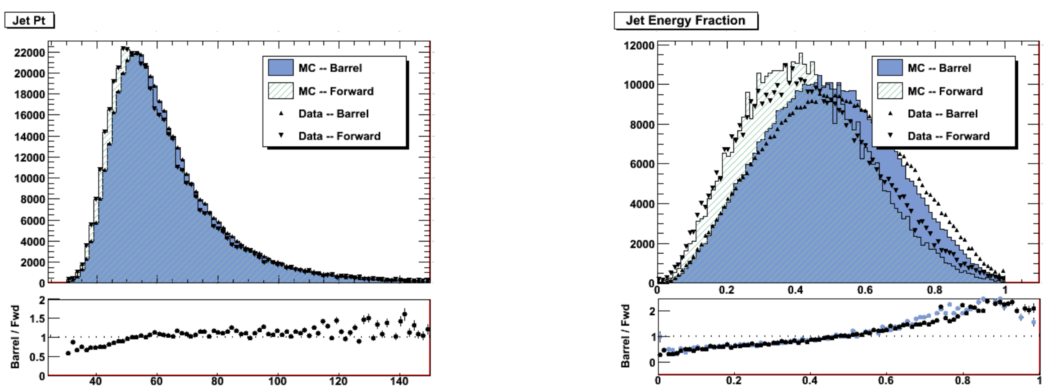


Figure A.2: Left: Jet p_{t} [GeV], right: vertex energy / jet energy.

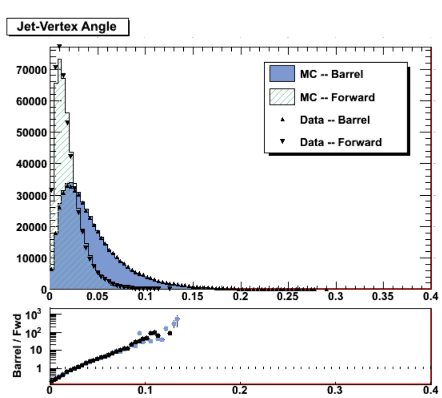


Figure A.3: Angle between vertex and jet axis [rad].

Appendix B

Cross Check Plots: Jet p_{t}

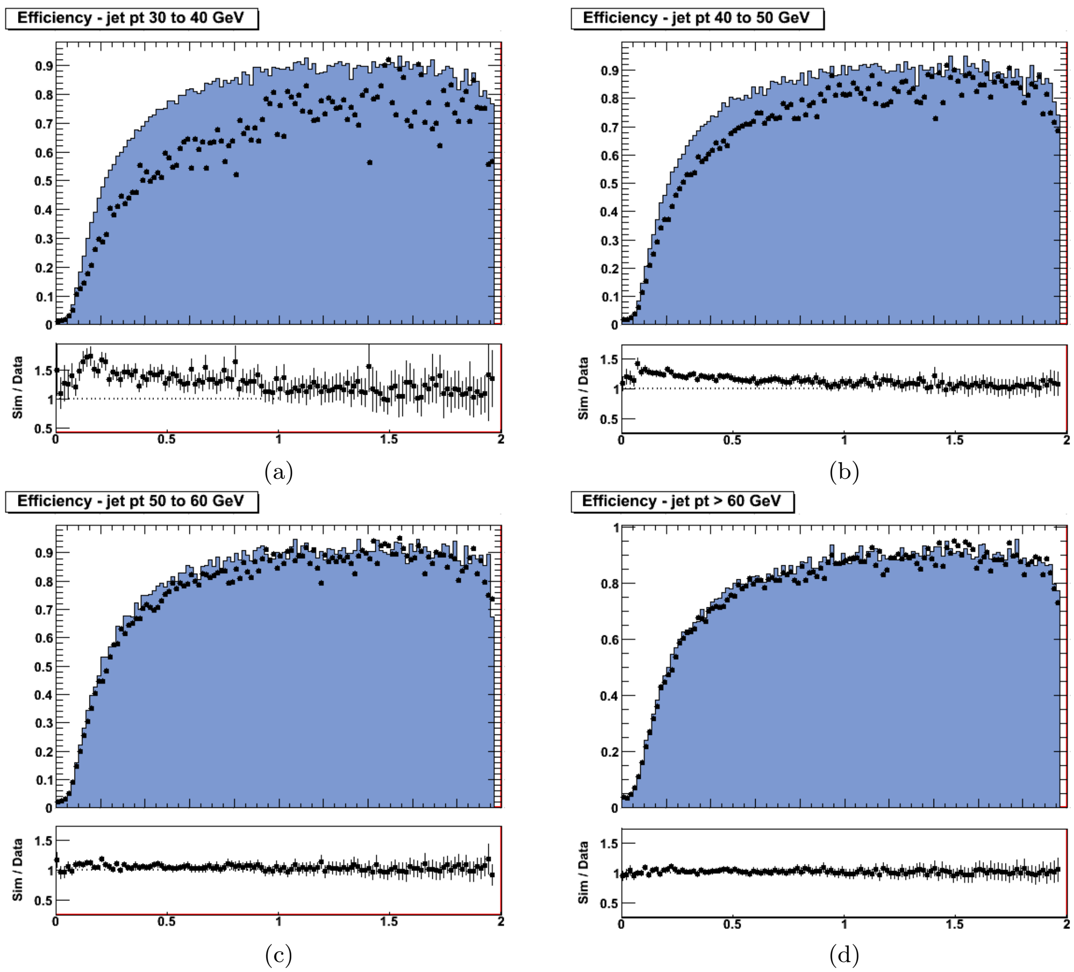


Figure B.1: Reconstruction efficiency vs. flight distance [cm].

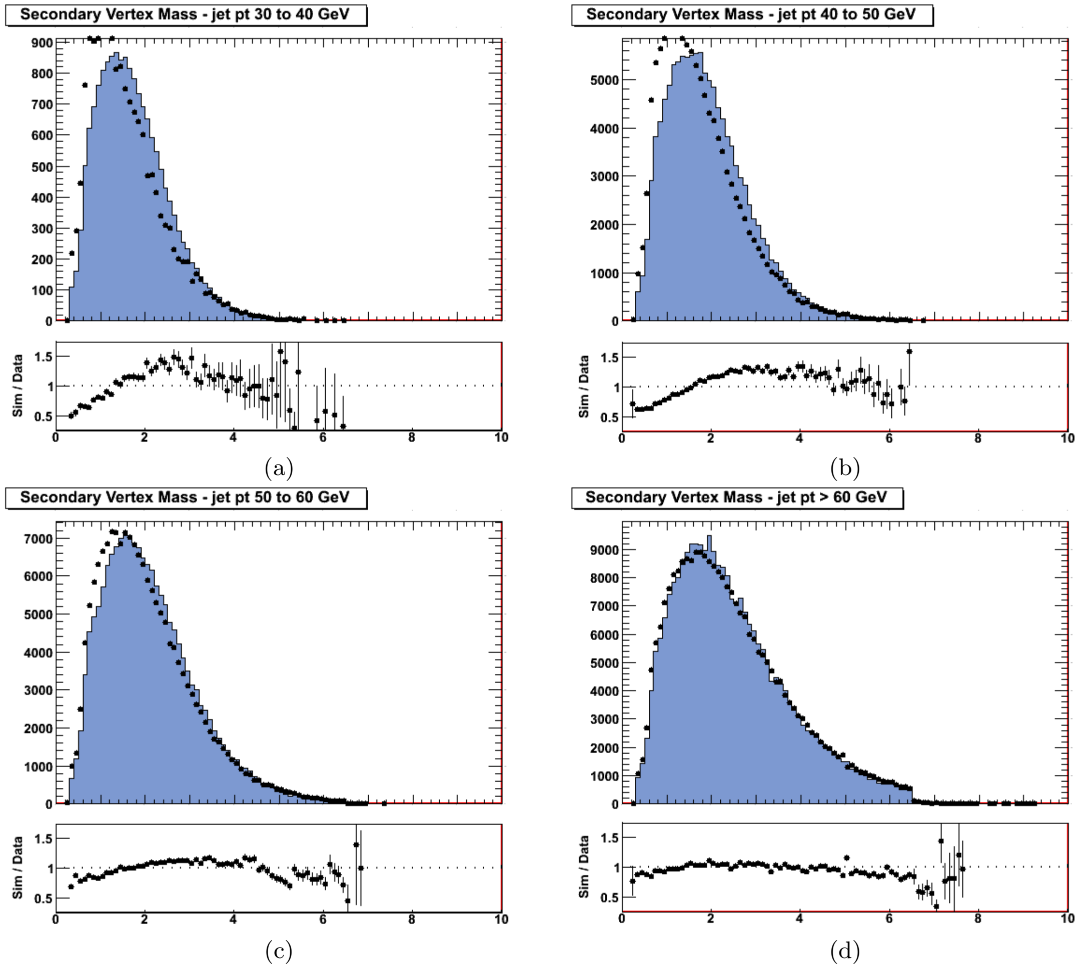


Figure B.2: Secondary vertex mass [GeV].

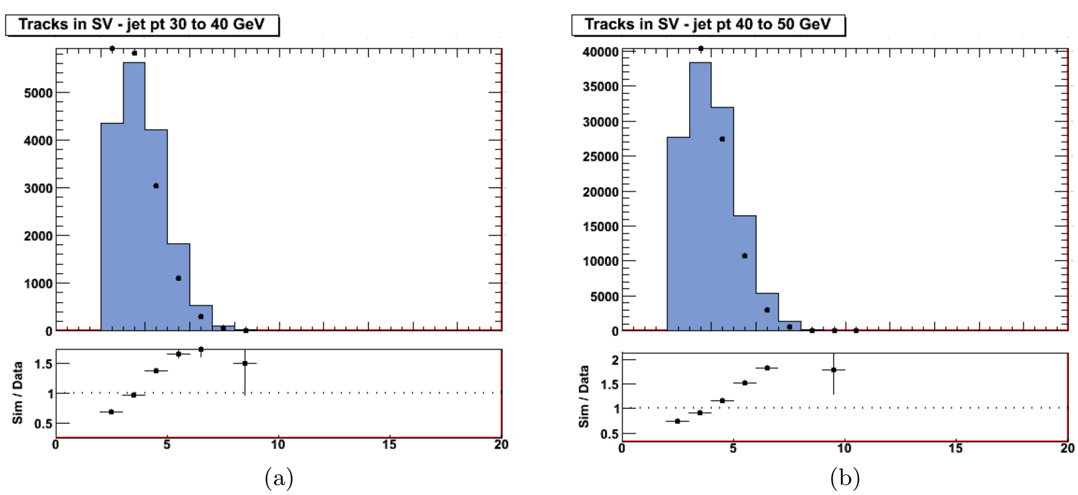
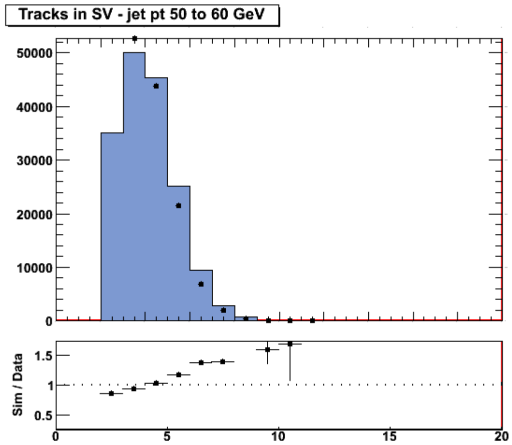
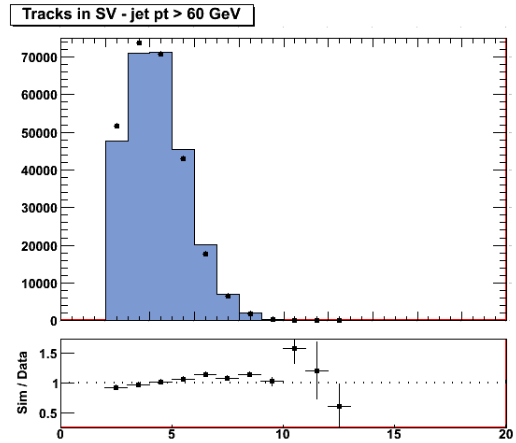


Figure B.3: Number of tracks in secondary vertex.

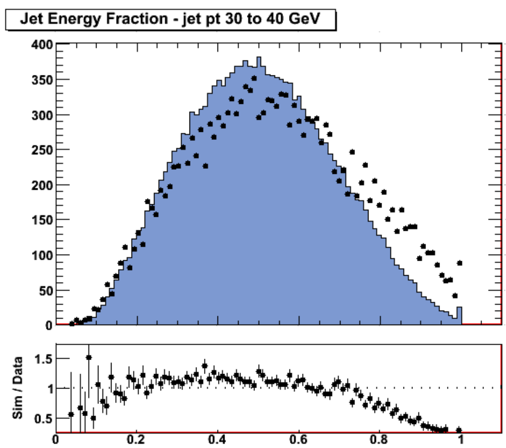


(a)

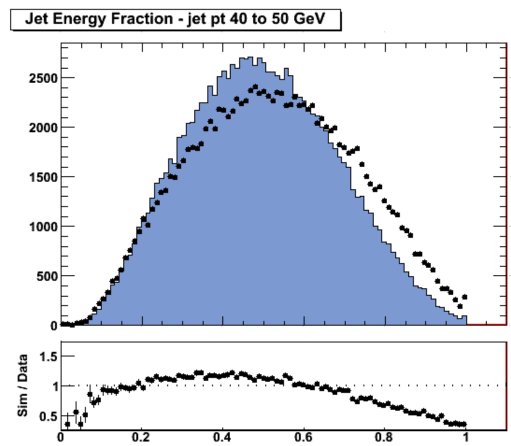


(b)

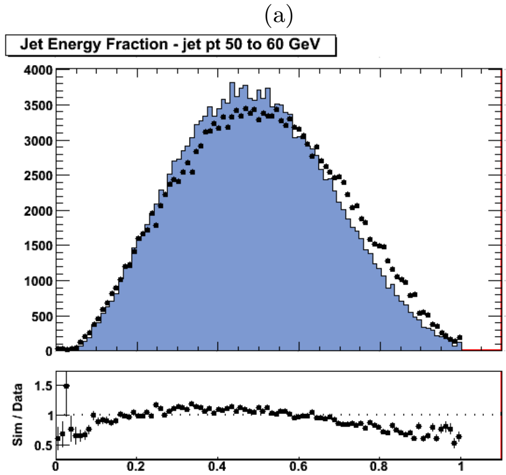
Figure B.4: Number of tracks in secondary vertex.



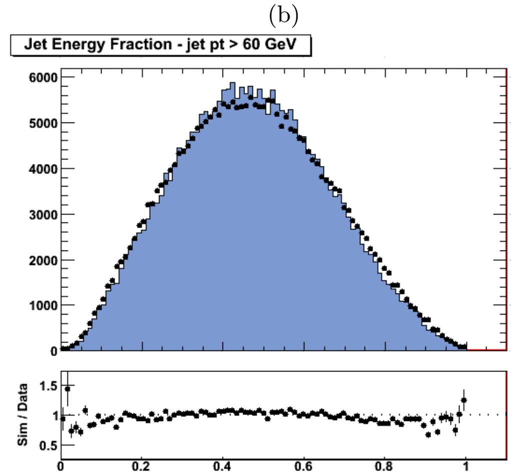
(a)



(b)



(c)



(d)

Figure B.5: SV energy divided by jet energy.

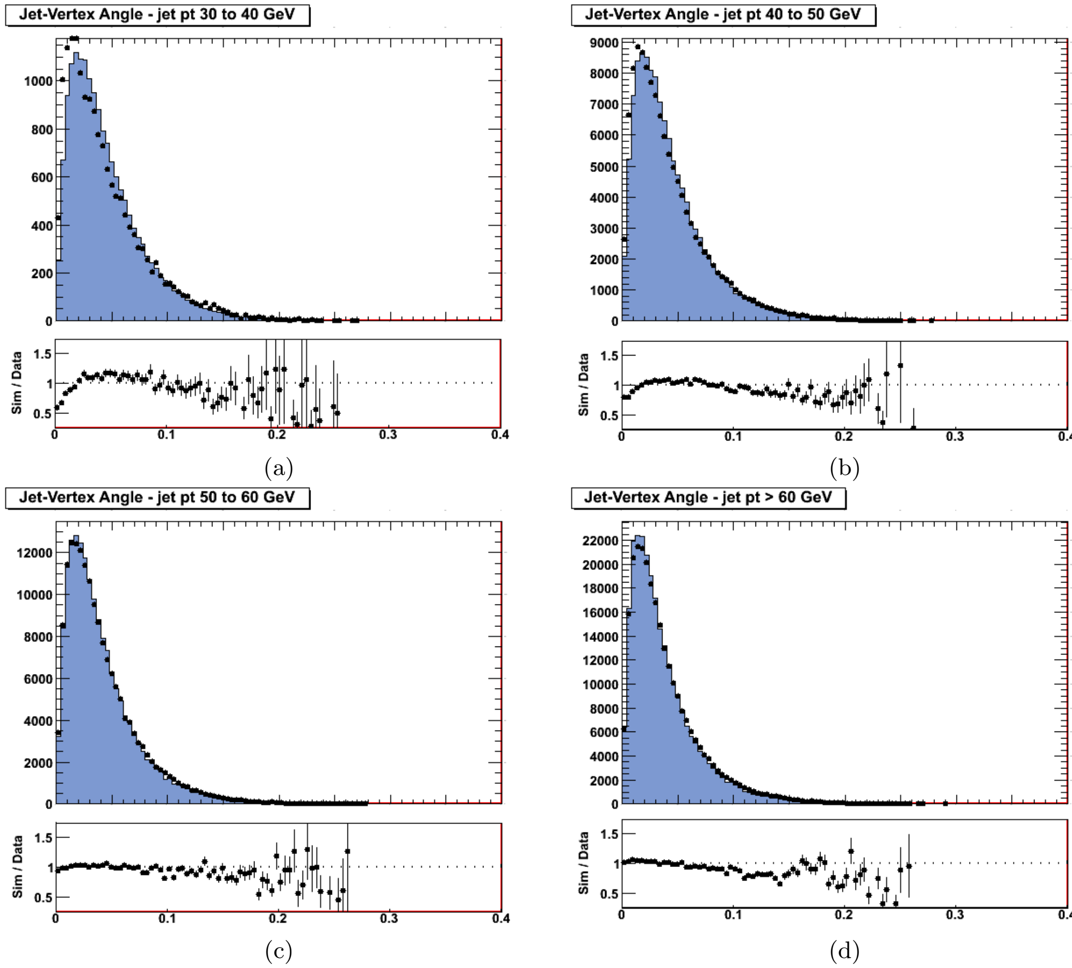


Figure B.6: Angle between vertex and jet axis [rad].

Appendix C

Cross Check Plots: Displaced Tracks

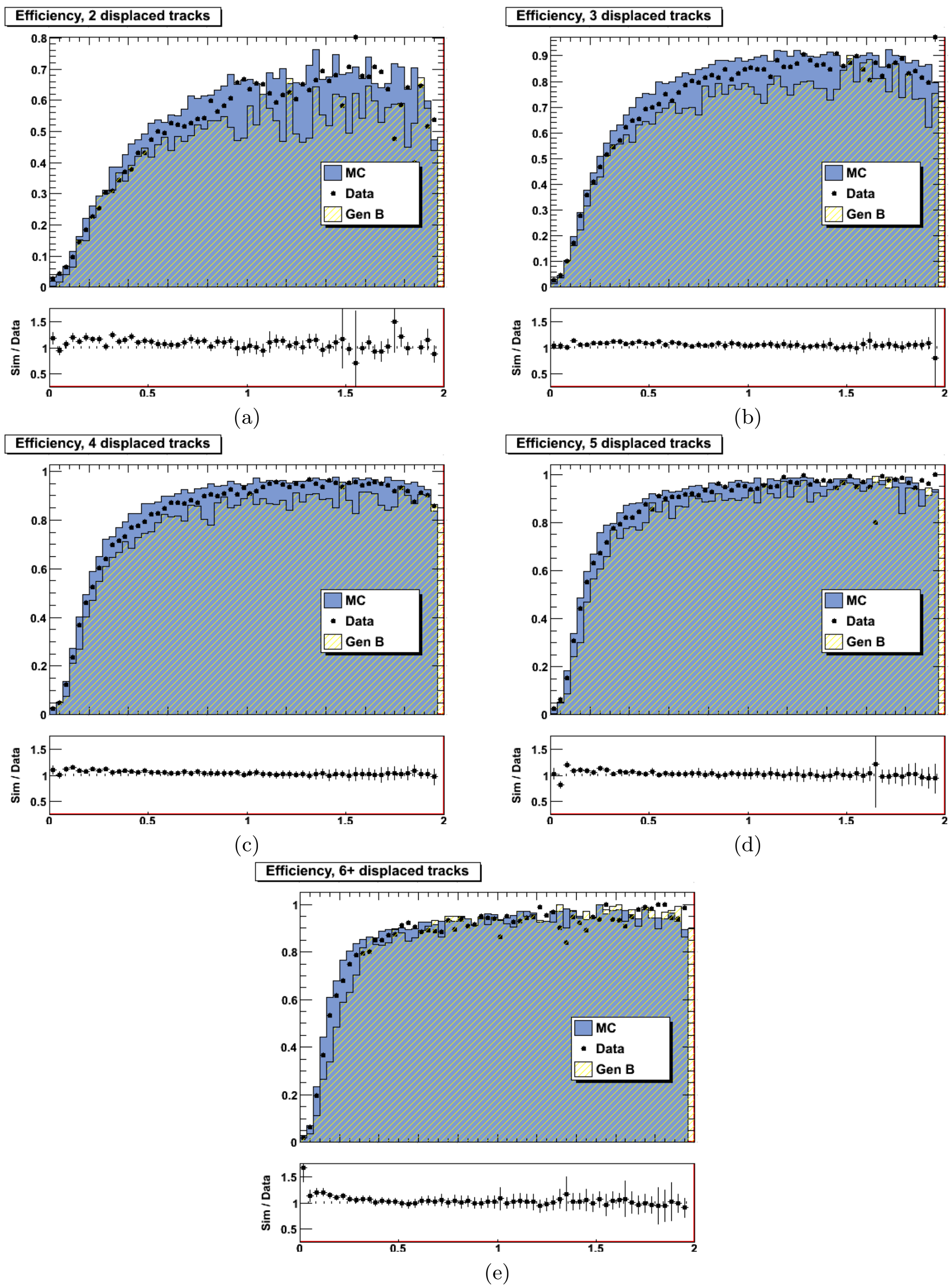
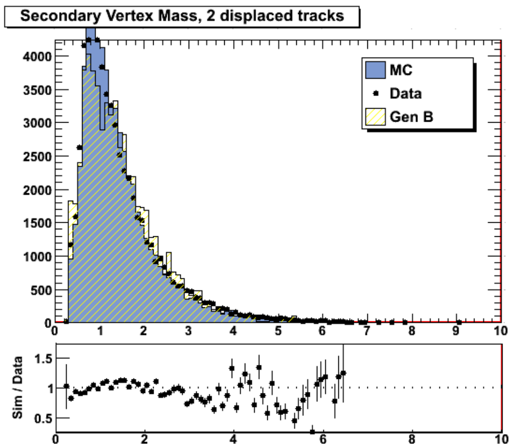
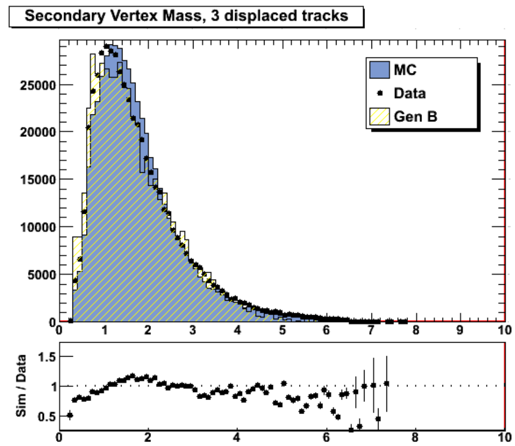


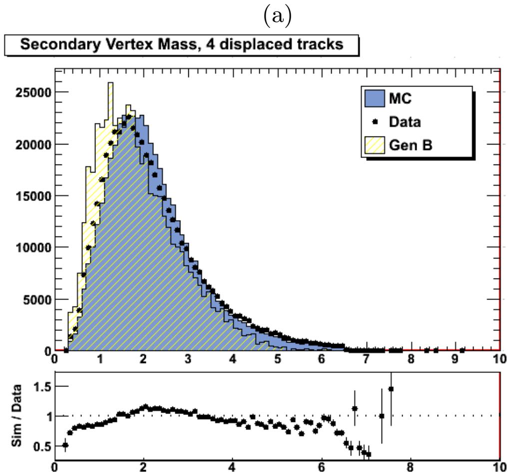
Figure C.1: Reconstruction efficiency vs. flight distance [cm].



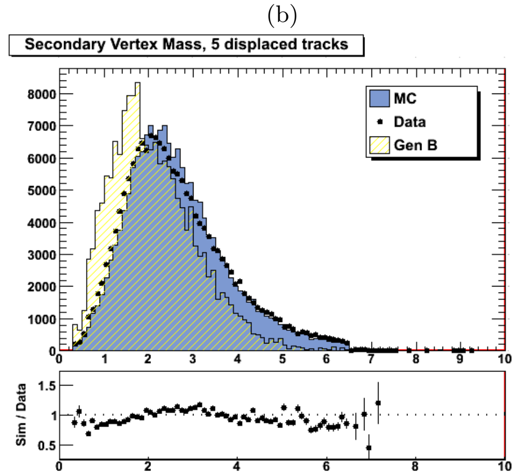
(a)



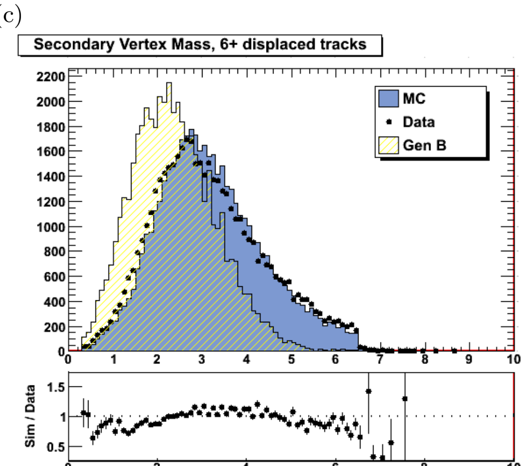
(b)



(c)



(d)



(e)

Figure C.2: Secondary vertex mass [GeV].

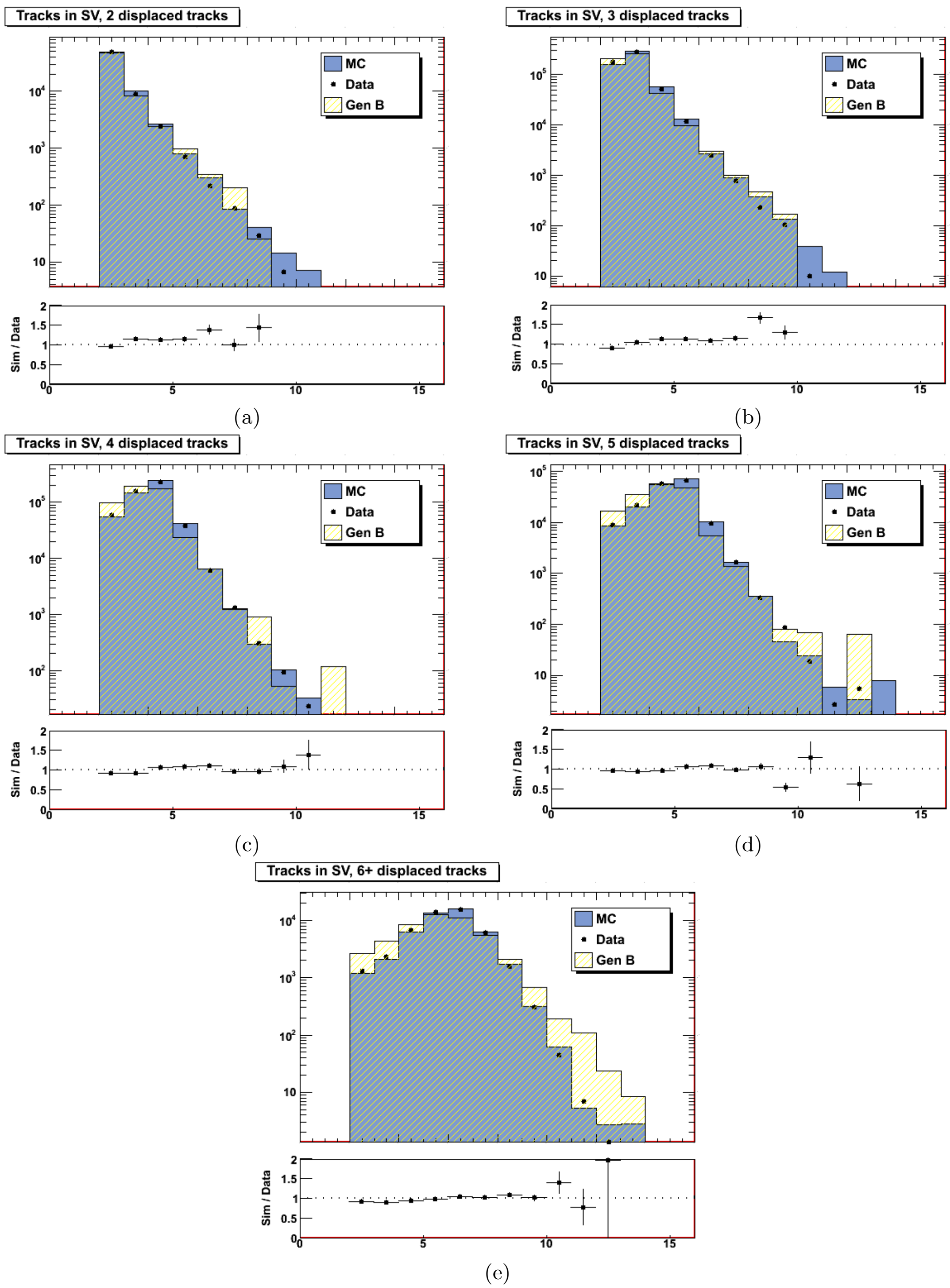
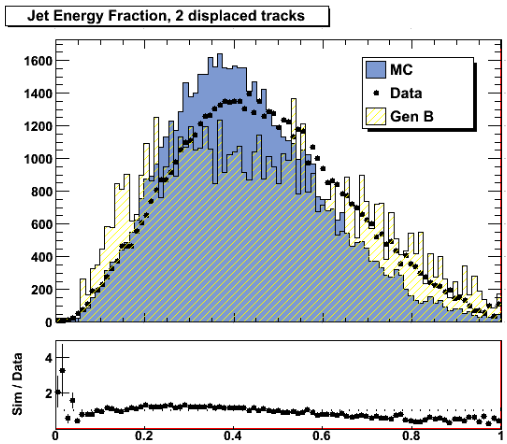
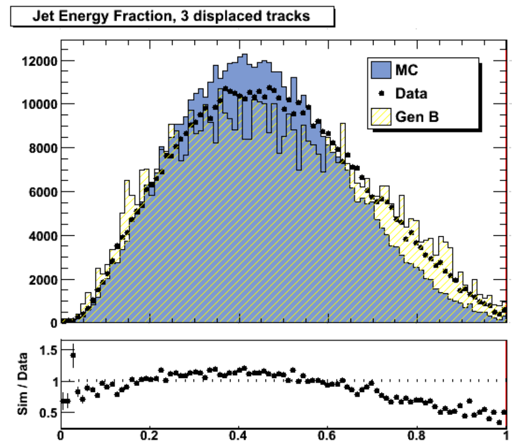


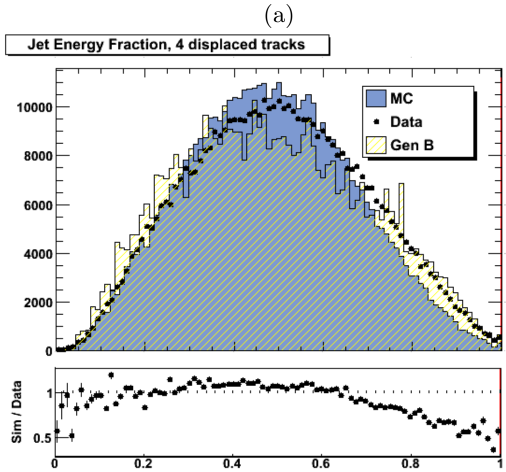
Figure C.3: Number of tracks in secondary vertex.



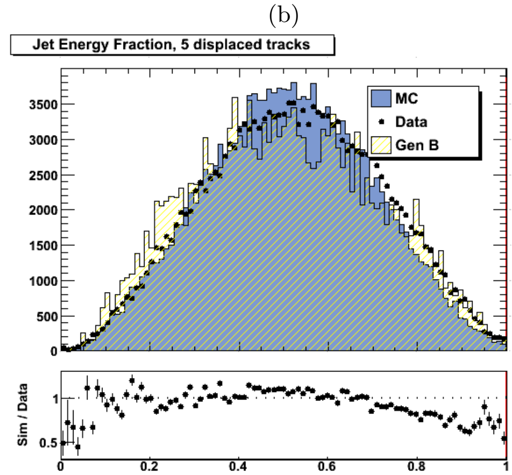
(a)



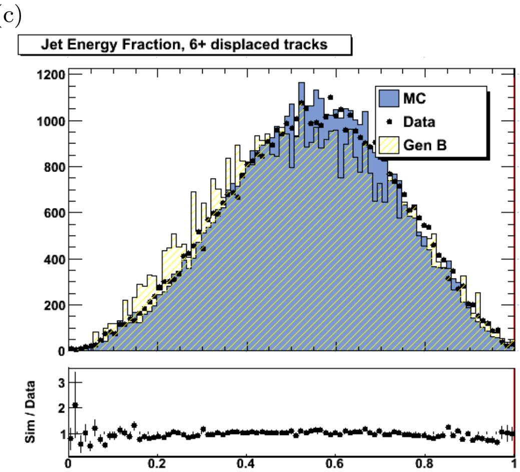
(b)



(c)



(d)



(e)

Figure C.4: SV energy divided by jet energy.

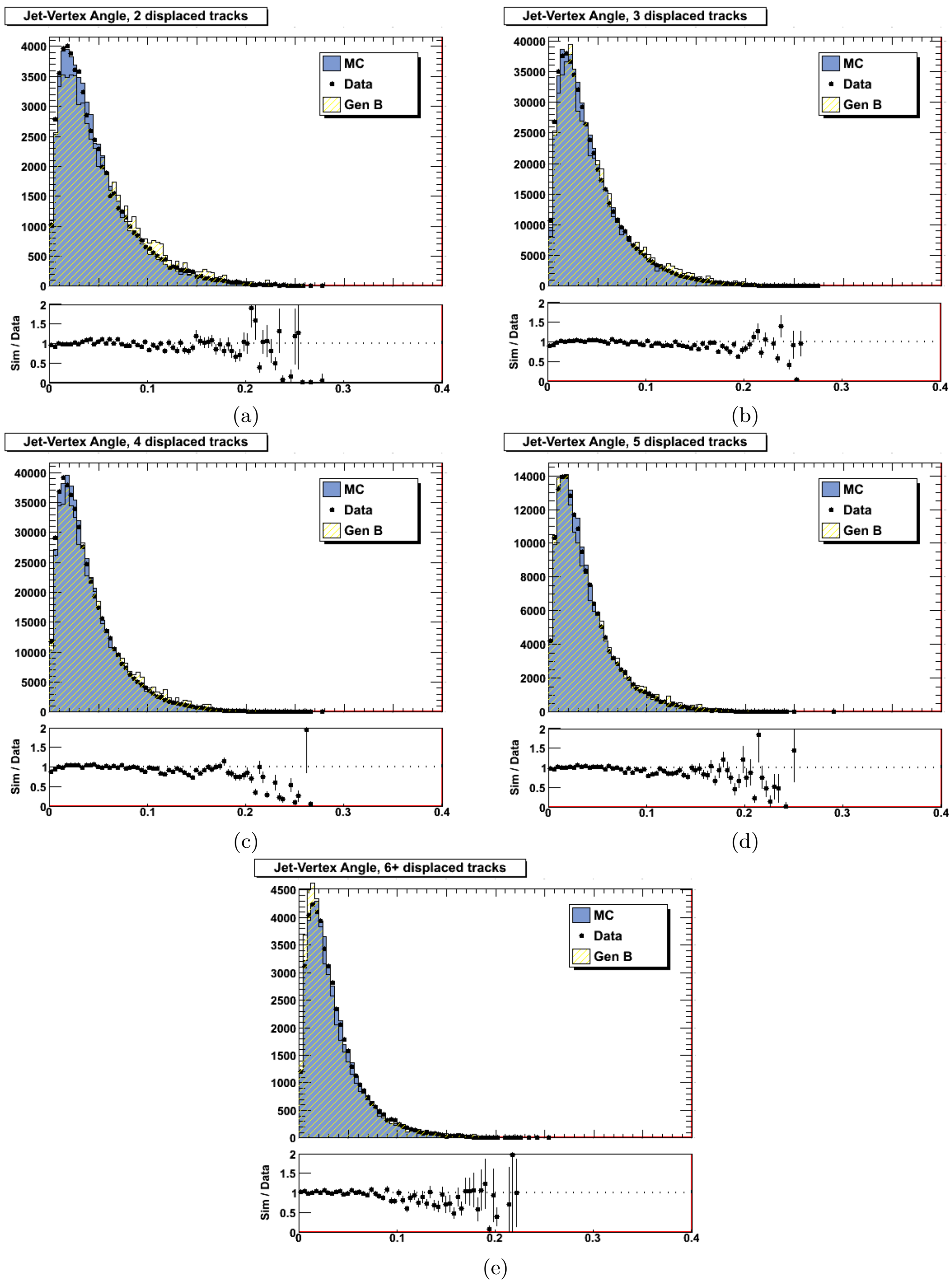


Figure C.5: Angle between vertex and jet axis [rad].

Appendix D

German Abstract / Zusammenfassung

Der Large Hadron Collider (LHC) am CERN ist seit 2010 in kontinuierlichem Betrieb, und die verschiedenen Experimente (darunter CMS) haben seit Beginn der Aufnahme bereits eine große Menge an Daten gesammelt. Die Analyse dieser stets wachsenden Datenmenge ist ein entscheidender Vorgang in der Suche nach neuer Physik und in der Überprüfung des Standardmodells der Elementarteilchenphysik.

Die Identifizierung von Punkten gemeinsamen Ursprungs (*Vertizes*) in einer Menge von rekonstruierten Teilchenspuren ist ein wichtiges Werkzeug in beinahe allen Analyseszenarien am CMS. Die korrekte Identifizierung des Kollisionspunktes der Strahlen (der *Primärvertex*) ist ein essentieller Startpunkt für die weitere Analyse des Ereignisses, und die Identifizierung der Zerfallspunkte (*sekundäre Vertizes*) von in der Kollision erzeugten Teilchen ist ein wichtiger Bestandteil in vielen Analyseszenarien, sowohl in der Untersuchung des Standardmodells als auch in der Suche nach neuer Physik. Von besonderem Interesse sind Zerfälle, die von Hadronen stammen, die Bottom-Quarks beinhalten; diese kommen in vielen der Prozesse vor, auf deren Untersuchung der CMS-Detektor ausgelegt ist.

Diese Diplomarbeit wurde am Institut für Hochenergiephysik (HEPHY) der Österreichischen Akademie der Wissenschaften durchgeführt. Es wird in der Arbeit ein Überblick gegeben über den LHC, den CMS-Detektor und seine verschiedene Teilsysteme, und die Analysealgorithmen im Software-Framework von CMS. Schließlich wird eine neue, daten-basierte Methode zur Messung und Kalibrierung der Effizienz der Vertexrekonstruktion vorgestellt.

Appendix E

Acknowledgements

First, I would like to thank Christian Fabjan and the entire HEPHY for providing me with the opportunity to work in this field during the exciting first period of LHC data taking, and Manfred Krammer in particular for sparking my interest in my first practical course at the institute and for his support of my initial stint at CERN (together with Beatrix Hiesmayr, to whom I also extend my thanks here).

Furthermore, I would like to thank Wolfgang Adam, Rudolf Frühwirth and Dietrich Liko for their guidance and many fruitful discussions during my thesis work. I would also like to extend my thanks to my colleagues in CMS analysis, in particular Wolfgang Kiesenhofer, Robert Schöfbeck, Florian Teischinger and Wolfgang Waltenberger, who were always there for questions and discussions, be they physical, computational or philosophical in nature.

On a personal note I would like to thank Viola Thomay, my mother, who has never failed to support me and even when the going was slow or the direction uncertain, always stood by my decisions. Without her, I would not be where I am today.

Finally, I want to thank the Austrian public, without whose continued interest in, and support of, fundamental research – and their provident wisdom in not quitting CERN when the long years of preparation finally had begun to pay off – neither this work nor the scientific matrix in which it was created would have been possible.

Bibliography

[1] E. Lyndon and P. Bryant, *LHC Machine*, J. Instrum. 3 (2008) S08001.

[2] The CMS Collaboration, *CMS technical design report, Volume II: Physics Performance*, J. Phys. G 34 (2007).

[3] K. Nakamura *et al.* (Particle Data Group), J. Phys. G 37, 075021 (2010).

[4] The CDF, the D0 Collaborations: T. Aaltonen *et al.*, *Combination of Tevatron searches for the standard model Higgs boson in the W+W- decay mode*, Phys. Rev. Lett. 104, 061802 (2010).

[5] The CMS Collaboration, *The CMS experiment at the LHC*, J. Instrum. 3 (2008) S08004.

[6] K. Desch, J. Kalinowski, G. Moortgat-Pick, K. Rolbiecki, W.J. Stirling, *Combined LHC/ILC analysis of a SUSY scenario with heavy sfermions*, CERN-PH-TH/2006-070.

[7] S. P. Martin, *A Supersymmetry Primer*, arXiv:hep-ph/9709356.

[8] A. Schopper, *Flavour Physics and CP Violation at LHC*, Flavour Physics and CP Violation Conference, Vancouver, 2006.

[9] B. Müller, *Heavy Ions – prospects at the LHC*, Czechoslovak Journal of Physics, Vol 55, Part I, B313 (2005).

[10] P. Wesson, *Five-Dimensional Physics: Classical and Quantum Consequences of Kaluza-Klein Cosmology*, World Scientific, 2006.

[11] K. Becker, M. Becker, J. H. Schwarz, *String Theory and M-Theory: A Modern Introduction*, Cambridge University Press, 2006.

[12] K. S. Thorne, in *Magic without Magic: John Archibald Wheeler*, edited by J. Klauder Freeman, San Francisco, 1972.

[13] S. C. Park, *Black Holes at the LHC: Progress since 2002*, arXiv:0809.2571v1

[14] http://cmsinfo.cern.ch/Welcome.html/CMSdocuments/DetectorDrawings/Slice/CMS_Slice.gif

[15] http://cms.web.cern.ch/cms/Resources/Website/Media/Publications/Posters/DetectorPosters/files/Trackers.pdf

[16] M. Krammer, *Detektoren in der Hochenergiephysik*, lecture notes summer 2007.

[17] The CMS Collaboration, *Measurement of Tracking Efficiency*, CMS PAS TRK-10-002.

[18] The CMS Collaboration, *Tracking and Primary Vertex Results in First 7 TeV Collisions*, CMS PAS TRK-10-005.

[19] W. Adam, Th. Speer, B. Mangano, T. Todorov, *Track reconstruction in the CMS tracker*, CMS NOTE-2006/41.

[20] http://cms.web.cern.ch/cms/Resources/Website/Media/Publications/Posters/DetectorPosters/files/ECALncol.pdf

[21] http://cms.web.cern.ch/cms/Resources/Website/Media/Publications/Posters/DetectorPosters/files/HCALncol.pdf

[22] http://cms.web.cern.ch/cms/Resources/Website/Media/Publications/Posters/DetectorPosters/files/MuDetectors.pdf

[23] The CMS Collaboration, *The TriDAS Project – The Level-1 Trigger Technical Design Report*, CERN/LHCC 2000-38.

[24] The CMS Collaboration, *The TriDAS Project – Data Acquisition and High-Level Trigger Technical Design Report*, CERN/LHCC 2002-26.

[25] V.M. Abazov *et al.* D0 Collaboration, *Observation of Single Top-Quark Production*, Phys. Rev. Lett. 103, 092001 (2009).

[26] The CMS Collaboration, *Impact of Tracker Misalignment on the CMS b-Tagging Performance*, CMS PAS BTV-07-003.

[27] The CMS Collaboration, *Algorithms for b Jet Identification in CMS*, CMS PAS BTV-09-001.

[28] The CMS Collaboration, *Commissioning of b-jet identification with pp collisions at* $\sqrt{s} = 7$ TeV, CMS PAS BTV-10-001.

[29] A. Rizzi, F. Palla, G. Segneri, *Track impact parameter based b-tagging with CMS*, CMS Note 2006/019.

[30] Delphi Collaboration, *b-tagging in DELPHI at LEP*, CERN-EP/2002-088.

[31] A. Strandlie, R. Frühwirth, *Track and vertex reconstruction: From classical to adaptive methods*, Review of Modern Physics, Vol. 82, April-June 2010.

[32] R. Frühwirth, *Application of Kalman Filtering to Track and Vertex Fitting*, Nuclear Instruments and Methods in Physics Research A262 (1987) 444-450.

[33] R. Frühwirth, P. Kubinec, W. Mitaroff, M. Regler, *Vertex reconstruction and track bundling at the LEP collider using robust algorithms*, Computer Physics Communications 96 (1996) 189-208.

[34] W. Waltenberger, *Adaptive Vertex Reconstruction*, CMS NOTE-2008/033.

[35] M. Cacciari, G. Salam, G. Soyez, *The anti-k_t jet clustering algorithm*, JHEP 04 (2008) 063.

[36] The CMS Collaboration, *Particle–Flow Event Reconstruction in CMS and Performance for Jets, Taus, and E_T^{miss}*, CMS PAS PFT-09-001.

[37] The CMS Collaboration, *Tracking and Vertexing Results from First Collisions*, CMS PAS TRK-10-001.

[38] T. Sjöstrand, S. Mrenna, P. Skands, *A Brief Introduction to PYTHIA 8.1*, arXiv:0710.3820.

[39] A. Rybkin *et al.*, *GEANT 4 – a simulation toolkit*, Nuclear Instruments and Methods in Physics Research A 506 (2003) 250-303.

Curriculum Vitae

PERSONAL DETAILS

Christian Thomay
Wallgasse 12/16b
A-1060 Wien

christian_thomay@hotmail.com
Mobile: 0650/8607631

Date of Birth: 21.12.1982

EDUCATION

Mar. 2005 – Feb. 2011	Study of Physics, University of Vienna
Jul. 09 – Aug. 09	Summer Student at CERN, Geneva
Nov. 09 – Jan. 11	Diploma thesis at the Institute of High Energy Physics (HEPHY) Vienna
Feb. 2004 – Feb. 2005	Austrian civilian service (Zivildienst)
Mar. 2002 – Jan. 2004	Study of Physics, University of Vienna
Oct. 2001 – Jan. 2002	Study of Molecular Biology, University of Vienna
Sep. 1992 – Jun. 2001	Bundesgymnasium Dornbirn

ADDITIONAL SKILLS

Languages	Fluent in German, English and Hungarian, fair in French
Computer Skills	Windows, Mac OS, MS Office, LaTeX, Mathematica
Programming	C++, Python