Universität Karlsruhe (TH)
Forschungsuniversität • gegründet 1825

Algorithms for the Identification of b-Quark Jets with First Data at CMS

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

DISSERTATION

von

Dipl. Phys. Armin Scheurer
aus Reutlingen

Tag der mündlichen Prüfung: 18.07.2008

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

Korreferent: Prof. Dr. M. Feindt
Institut für Experimentelle Kernphysik

Probleme kann man niemals mit derselben
Denkweise lösen, durch die sie entstanden sind.

Albert Einstein, $*$ 1879 $-$ † 1955

Contents

1 Introduction **9**

2 Theoretical Bases **13**
2.1 The Standard Model of Particle Physics . . . 13
2.1.1 Fundamental Bosons . . . 14
2.1.2 Fundamental Fermions . . . 15
2.1.3 Hadrons: Mesons and Baryons . . . 16
2.1.4 b-Quark Jets at the LHC . . . 17

3 The Experiment **25**
3.1 The Large Hadron Collider . . . 25
3.2 The LHC Experiments . . . 31
3.3 The Compact Muon Solenoid . . . 32
3.3.1 Coordinate Conventions . . . 32
3.3.2 Detector Requirements and Overall Layout . . . 33
3.3.3 Solenoid Magnet . . . 34
3.3.4 Silicon Tracking System . . . 35
3.3.5 Electromagnetic Calorimeter . . . 37
3.3.6 Hadronic Calorimeter . . . 38
3.3.7 Muon Tracking System . . . 39
3.3.8 Data Acquisition and Trigger System . . . 39
3.4 The CMS Software Framework . . . 41
3.4.1 Software Components . . . 42
3.4.2 Event Generation and Detector Simulation . . . 43
3.4.3 Reconstruction of Physics Objects . . . 45

4 The CMS Computing Model **49**
4.1 The Worldwide LHC Computing Grid . . . 49
4.1.1 WLCG Tier Structure . . . 50
4.1.2 The Grid Middleware . . . 51
4.1.3 WLCG Services . . . 52
4.2 Operation of the Tier1 Centre GridKa . . . 55
4.2.1 Computing, Software and Analysis Challenge 2006 . . . 55

4.2.2 Lessons Learned and Current Status 63

5 b-Quark Jet Identification and Alignment Studies 67
5.1 b-Quark Jet Identification 68
5.2 b-Tagging at CMS 70
5.2.1 Available Algorithms 70
5.2.2 Jet, Track and Vertex Reconstruction 74
5.2.3 Input Variables 77
5.2.4 Algorithm Performance 80
5.3 Alignment Studies 84
5.3.1 Misalignment Scenarios 84
5.3.2 Technical Implementation 86
5.3.3 Refitting Versus Re-Reconstruction of Tracks 86
5.3.4 Impact on the Input Variables 89
5.3.5 Expected Algorithm Performance 94
5.4 Secondary Vertex Only Based Algorithm 98
5.4.1 SimpleSecondaryVertex Candidates 101
5.4.2 Secondary Vertex Reconstruction Optimisations 106
5.4.3 Secondary Vertex Finder Variations 109
5.4.4 Alignment Position Error Variations 109
5.4.5 Association Cone Size Variations 112
5.4.6 Performance Comparison 113

6 Conclusion and Outlook 117

A Configuration Files 121
A.1 Sample CMSSW Configuration File 121
A.2 Configuration of a b-Tagging and Misalignment Job 124

B Used Data Sample 133
B.1 Inclusive $\mathbf{t\bar{t}}$ - Events 133

C Additional Figures 135
C.1 b-Tagging Input Variables 135
C.2 b-Tagging Discriminator Distributions 140
C.3 b-Tagging Performance Distributions 142
C.4 Track Refit Versus Track Re-Reconstruction 145
C.5 Impact of Misalignment on the Input Variables 148
C.6 Impact of Misalignment on the Algorithm Performance 150
C.7 SimpleSecondaryVertex Candidates 153
C.8 Flavour Efficiencies Versus Discriminator Cut 157
C.9 SV Optimisation Studies 159
C.10 Variations of the Alignment Position Error 163

C.11 Variations of the Track Association Cone Size 168
C.12 Performance of All Lifetime Based b-Tagging Algorithms 170

D CMS Tracking Efficiencies **173**
D.1 Track Reconstruction Efficiencies and Resolutions 173

E Old Versus New Software Framework **177**
E.1 Results From the Physics TDR 177

List of Figures **185**

List of Tables **187**

List of Abbreviations **192**

Bibliography **197**

Chapter 1

Introduction

The Standard Model of particle physics is a compendium of several different quantum field theories following the laws of quantum mechanics and special relativity. In a very precise way, it describes all known particles together with three out of four fundamental forces: the strong, the weak and the electromagnetic force. During the last decades, various experiments have verified the predictions of the Standard Model with incredible precision up to energies of the order of 1 TeV, strengthening our comprehension of the universe. Although the lack of the gravitational force prevents the Standard Model from being a complete theory of fundamental interactions it nevertheless became an effective theory due to the negligible strength of the gravitation with respect to the other interactions at the small scales at which particle physics experiments are taking place.

Originally, the Standard Model predicted massless bosons as mediating particles for the electromagnetic, strong and weak force. However, the limited strength of the latter one was a sign on massive vector bosons which was confirmed at CERN in 1983 through the discovery of heavy W and Z bosons. The simplest approach in order to explain massive gauge bosons was the introduction of a scalar field which is assumed to be responsible for creating the masses through spontaneous symmetry-breaking of the electroweak gauge symmetry. The mediator of this field, the Higgs boson, is the only particle predicted by the Standard Model that has not been observed yet. Its discovery at former and present particle accelerators was quite unlikely due to its relatively high mass, low production probability and inconsiderable decay modes.

The next consequent step on a way to fully confirm the Standard Model is the design and construction of a more powerful machine in order to enable the investigation of a larger energy scale and to detect and analyse yet unknown or theoretical expected particles. After a long phase of development, the Large Hadron Collider (LHC) is currently being assembled in the former LEP tunnel at CERN. Its completion is expected for summer 2008 and proton-proton collisions at an energy of 14 TeV will take place after the initial phase at lower energies is over. The LHC will exceed the world's most energetic particle

accelerator, the TEVATRON, by a factor of seven in energy and more than two orders of magnitude in collision rate.

The LHC design leads to about 40 million interactions per second and thus an enormous amount of data has to be handled by each of the six participating experiments. Since they will produce roughly 15 Petabytes of data per year, a new concept in terms of storage resources and computing power was developed. In order to ensure a proper processing and analysis of the collected data by thousands of scientists all around the globe, the LHC Computing Grid (LCG) was born. It consists of so-called computing centres ordered in tiers which are hierarchically linked. The Tier-0 centre at CERN is followed by eight Tier-1 centres equally distributed among the participating countries and several Tier-2 and Tier-3 centres attached to them.

Another demanding part of the LHC experiments is a proper reconstruction and identification of physically interesting collision events out of a huge amount of background. Among other things, various physics analyses including top quark physics, the search for the Higgs boson or for physics beyond the Standard Model, rely on the presence of so-called particle jets originating from the hadronisation of bottom quarks (b-tagging) in their decay chain. Thus, the identification of such jets is a crucial tool in order to reach many physics goals presented by the LHC. Most of the used b-tagging algorithms exploit the fact that b-hadrons have a significant lifetime and therefore create a secondary decay vertex displaced from the primary interaction point of the event. The charged particles emerging from the b-hadron decay provide a measurable distance between their linearised track and the primary vertex. The achievable performance of the b-tagging algorithms therefore strongly depends on the efficiency and precision of the silicon tracking system and a good secondary vertex reconstruction. In return, alignment positions and orientations of the tracking detector modules directly affect the performance of the tracking device and thus have a large influence on the b-tagging algorithm behaviour. Although the knowledge of the alignment is increased during the time of data taking, a good understanding of the algorithm behaviour with respect to misalignment is desirable in order to use these methods reliably even with first data.

This work is related to the Compact Muon Solenoid (CMS) experiment which is one of the two large general-purpose detectors situated at the LHC main ring. The studies presented in this thesis can be divided into two parts: part one deals with the operation and maintenance of one of the WLCG Tier-1 centres, GridKa, which is situated at Forschungszentrum Karlsruhe, whereas part two is related to the identification of b-quark jets at CMS. The performance of all b-tagging algorithms is described and their behaviour concerning several degrees of alignment precision of the silicon tracking device is studied. Furthermore, the development of a b-tagging algorithm which is almost insensitive with respect to tracker alignment is presented here.

This thesis is structured as follows: Chapter 2 gives a short introduction to the Standard Model and summarises some important LHC physics searches with b-jets in the final state.

Chapter 3 describes the Large Hadron Collider, the Compact Muon Solenoid experiment and parts of the CMS offline software framework. Chapter 4 focuses on the LHC Computing Grid and presents the efforts on operating and maintaining one of its major Tier-1 centres, GridKa. In chapter 5, the available CMS b-tagging algorithms are introduced and their behaviour under certain aspects of a misaligned tracking detector is studied. The chapter presents also the development and implementation of an alignment insensitive b-tagging algorithm at CMS. Chapter 6 finally gives a conclusion of the presented work and depicts an outlook on possible future studies.

Chapter 2

Theoretical Bases

Several theoretical concepts described nature and especially observations in the field of particle physics quite well during the past few centuries. In the beginning, an idea of small and indivisible particles, called atoms, became one of these models which had to be adjusted in the same manner as the experiments trying to prove these theories evolved during time. Due to higher probing energies, new phenomena arose and modifications had to be made to the theories in order to reflect these new observations correctly. Today, a collection of different quantum field theories called the Standard Model (SM) of particle physics is widely used and precisely describes three out of four known fundamental forces and the participating particles. The lack of gravity prevents the Standard Model from being a complete theory of fundamental interactions, but since the strength of the gravitational force is negligible with respect to microscopic phenomena arising in particle collisions it can be disregarded for our purpose. This chapter gives a short overview of the Standard Model and describes the necessity for a precise identification of so-called particle jets in order to verify and measure several SM parameters, get evidence for the Higgs boson or even indications for new physics beyond the Standard Model. More details can be found in [1] and [2].

2.1 The Standard Model of Particle Physics

The knowledge of our environment and all observed processes in nature can be reduced to four fundamental forces: gravitational, electromagnetic, weak and strong force. The Standard Model combines the latter three underlying models as well as all known elementary particles and forms a quantum field theory which is consistent with quantum mechanics and special relativity. The mathematical framework of the SM is based on group symmetries with the gauge groups $SU(3) \otimes SU(2) \otimes U(1)$. The gauge group $U(1)$ corresponds to quantum electrodynamics (QED) and provides one generator for the represented gauge field. The $SU(2)$ and $SU(3)$ groups are a bit more complicated and belong

to the weak interaction (quantum flavourdynamics, QFD) and to the strong interaction (quantum chromodynamics, QCD) and provide three and eight generators for the quantum fields, respectively. Table 2.1 gives an overview of these four fundamental forces ordered by their strength, relative to each other[1].

Table 2.1: Fundamental forces ordered by decreasing strength [1]

Funamental Force	Strength	Underlying Theory	Mediating Particles
Strong	10	Quantum chromodynamics	8 Gluons
Electromagnetic	10^{-2}	Quantum electrodynamics	Photon (γ)
Weak	10^{-13}	Quantum flavourdynamics	$W^{\pm}$ and Z^0
Gravitational	10^{-42}	Quantum geometrodynamics	Graviton

There are only two types of particles forming the world we live in. Following the laws of quantum mechanics, only particles with whole- or half-numbered spin quantum numbers J are allowed.

2.1.1 Fundamental Bosons

Particles with an integer spin, $J = n \cdot \hbar$ ($n \in \mathbb{N}_0$) are called bosons. Among them are four fundamental bosons which act as mediating particles for the weak, strong and electromagnetic force. They are called gauge bosons and are mathematically described as field equations for massless particles in the first place. All bosons are indistinguishable particles which follow the Bose-Einstein statistics and therefore can populate the same quantum states at the same time. The gauge bosons which act as force carriers in case of the electromagnetic interaction are called photons, γ. The weak and strong interactions provide three and eight mediating particles according to their field generators, respectively. The three heavy vector bosons $W^{\pm}$ and Z^0 correspond to the weak interaction, whereas eight gluons g act as mediators for the strong force. Table 2.2 gives an overview of these field quanta and shows some properties of these elementary particles.

[1]The absolute strength of the different interactions should not be taken too literally, since it depends on various aspects like the nature of the source or the distance to it. It nevertheless gives an impression on the strength of the forces relative to each other.

Table 2.2: Fundamental bosons which act as field quanta for the weak, strong and electromagnetic interaction and several particle properties [3] like their mass m, electric charge q, spin parity J^P and third component of the weak isospin T_3

Bosons	mass m [GeV/c^2]	el. charge q	J^P [$\hbar$]	T_3 [$\hbar$]
Photon (γ)	0	0	1^-	0
$W^\pm$	80.40	$\pm e$	1^+	± 1
Z^0	91.19	0	1^+	0
Gluon (g)	0	0	1^-	± 1

2.1.2 Fundamental Fermions

On the other hand, particles with spin $J = (n + \frac{1}{2}) \cdot \hbar$ are called fermions. They follow the Pauli exclusion principle, thus rely on the Fermi-Dirac statistics. The fundamental fermions are divided into two subgroups, namely quarks and leptons and are summarised in table 2.3.

There are 12 fundamental fermions, 6 quarks and 6 leptons as well as the same number of anti-particles associated to them which have the same mass but opposite quantum numbers. The quarks and leptons occur pair-wise and therefore can be grouped into three generations (also called families), where particles within one family have similar properties.

Table 2.3: The fundamental fermions are divided into quarks and leptons. Each of them consists of three generations with similar properties, e.g. electric charge q, spin J and third component of the weak isospin T_3

Fermions	Generation 1	2	3	el. charge q [e]	J [$\hbar$]	T_3 [$\hbar$]
Quarks	u	c	t	$+2/3$	$1/2$	$+1/2$
	d	s	b	$-1/3$	$1/2$	$-1/2$
Leptons	ν_e	ν_μ	ν_τ	0	$1/2$	$+1/2$
	e	μ	τ	-1	$1/2$	$-1/2$

The electron e, muon μ and tauon τ together with their three corresponding neutrinos ν_e, ν_μ and ν_τ represent the three known lepton flavours. According to their generation,

each flavour has a characteristic quantum number Q_e, Q_μ and Q_τ assigned to the different leptons. Due to differences in the mass and energy eigenstates of the neutrinos, oscillations occur between the different families which lead to non-conserved lepton numbers. However, the sum of the lepton numbers is conserved by all interactions.

The six quarks, up, down, charm, strange, top and bottom can also be grouped into pairs and carry a charge of either $+2/3$ or $-1/3$, respectively. Corresponding to their names, each quark flavour has an assigned quantum number U, D, C, S, T and B which are each conserved by all known interactions, except the weak force. Responsible for possible weak interactions which violate the quark flavour quantum numbers are the differences between the mass eigenstates $|d\rangle$, $|s\rangle$, $|b\rangle$ and the eigenstates $|d'\rangle$, $|s'\rangle$, $|b'\rangle$ of the weak interaction. This issue is described by the Cabibbo-Kobayashi-Maskawa (CKM) matrix which is defined as

$$\begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \cdot \begin{pmatrix} |d\rangle \\ |s\rangle \\ |b\rangle \end{pmatrix} = \begin{pmatrix} |d'\rangle \\ |s'\rangle \\ |b'\rangle \end{pmatrix} \tag{2.1}$$

and connects both representations. The matrix elements V_{ij} represent the weak coupling constants for transitions from up-type quark flavours i to down-type flavours j under emission of a W^+-boson, normalised to the square root of the Fermi coupling constant, $\sqrt{G_\mathrm{F}}$. The diagonal elements have much higher values [3] and thus, transitions within the same families are much more privileged. However, the non-zero values of the remaining elements reflect the above mentioned violation of the quark flavour quantum number conservation.

2.1.3 Hadrons: Mesons and Baryons

Hadrons are strongly interacting particles which are composed of quarks. Their constituents, together with gluons and sea quarks (virtual quark-anti-quark pairs) are called partons. There are two types of hadrons, namely baryons and mesons which are distinguishable by their so-called baryon quantum number, B. Compounds of three quarks carry a baryon number of 1 and are simply called baryons whereas compounds of one quark and one anti-quark are denoted as mesons with a baryon number of 0. There are some theories which predict compounds of more than three quarks, e.g. so-called penta-quarks, which are not discussed here.

Since mesons consist of an even number of elementary fermions they act as bosons on particle level, whereas baryons act as fermions. The discovery of the Δ^{++} resonance, which is a baryon composed of three up quarks in the same state, required the introduction of a further quantum number, otherwise the Pauli exclusion principle would have been violated. This quantum number is called colour charge with the three possible values red, green and blue and their corresponding anti-colours. Each observed particle provides a neutral

colour charge which can only be achieved by e.g. three particles of different colours or a particle and an anti-particle with colour and anti-colour, respectively. The fact that no isolated quarks are allowed is denoted as confinement of colour charged particles. Gluons, as mediators of the strong interaction, are colour charged particles as well.

2.1.4 b-Quark Jets at the LHC

The LHC, a hadron collider which is introduced in the next chapter, collides two high energetic proton beams at several distinct interaction points. Due to the partonic substructure of the protons, a hard proton-proton interaction results in the separation of the constituent partons. As a result of the increased strong interaction energy due to the confinement, new quark-anti-quark pairs are produced and subsequently combined into hadrons during a process called fragmentation (or hadronistation). The charged hadrons and their decay products can then be measured in a detector in form of collimated streams of charged particle tracks along the flight directions of the initial partons, called particle jets.

The majority of collision events at the LHC contains a large number of particle jets in the final states. Figure 2.1 presents various production cross sections, calculated with next-to-leading order precision, for hard scattering processes with respect to the centre-of-mass energy in proton-(anti-)proton collisions. It can be seen that the LHC production cross section for jets with transverse energies greater than 100 GeV is increased by more than two orders of magnitude with respect to the production at the TEVATRON[2]. In addition, the production of b-hadrons is as well increased by about a factor of 10.

The identification of such particle jets is an important task in order to compete for possible innovative discoveries at the LHC. Light flavour (u-, d- and s-hadrons) and gluon jets are almost not distinguishable, since they provide similar event topologies in the final state. However, the identification of b-jets due to particular properties of the weakly decaying b-hadron is feasible.

One possibility is to exploit the non-negligible lifetime of b-hadrons and the fact that tracks of associated decay particles are therefore not compatible with the primary vertex of the event. The displacement of such tracks is measured as the impact parameter, which denotes the distance between the point of closest approach of a charged particle track and the primary vertex. Even though b-quark jets can be identified by algorithms which exploit these lifetime-based parameters, these methods are ambiguous as several processes exist which also produce displaced decay vertices and therefore tracks with measurable impact parameters inside a track detector, for example:

[2] Proton-anti-proton collider with a centre-of-masse energy of 1.96 TeV. The TEVATRON is situated at the Fermi National Accelerator Laboratory near Batavia/Illinois, USA

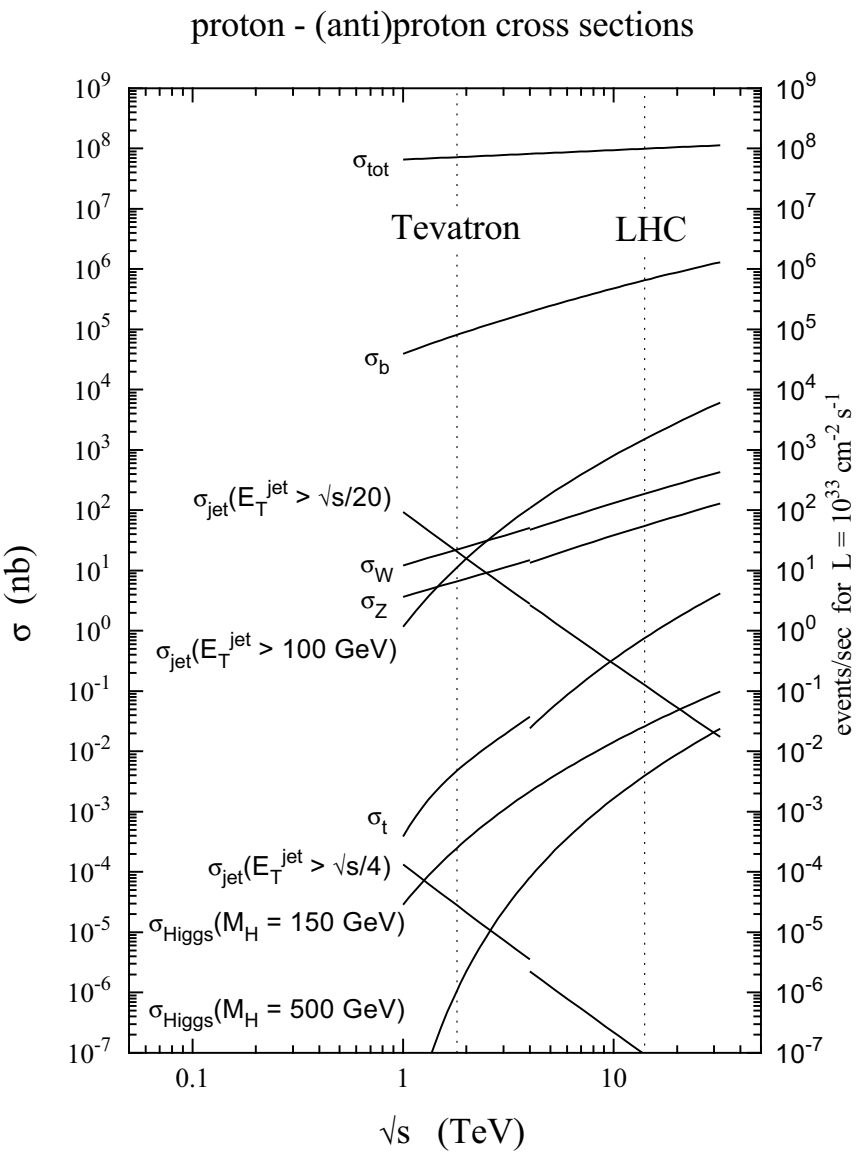


Figure 2.1: Cross sections (based on next-to-leading order calculations) for hard scattering processes with respect to the centre-of-mass energy in proton-(anti-)proton collisions. The dashed lines on the left and right indicate the design centre-of-mass energy of the TEVATRON and the LHC, respectively [4]

- c-hadrons (like b-hadrons) have a distinct lifetime and therefore produce similar event topologies in the final state.

- Some neutral hadrons containing strange quarks, like the K_s^0-meson, form a displaced two-body decay, called V0, inside the track detector.

- Mismeasurements of tracks from light flavour or gluon jets can lead to measurable impact parameters, too.

Another method to identify b-jets uses the fact that some b-hadrons (as well as some c-hadrons) decay into a lepton and its corresponding neutrino. The identification of electrons or muons within a jet indicates a b-jet. However, the overall usability of such an algorithm is limited by the fraction of b-hadrons decaying leptonically. Since the main topic of this thesis is a study of available b-quark jet identification algorithms at the CMS experiment and their behaviour during the first data taking phase of the LHC, more details about this issue are presented in chapter 5.

In the following subsections, a short description of some interesting physics processes with b-jets in the final state are described to emphasise the importance of a proper identification of b-quark jets in order to compete for possible innovative discoveries at the LHC. Besides the interesting field of B-hadron physics, many physics channels with decays of top quarks, Higgs bosons or supersymmetric particles produce b-jets in the final state.

Bottom Physics

An important field of particle physics is related to the precise measurement of the properties of B-hadrons. CP violating processes in B-hadron decays were first measured during the last view years. Oscillations between neutral B-mesons due to differences in the mass eigenstates of the particle and its anti-particle are of prime importance for many projected analyses at the LHC. The measurement of the oscillation frequencies of transitions between $B_0 - \bar{B}_0$ or $B_s - \bar{B}_s$ helps to constrain and measure CKM coupling constants.

Especially the measurement of the high B_s oscillation frequency was performed recently by the D0 and CDF-II experiment at the TEVATRON. Therefore, improved measurements due to the increased production cross section at the high centre-of-mass energy and thus lower statistical fluctuations are important topics at the LHC. Observed variances in these measurements or the identification of other violating decays may indicate on processes of new physics beyond the Standard Model. The Large Hadron Collider beauty (LHCb) experiment is dedicated to B-physics at the LHC. However, important discoveries could be verified as well by searches performed at the CMS or ATLAS experiment.

Top Physics

The top quark has by far the largest mass among all fermions described by the Standard Model and it interacts primarily by the strong interaction, but can only decay via weak processes. These are indications to a broken electroweak symmetry and therefore, the investigation of its properties is important in various current particle physics analyses. Top quarks are mainly produced as $t\bar{t}$-pairs, where gluon-gluon fusion and quark-anti-quark annihilation are the dominant production processes. The lifetime of the top quark is extremely short, and therefore no hadrons are formed. As a result, all spin information of the top quark are directly transferred to its decay particles whose energy-angular distributions can be calculated as purely partonic processes.

The Standard Model top quark decays almost exclusively to a W-boson and a b-quark. The different decay processes are therefore only classified by the subsequent decay of the W into either a lepton and and its corresponding neutrino or into one quark and another anti-quark. Figure 2.2 shows the Feynman diagrams of a leptonically and hadronically decaying top quark. Therefore, the fully hadronic decay of a $t\bar{t}$-pair results in six jets in the final state of the event, four light flavour or c-jets and two b-jets.

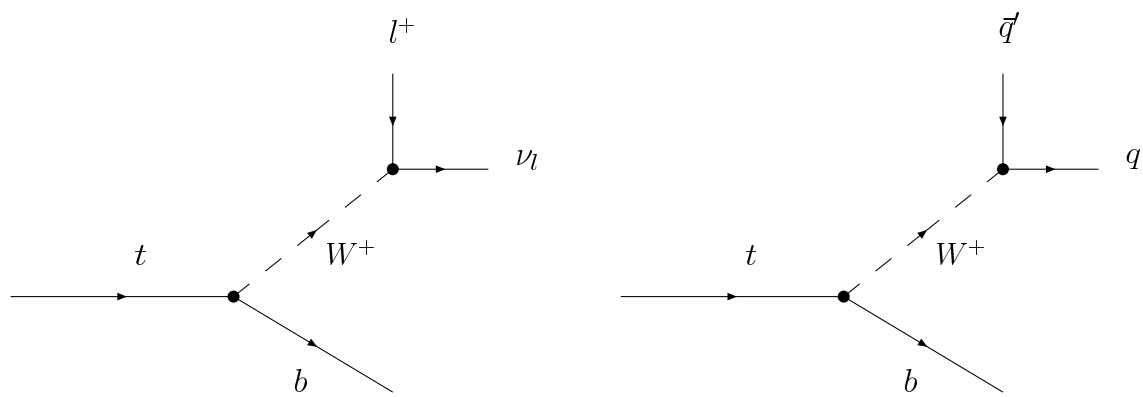


Figure 2.2: Feynman diagram of a leptonically (left) and hadronically (right) decaying top into a lepton and neutrino, l^+ and ν_l, and a quark and an anti-quark, respectively

High precision measurements with top quarks will lead to a deeper understanding of symmetry breaking physics and the origin of the flavour structure. Important topics at the LHC are for example:

- Measurements of the production cross section of $t\bar{t}$-pairs.
- Measurements of the top quark mass in di-leptonic, semi-leptonic and fully hadronic top quark decays as well as by identified J/Ψ resonance in the b-decay.

- Measurements of the decay width of the top quark and CKM coupling constants may indicate new physics.
- Measurements of spin correlations in the top quark pair production.
- More precise measurements of parameters related to the recently discovered single top production mechanism due to the exchange of a virtual W boson.
- Search for flavour changing neutral currents in top decays may lead to indications on physics beyond the Standard Model.

Searches for the Higgs Boson

The Standard Model was not able to explain the generation of mass of fundamental particles. It predicts massless mediating particles for the fundamental forces, but due to the discovery of the massive vector-bosons, Z^0 and $W^{\pm}$, a modification of the theory has to be made. The Higgs mechanism, postulated in 1964, was introduced as an extension to the Standard Model. A further quantum field, called the Higgs field, could explain the generation of mass by coupling to the fundamental particles as well as to itself by spontaneously breaking the electroweak gauge symmetry. The mediating particle of this field is the Higgs boson which is the only Standard Model particle that has not been observed yet.

The only unknown parameter in this model is the mass, m_H, of the Higgs boson itself. Today, only a lower exclusion limit on this mass of 114.4 GeV/c^2 at a confidence level of 95% [3] is given by the former LEP experiments. However, several precision measurements indicate an upper bound of 144 GeV/c^2 at a confidence level of 95%. Figure 2.3 shows the production cross sections as well as the decay branching fractions for different processes for the Standard Model Higgs boson with respect to its mass.

The main decay channel of a light SM Higgs boson below 135 GeV/c^2 is the decay to a $b\bar{b}$-pair. However, the LHC QCD background of this channel with b-dijet events is huge and makes a discovery almost impossible.

The Minimal Supersymmetric Standard Model (MSSM) extends the Standard Model to a supersymmetric theory. After electroweak symmetry breaking, five scalar Higgs mass eigenstates remain in this model: a CP-odd neutral scalar boson A, two charged scalar bosons $H^{\pm}$ and two CP-even neutral scalar bosons h and H. For high values of the introduced MSSM parameter $\tan\beta$, the two heaviest neutral Higgs bosons, A and H, can be produced in association with two b-quarks. A possible decay of these two Higgs bosons into a $b\bar{b}$-pair may result in four b-jets in the final state. This process is the main production channel of a MSSM Higgs boson at the LHC and is discussed in [6]. Since the Higgs boson couples strongly to the top quark, it also may decay into a $t\bar{t}$-pair and these subsequently into a b- and $\bar{b}$-quark in addition to the two W bosons.

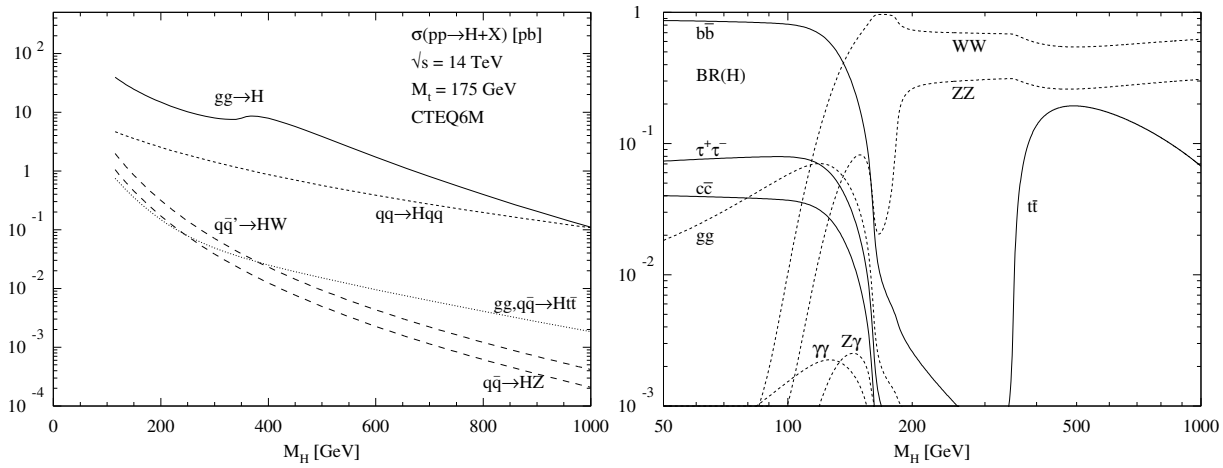


Figure 2.3: Production cross sections (left) and decay branching fractions (right) for a Standard Model Higgs boson depending on its mass, m_H [5]

Searches Beyond the Standard Model

The most promising model for physics beyond the Standard Model is the introduction of supersymmetry (SUSY) where each particle has a so-called superpartner which differs in spin by at least $\hbar/2$. However, up to now, no direct evidence for supersymmetry could be found, but this theory would help to solve two fundamental open questions of the Standard Model: the hierarchy problem of different couplings or masses in theory and experimental measurements and the unification of the strong, weak and electromagnetic force.

The already mentioned MSSM is one of the best studied candidates for supersymmetry. The superpartners associated to quarks or leptons are called squarks, $\tilde{q}$, and sleptons, $\tilde{l}$. In case of gluons, for example, the superpartners are Majorana particles called gluinos, $\tilde{g}$. All supersymmetric decay chains contain a neutral and only weakly interacting, but massive particle, $\tilde{N}_1^0$. Therefore, measurable signatures of SUSY events in a detector are characterised by a fair amount of missing transverse energy, $E_{\mathrm{T}}^{\mathrm{miss}}$, in the final state. The following list gives some typical event signatures of processes with decaying squark and gluino pairs:

- $\tilde{q}\tilde{\bar{q}} \rightarrow q\bar{q}\tilde{N}_1^0\tilde{N}_1^0$, squark-anti-squark $\rightarrow$ 2 jets and $E_{\mathrm{T}}^{\mathrm{miss}}$
- $\tilde{q}\tilde{\bar{q}} \rightarrow q\bar{q}l\bar{l}\tilde{N}_1^0\tilde{N}_1^0$, squark-anti-squark $\rightarrow$ 2 jets, 2 leptons and $E_{\mathrm{T}}^{\mathrm{miss}}$
- $\tilde{g}\tilde{g} \rightarrow q\bar{q}q\bar{q}\tilde{N}_1^0\tilde{N}_1^0$, gluino-gluino $\rightarrow$ 4 jets and $E_{\mathrm{T}}^{\mathrm{miss}}$

In SUSY, there are various decay channels with jets in the final state. Since the identification of particular jet flavours is only possible in case of b- and c-jets, the identification of

b-quark jets will be an important tool for providing additional information in this SUSY sector where several supersymmetric decay modes contain b-jets in the final state.

Another topic for possible discoveries at the LHC are SUSY decays with top quarks and therefore subsequent b-jets in the decay chain, which can be identified. In addition to the MSSM, there are much more topics studied at the LHC which deal with physics beyond the Standard Model. Among them are searches for extra dimensions, the Z' boson, searches based on the mSUGRA model or on Technicolour. More information can be obtained from [5].

Chapter 3

The Experiment

The main focus of this chapter lies on the description of the Compact Muon Solenoid (CMS) experiment located at one of the interaction points of the Large Hadron Collider (LHC) currently built-up in the former LEP[1] tunnel at CERN[2] near Geneva, Switzerland. It introduces the Large Hadron Collider as well as its experiments, where the focus lies on CMS and its most important detector components and offline software.

3.1 The Large Hadron Collider

The Standard Model of particle physics is experimentally well confirmed and considered to be an effective theory up to an energy scale of $\Lambda = 1$ TeV. A limiting factor for possible discoveries of new physics beyond the SM is the achieved collision energy at present particle accelerators. In order to reach higher energy scales, a concept for a next generation particle accelerator had to be found. One constraint for the new machine was born in terms of financial aspects leading to the decision to re-use the already existing LEP tunnel at CERN. Another limitation was given due to physical reasons. The maximum energy of circular particle accelerators is limited by the amount of energy loss due to synchrotron radiation [7]:

$$-\Delta E = \frac{4\pi\alpha\hbar c}{3R}\beta^3\gamma^4 \quad \text{with} \quad \beta = \frac{v}{c} \approx 1 \quad \text{and} \quad \gamma = \frac{E}{mc^2} \tag{3.1}$$

Since the radius of the ring is fixed, the only way of significantly increasing the centre-of-mass energy of the system is to use heavier particles for acceleration. Due to large differences in rest mass, the loss through radiation for electrons is increased by a factor of

[1] Large Electron-Positron Collider, Operation: 1989 - 2000, Energy: $\sqrt{s} \approx 91$ - 210 GeV

[2] European Organization for Nuclear Research, formerly Conseil Européen pour la Recherche Nucléaire

10^{13} with respect to the negligible energy loss in case of protons at the same beam energy at parton level. Basically designed as a proton-proton collider, the LHC is currently being built in the 27 km long circular former LEP tunnel up to 175 m below the surface, crossing the Franco-Swiss border. Figure 3.1 gives an aerial view of the countryside above the LHC and shows the geographical position of the four main LHC experiments described below. The energy of the two counterrotating proton beams of the LHC will reach 7 TeV each, resulting in collisions at a centre-of-mass energy of $\sqrt{s} = 14$ TeV. On this energy scale, physicists hope to verify well known SM parameters as well as to get indications on electroweak symmetry-breaking as a result of the Higgs mechanism or even to find new physics beyond the Standard Model.

In order to achieve such high energies, the protons have to pass through different acceleration stages whereas the LHC itself is just the final part. Figure 3.2 shows a schematic view of the acceleration. Further details beyond the scope of this chapter can be obtained by the LHC design reports [8], [9] and [10].

Figure 3.1: Aerial view of the countryside above the LHC and geographical position of the four main LHC experiments ALICE, ATLAS, CMS and LHCb

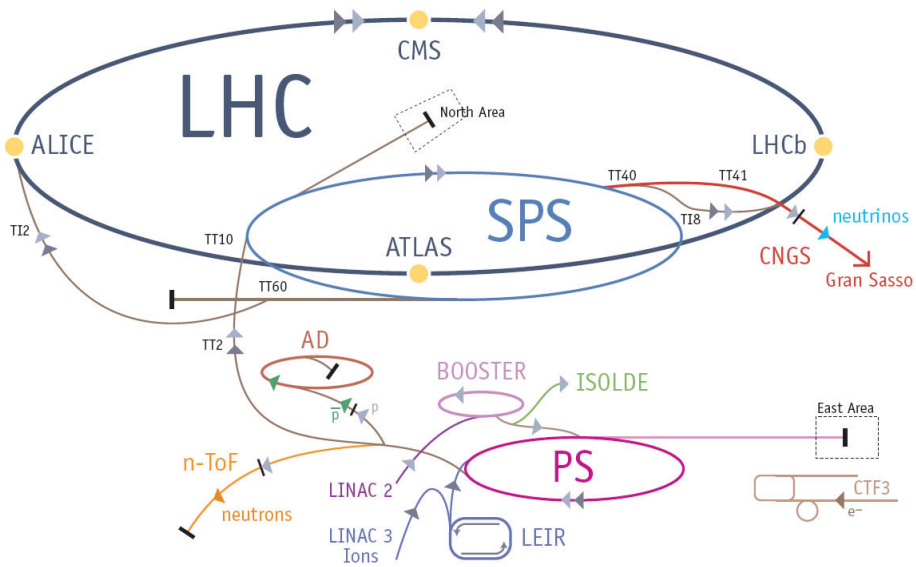


Figure 3.2: Schematic view of the CERN accelerator chain from the proton injection up to the final collision energy inside the LHC main ring [11]

Proton Production and Linear Acceleration

The LHC has the ability to collide either protons or heavy ions (e.g. lead) at several given interaction points around the main ring. Since the main LHC operation mode will be the collision of two proton beams, the production and acceleration of heavy ions is to be ignored here. The first step of the proton production is the ionisation of hydrogen gas through a Radio Frequency Quadrupole (RFQ) duoplasmatron creating a negatively charged ion beam of 750 keV. This beam is furthermore accelerated by a linear accelerator (LINAC2) and passes through a carbon foil at the end of the machine. The foil strips off all orbiting electrons to generate a pure proton beam at an energy of 50 MeV and a current of 180 mA which is the requirement for the LHC at the stage of the next injection step.

Booster and Proton Synchrotron

The proton beam is fed into the Proton Synchrotron Booster (PSB) increasing its energy to about 1.4 GeV and then, in the next step of the acceleration chain, injected to the Proton Synchrotron (PS) itself, a circular accelerator with a circumference of about 630 meters. The PS served as beam source for several CERN experiments, e.g. ISR[3] and LEP. Since

[3]Intersecting Storage Rings, the worlds first hadron collider

its completion in 1959 it has passed through several upgrade stages and is now capable of accelerating the protons to an energy of 26 GeV.

Super Proton Synchrotron

The last step of the LHC pre-injector chain is the Super Proton Synchrotron (SPS). Since its commissioning in 1976, the SPS has been used to accelerate electrons, positrons, protons, antiprotons and heavy ions for a variety of experiments. Today it provides proton beams for several fixed-target experiments like Compass[4] and NA48[5]. With the start-up of the LHC, the SPS picks up the proton beam from the PS and accelerates it to a final energy of 450 GeV. Afterwards, the beam is split and extracted via two transfer lines into the LHC main ring in opposite directions.

LHC Main Ring

The final part of the proton acceleration is the velocity increase through radio frequency cavities implemented in 1232 superconducting dipole magnets with a length of 15 m inside the LHC. Within one circulation, the beam gains 0.5 MeV of energy and reaches its final state after about 20 minutes of operation. The dipole magnets have to provide a nominal magnetic field of 8.33 T to keep the 7 TeV protons on their circular path around the ring. Reaching such high magnetic fields required the development of new types of superconducting magnets. The coils of the magnets are made up of copper-clad niobium-titanium (NbTi) cables which are chilled to the temperature of superfluid helium of about 1.9 K. In addition to these bending magnets, several other types are required for a proper operation of the LHC, e.g. 392 quadrupole magnets counteracting a defocusing of the beam.

Table 3.1 gives a short summary of the accelerator chain and shows the corresponding beam energies in each of the different acceleration stages.

LHC Machine Parameters

The LHC experiments aim at identifying interesting collisions at the different interaction points (IP) around the ring. Most of these events are very rare due to a very small production cross section, σ_{event}. The cross section is a measure for the probability to observe a desired interaction with respect to all events generated by the accelerator. The

[4]Common Muon and Proton Apparatus for Structure and Spectroscopy
[5]Fixed target experiment in the field of kaon physics

Table 3.1: Energy of the proton beam at each acceleration step

Accelerator	Injection Energy	Final Energy
LINAC2	750 keV	50 MeV
PSB	50 MeV	1.4 GeV
PS	1.4 GeV	26 GeV
SPS	26 GeV	450 GeV
LHC	450 GeV	7 TeV

amount of such events produced per second is given by

$$\dot{N}_{events} = \mathcal{L}\ \sigma_{event} \tag{3.2}$$

and directly linked to the luminosity $\mathcal{L}$ of the machine. In case of colliders, each beam has a well defined structure and is composed of a certain amount of bunches k_b, where each bunch contains N_p particles. The luminosity is then given by

$$\mathcal{L} = \frac{\gamma f k_b N_p^2}{4\pi\epsilon_n\beta^\star} F \quad \text{with} \quad \gamma = \frac{1}{\sqrt{1-\beta^2}} \tag{3.3}$$

for beams with a rotation frequency f. The equation contains as well the Lorentz factor γ, the betatron function $\beta^\star$ at the intersection point, the normalized emittance in transverse direction ϵ_n and the reduction factor F due to the crossing angle of the beams. In order to achieve a fair amount of desired high energy collision events, the luminosity of a hadron collider has to be increased since the cross section of such events decreases with the squared energy of the hard parton-parton interaction, E^2. The LHC is therefore designed to achieve a luminosity of $\mathcal{L}_{high} = 10^{34}$ cm^{-2}s^{-1} during the so-called high luminosity phase, but will be operated at a reduced luminosity of $\mathcal{L}_{low} = 2 \times 10^{33}$ cm^{-2}s^{-1} during the first three years of operation. The two beams contain 2808 bunches each with a density of 1.15×10^{11} protons per bunch. The nominal spacing between two bunches is 25 ns and is required by the front-end electronics of the different detector components in order to recover properly. Based on this configuration, the estimated collision frequency is about 40 MHz at each interaction point. Since the number of protons per bunch is very large compared to current particle accelerators, a main drawback in case of the LHC is the number of about 20 concurrent collisions within one bunch crossing at design luminosity. This effect is called *pile-up* and complicates the detection and reconstruction of events and therefore makes the possibility

of new discoveries more challenging. Table 3.2 gives a summary of some important LHC parameters.

Table 3.2: Important LHC parameters in case of proton-proton collisions

Parameter	Value
Energy per proton	$E = 7$ TeV
Magnetic field of bending dipoles	$B = 8.33$ T
Cooling temperature of dipoles	$T = 1.9$ K
Ring circumference	$R = 26.658$ km
Rotation frequency	$f = 11.2$ kHz
Luminosity (low)	$\mathcal{L}_{low} = 2 \times 10^{33}$ cm^{-2}s^{-1}
Luminosity (high)	$\mathcal{L}_{high} = 10^{34}$ cm^{-2}s^{-1}
Number of bunches	$k_b = 2808$
Number of protons per bunch	$N_p = 1.15 \times 10^{11}$
Bunch spacing	$t = 25$ ns
Betatron amplitude value at IP	$\beta^\star = 0.55$ m
Normalised transverse emittance	$\epsilon_n = 3.75$ µm
RMS beam radius at IP	$\sigma^\star = 16.7$ µm
Number of concurrent collisions at $\mathcal{L}_{high}$	$n_c \approx 20$
Crossing angle of the beams	$\alpha = 300$ µrad

Another important measure for the collected data at particle accelerators is the integral of the luminosity with respect to time, $L = \int \mathcal{L} dt$. (The unit of L is the reciprocal value of barn which is defined as 1 b $= 10^{-28}$m^2.) The projected integrated luminosity of the LHC is about $10 - 30$ fb^{-1} for the first years of operation. Later, it is planned to increase the amount to $100 - 300$ fb^{-1} per year.

3.2 The LHC Experiments

The LHC hosts six experiments at its four underground caverns distributed around the main ring. There are two general-purpose detectors ATLAS and CMS. The CMS experiment will be described in more detail later in this chapter. The other four experiments namely ALICE, LHCb, LHCf and TOTEM are dedicated to specific analyses.

- **ALICE - A Large Ion Collider Experiment:**
  The experiment is focused on the analysis of heavy lead-ion collisions in order to study properties of a quark-gluon plasma. The conditions in the plasma are supposed to be equal to the state of matter just after the Big Bang where quarks and gluons can be considered as free particles and are not yet confined inside hadrons, like protons or neutrons. The detector consists of a central barrel and a single arm forward muon spectrometer. Its size adds up to 26 m in length, a diameter of 16 m and the weight is about 10 000 t.

- **ATLAS - A Toroidal LHC ApparatuS:**
  ATLAS is, besides CMS, one of the two general-purpose experiments designed to cover the widest range of physics which can possibly be discovered at the LHC. Main feature of the detector is the toroidal-shaped magnet system which provides a homogeneous magnetic field inside the torus. ATLAS is the largest-volume collider detector ever built with a total length of 46 m and a diameter of 25 m for the barrel and endcap system.

- **LHCb - Large Hadron Collider beauty:**
  Main purpose of LHCb is the study of a slight asymmetry between matter and antimatter present in interactions of particles containing bottom quarks and is intended to answer fundamental questions on the composition of our universe. The experiment is built to detect particles with a low transverse momentum going forward close to the beam line. The detector consists of a forward spectrometer with planar detectors lined up one behind the other over a length of about 21 m.

- **LHCf - Large Hadron Collider forward:**
  LHCf is a small experiment built inside the ATLAS underground cavern 140 m off the interaction point. The motivation for this experiment is to test several models used to estimate the primary energy of ultra high-energy cosmic rays. It will measure forward directed particles like neutral pions produced at the ATLAS interaction point close to the beam pipe and consists of two 30 cm long detector components.

- **TOTEM - TOTal Elastic and Diffractive Cross-Section Measurement:**
  TOTEM is designed to measure the effective size (cross-section) of the proton at the LHC and is also used to monitor the actual luminosity of the LHC machine. The detector is hosted inside the CMS cavern. In order to achieve a high precision in

detecting particles very close to the beam pipe, specially designed vacuum chambers, so-called "Roman pots", host the detector components and are directly connected to the beam pipe. Eight of them are placed in pairs at four locations distributed near the interaction point.

3.3 The Compact Muon Solenoid

The second general-purpose detector at the LHC is the Compact Muon Solenoid. The studies presented in this thesis are directly related to this experiment. Therefore, the main detector components are described in detail. Further information beyond this chapter about the composition of the detector subsystems, software framework, computing model and expected detector and physics performance are discussed in [12], [5], [13] and [14]. Although designed to reach the same physics goals as ATLAS, the set-up of CMS differs from the one of its competitor. Based on a cylindrical solenoid, the onion-like layout of the detector provides an almost hermetic coverage around the main interaction point. Several subdetector components required for the measurement of different kinds of particles and their properties are arranged in a layered structure forming a central barrel enclosed by two flat endcaps. The overall length of CMS is about 21 m with a diameter of 15 m which adds up to a weight of 12 500 t.

3.3.1 Coordinate Conventions

The coordinate system of CMS has its origin inside the detector at the primary interaction point. The x-axis points radially towards the centre of the LHC, whereas the y-axis points vertically upward. Thus, the z-axis shares the same direction with the beam line and is positive defined towards the Jura mountains. The azimuthal angle ϕ is measured from the x-axis in the $x - y$ plane whereas the polar angle θ is measured from the z-axis. Particle physicists often use a Lorentz invariant quantity called rapidity y instead of θ. It is defined as

$$y = \frac{1}{2} \ln \left(\frac{E + p_z}{E - p_z} \right) \tag{3.4}$$

and equals, in case of massless particles, the so-called pseudorapidity η given by

$$\eta = -\ln \left(\tan \left(\frac{\theta}{2} \right) \right) \tag{3.5}$$

The angular distance between two particles observed from the origin of the coordinate system is $\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}$. Measurable quantities like momentum and energy

transverse to the beam line are denoted by p_{T} and E_{T}, respectively, and can be derived from its x and y components.

3.3.2 Detector Requirements and Overall Layout

Since CMS is tend to detect a wide variety of physics processes, the detector was designed to fulfil several benchmark requirements in order to be ready for new discoveries. During the design phase of CMS and ATLAS in the early 1990s, the proof of electroweak symmetry breaking through detection of one or more Higgs bosons was defined as a goal of these new experiments. The diversity of decay modes depending on the mass of the Higgs boson makes this a particularly appropriate benchmark. The current lower limit of $m_H = 114.4$ GeV/c^2 on the Higgs boson mass with a confidence level of 95% is obtained from former LEP experiments and points, together with theoretical predictions, at dominating hadronic final states in this energy region. Due to the large QCD background and poor mass resolution that can be obtained with jets, a clear focus of CMS lies on the detection of final states containing isolated leptons, although the branching fractions of these processes are much smaller. This results in the requirement for an excellent muon tracking system capable to cover a wide range of momenta up to about 1 TeV, and to provide a good dimuon mass resolution of $\approx 1\%$ at 100 GeV/c^2 as well as a precise reconstruction of the production vertex. The electromagnetic calorimeter has to yield an accurate electromagnetic energy resolution as well as good diphoton and dielectron invariant mass resolutions comparable to the one of the muon system.

Another main pillar is the search for supersymmetric particles leading to event topologies with a significant amount of missing transverse energy $E_{\mathrm{T}}^{\mathrm{miss}}$ in the final states. This is a result of weakly interacting lightest supersymmetric particles present in all supersymmetric decay modes (including cascade decays). In order to achieve a good measurement of the missing transverse energy, the coverage of the hadronic calorimeter has to be almost hermetic and the obtained dijet mass resolution has to be good.

The LHC will also allow to study QCD effects, electroweak and flavour physics like multiple jet events and top-quark production, CP-violation in B-hadron decays, lepton flavour conservation measurements as well as precise measurements of the CKM mixing matrix. These precision studies and the comparison to direct searches may lead to indications for physics beyond the Standard Model as well. All these studies require in addition to the detector subsystems mentioned above a very good silicon based tracking system in order to achieve a good momentum resolution of charged particles as well as efficient track reconstruction abilities. The silicon tracking system is a crucial part with respect to the b-tagging studies presented later in this thesis.

There are several other topics that will be studied with needs similar to the mentioned ones. Among them are searches for extra dimensions, mini black holes, heavy vector bosons and studies concerning quark gluon plasma as a result of heavy ion collisions. Figure 3.3

presents a schematic view of CMS containing the main subdetector components which will be described in the following sections. Figure 3.4 shows the interaction behaviour of different kinds of particles including muons, electrons, photons and hadrons traversing the detector.

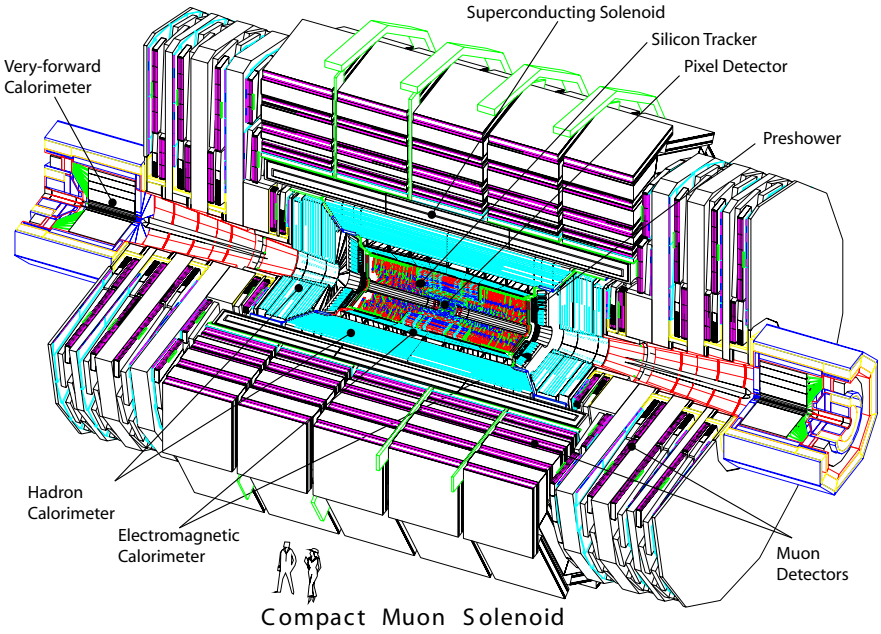


Figure 3.3: Schematic view of CMS including the main detector subcomponents [12]

3.3.3 Solenoid Magnet

Eponym for the CMS detector is a superconducting solenoidal magnet producing an almost homogeneous field of $B = 4$ T along the beam axis. The magnet type and field strength was chosen due to the bending power needed for muons with energies up to 1 TeV, to determine their sign and to provide a good momentum resolution. The solenoid has a length of 12.9 m and a diameter of 5.9 m which results in 2.7 GJ of stored energy inside the chilled superconducting niobium-titanium coil. The magnetic field in the iron return yoke, which hosts several layers of muon chambers, is saturated at about 2 T. The silicon tracking system and the electromagnetic and hadronic calorimeters are situated just inside the solenoid making the detector compact in order to reduce disturbing interactions of particles with the coil material.

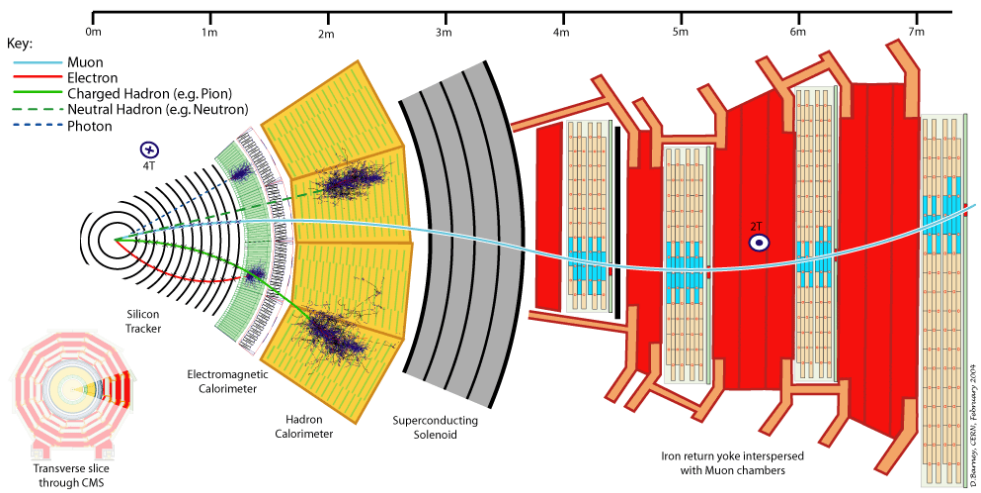


Figure 3.4: A slice through the CMS detector in $r - \phi$ view. Sample trajectories and the interaction behaviour of different kind of particles are shown for muons, electrons, photons and hadrons

3.3.4 Silicon Tracking System

The innermost part of CMS is its silicon tracking system [15]. The possibility to reconstruct charged particle tracks with high precision is needed by many parts of the CMS physics program. A high resolution on the momentum measurement, as well as the ability to distinguish tracks coming either from the primary interaction point or from a displaced secondary or tertiary vertex is indispensable. High reconstruction efficiencies for particles with low transverse momenta between 1 and 5 GeV/c are for example necessary to measure isolated objects coming from gauge boson decays, whereas an exact vertex reconstruction helps in case of identifying particles originating from decays of b-quarks or τ-leptons (so-called tagging). Studies concerning the CMS b-tagging performance and the impact of a misaligned silicon tracking system are presented in chapter 5.

Figure 3.5 gives a schematic view of the layout of the silicon tracking system in the $r - z$ plane and shows the position of the different pixel and strip layers in the barrel region and the endcaps.

- **Pixel Tracker:**
  Due to the high particle flux near the interaction point, a silicon pixel detector was installed closest to the beam pipe in order to keep the occupancy low despite the high collision rate of the LHC. The pixel detector is arranged in three barrel layers at radii of 4.4, 7.3 and 10.2 cm as well as two endcap disks in forward and backward direction. Overall length of the device adds up to 53 cm, covering an area of ≈ 1 m^2 with a total amount of 66 million pixels [16]. Each pixel has a size of 100×150 μm^2.

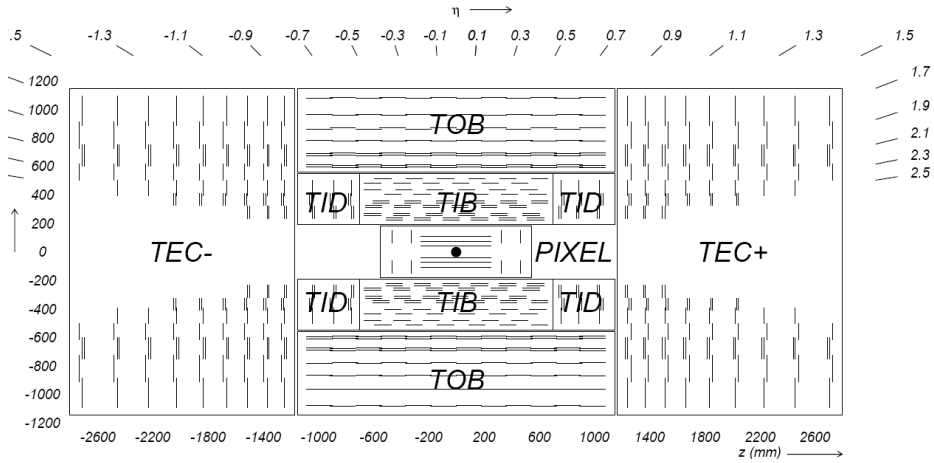


Figure 3.5: The CMS tracking system in $r - z$ view. Double lines indicate stereo modules whereas single lines correspond to the position of single-sided silicon strip modules. The coverage goes up to $|\eta| < 2.4$ [14]

The pixel barrel (PXB) consists of 768 modules, each combining several single pixels; always four of these modules are arranged into half-ladders. In case of the endcaps (PXE), a turbine-like structure, assembled by rotated blades, was chosen, where each blade contains 7 modules. The spatial resolution in $r - \phi$ is improved by the Lorentz effect due to the large Lorentz angle of 23° and averages out at about 10 µm whereas the resolution in z direction is about 20 µm.

- **Strip Tracker:**
  Just outside of the pixel detector, a silicon strip detector is attached. 9.6 million silicon strips are arranged in a central barrel region with the so-called tracker inner barrel (TIB) and the tracker outer barrel (TOB) as well as the tracker inner disks (TID) covering an intermediate region between the inner barrel and the two tracker endcaps (TEC). Depending on the location (see figure 3.5), either single- or double-sided modules with a stereo angle of 100 mrad are used in order to provide a measurement in both $r - \phi$ and $r - z$ coordinates. The strip pitch varies from 80 to 180 µm and allows a spatial resolution of 23 to 52 µm in $r - \phi$ and 230 to 530 µm in z. The entire strip detector is mounted on carbon-fibre structures inside a chilled outer support tube with an operation temperature of around −20 °C. The area covered by the strip detector adds up to about 200 m^2 and therefore allows measurements up to $|\eta| < 2.4$ for the whole system.

The functionality of a single-sided silicon strip detector is depicted in figure 3.6. Silicon detectors are p-n junction diodes operated at reverse bias. They consist of a n-type silicon wafer with strongly p-doped implants on the wafer surface forming a charge depletion area in between. Read-out stripes are attached through a thin insulator-film capacitively to

the p-doped implants whereas a strongly n-doped silicon layer followed by a metal coating encloses the wafer on the other side. Charged particles traversing the silicon stripes interact with the silicon and create electron-hole pairs inside the material. Amplified by the applied high-voltage, the holes and electrons travel to their corresponding electrodes and a current can be measured by the attached read-out electronics in order to determine the exact position of the charge deposit in the material, called hit. A fitting procedure combines several of these hits and calculates the trajectory of the incident charged particle. A measurement of the momentum and sign of the particle can be obtained by taking into account the curvature of the track caused by the magnetic field.

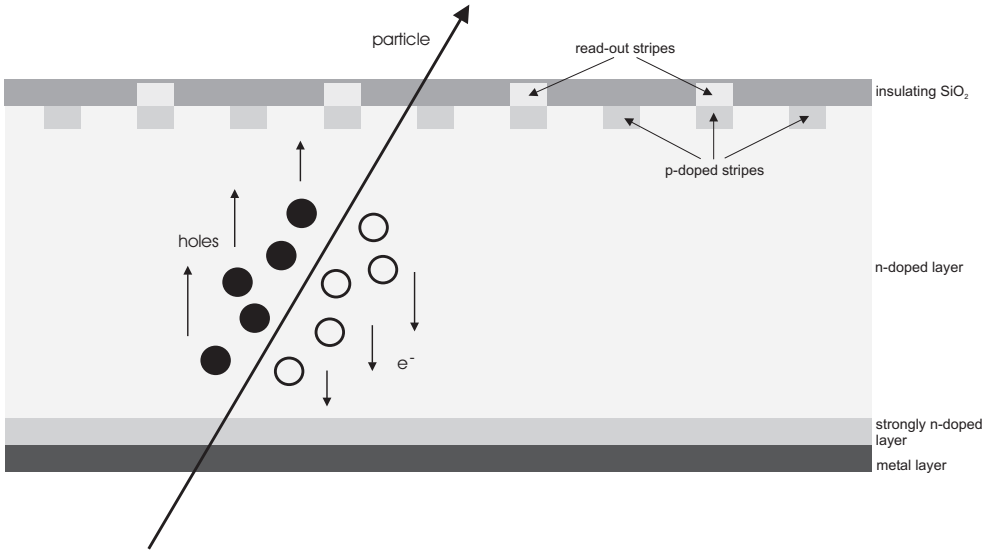


Figure 3.6: Schematic layout of a single-sided silicon strip detector and a denoted interaction of a charged particle traversing the material

3.3.5 Electromagnetic Calorimeter

The silicon detector is enclosed hermetically by the electromagnetic calorimeter (ECAL). Purpose of the ECAL is the precise energy measurement of electrons, photons and, together with the hadron calorimeter, jets. It consists of 61 200 lead tungstate ($PbWO_4$) crystals in the barrel (EB) part as well as another 14 648 crystals in the two endcaps (EE), resulting in a η-coverage of up to 3.0. An additional silicon preshower detector (SE) is situated between TEC and EE in order to separate single photons from photon pairs originating from the decay of neutral pions in the endcaps and to help to distinguish electrons from minimum ionising particles.

Figure 3.7 shows the layout of the CMS electromagnetic calorimetry system in a schematic $r - z$ view with respect to the other detector components inside the solenoid. Important

reasons for choosing $PbWO_4$ as scintillating material is the short radiation length and Moliere radius of $X_0 = 0.89$ cm and $r_M = 2.2$ cm, respectively. Due to this, a good coverage in interaction lengths and therefore a good energy resolution can be guaranteed and the ECAL nevertheless stays compact. Other advantages are the fast response time of $t = 25$ ns in which 80% of the light is emitted, and its radiation hardness of up to 10 Mrad. A drawback due to the use of $PbWO_4$ is the relatively low light emittance which requires the adoption of photodetectors capable to operate in magnetic fields. For this purpose, silicon avalanche photodiodes (APDs) are used in the barrel region since they are insensitive to high axial magnetic fields, whereas vacuum phototriodes (VPTs) are used in the endcaps due to the high particle flux. More detailed information concerning the ECAL can be found in [17].

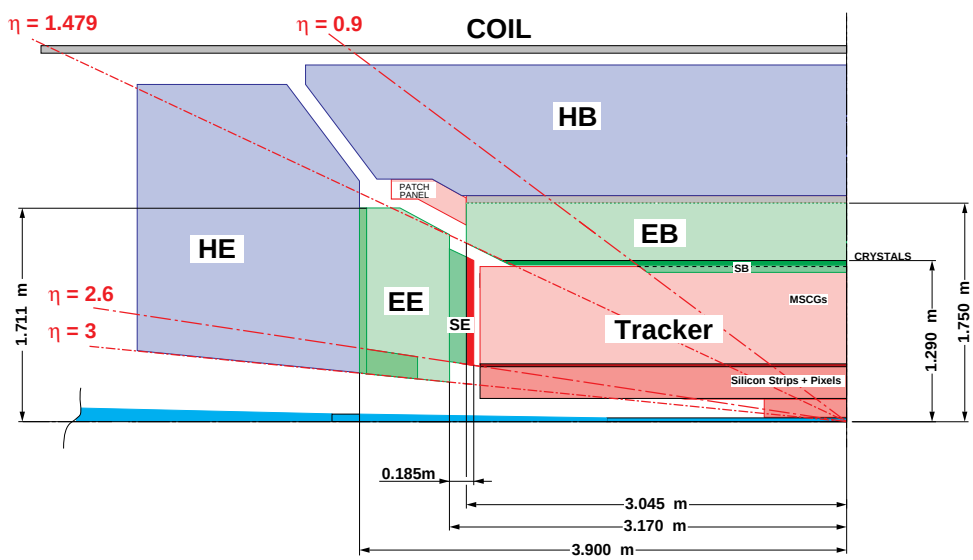


Figure 3.7: One quarter of the CMS detector subcomponents inside the magnetic coil. The schematic view includes the silicon tracking system and the electromagnetic and hadronic calorimeter system in $r - z$ view [17]

3.3.6 Hadronic Calorimeter

The CMS hadron sampling calorimeter (HCAL) [18] is directly bordered outside the electromagnetic one. It is used together with the ECAL to measure the energy and direction of hadronic particle jets and, due to a comparison between energy sum and centre-of-mass energy, able to give an estimate on the missing transverse energy (arisen from e.g. non-detectable neutrinos or supersymmetric decay modes). Since photons and electrons are totally absorbed in the ECAL, the HCAL indirectly allows the identification of these particles as well. In order to reach a good measurement of $E_{\mathrm{T}}^{\mathrm{miss}}$, the HCAL maximises the

material inside the coil in terms of interaction lengths. It is built of alternating layers of plastic scintillator tiles with embedded wavelength-shifting (WLS) fibres and brass plates serving as absorber material.

The hadron calorimeter also shows the characteristic barrel (HB) and endcap (HE) structure present in all over CMS (see figure 3.7). The scintillating material is combined in towers with a segmentation of $\Delta\eta \times \Delta\phi = 0.087 \times 0.087$ in the barrel and $\Delta\eta \times \Delta\phi = 0.175 \times 0.175$ in the endcap region. In addition to HB and HE, the hadron outer detector (HO) consisting of another scintillating layer is directly attached outside the magnetic coil in order to increase the accuracy for high energetic jets. Together with the hadron forward detector (HF), a coverage in pseudorapidity of $|\eta| < 5.0$ can be reached and the hermeticity with respect to $E_{\mathrm{T}}^{\mathrm{miss}}$ measurements can be maximised. Another benefit due to the HF is the possibility to tag or veto jets in forward direction.

3.3.7 Muon Tracking System

The name of the experiment already suggests that the muon system [19] of CMS plays an important role in order to meet the challenges presented by the LHC. Due to their high mass and long lifetime, muons provide a very clean and experimentally well measurable signature.

In order to reject muons originating from high energy cosmic rays and to focus on muons coming from the primary interaction point, a three-step-measurement in the inner tracking system, just outside the coil and inside the return flux, is necessary. The muon system itself consists of three different detector subsystems embedded in the magnetic return yoke, namely drift tube chambers (DT) in the barrel region, cathode strip chambers (CSC) in the endcaps as well as resistive plate chambers (RPC) in both, the endcaps and the barrel part. The overall pseudorapidity region covered by the muon system is $|\eta| < 2.4$ as can be seen in figure 3.8. There are 250 DT chambers organized in 4 layers (MB1-4) in the barrel part of CMS with $|\eta| < 1.2$ and 468 CSC in 4 layers (ME1-4) with $0.9 < \eta < 2.4$ in the two encaps. The resistive plate chambers provide a worse spacial resolution for the muon track reconstruction than the other two detector types, but this drawback is compensated by the very fast response time of the RPC.

3.3.8 Data Acquisition and Trigger System

As already mentioned in the first part of this chapter, the bunch crossing rate of the LHC leads to approximately 10^9 interactions per second in case of running at the design luminosity of $\mathcal{L}_{high} = 10^{34}$ cm^{-2}s^{-1}. The nominal event size is about 1.5 MB in case of proton-proton collisions which results in roughly 100 TB of data per second. This huge amount is too large to be handled by current storage and network solutions and therefore has to be trimmed to more reasonable values through a sophisticated cascaded trigger

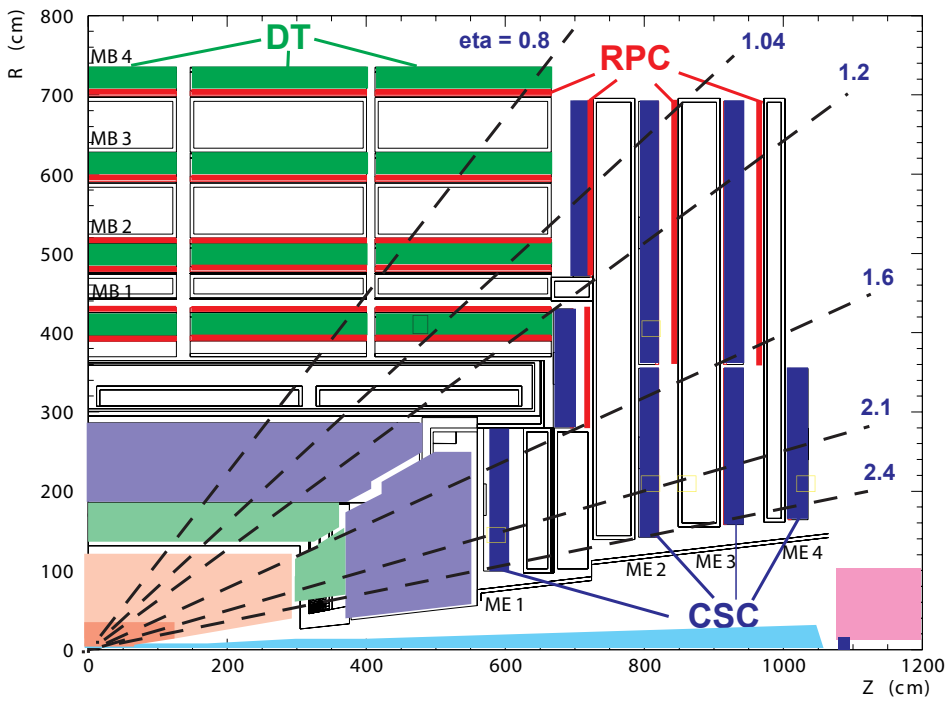


Figure 3.8: One quarter of the CMS muon system in $r - z$ view. The overall coverage in pseudorapidity is $|\eta| < 2.4$ for the DT, RPC and CSC muon detectors [12]

system. The main challenge is therefore the selection of interesting events among the huge amount of background which has to be reduced to a manageable amount. The event rate which can be recorded for offline processing and analysis is at the order of 150 Hz.

In order to reach this, the CMS trigger and data acquisition system (DAQ) includes four parts, namely the detector electronics, the first-level trigger processors (Level-1), the readout network and an online filter system which runs the high-level trigger (HLT) software on a PC farm. Figure 3.9 depicts the architecture of the CMS DAQ and trigger system.

The Level-1 trigger is implemented in hardware in order to achieve the needed response time. It is capable of reducing the data rate from about 40 MHz to 100 kHz by evaluating measurements made by the calorimeters and the muon system. In the meanwhile, the full event is stored in hardware buffers and passed to the next trigger step in case the Level-1 trigger made a positive decision. The readout network has to handle about 100 GB of data per second which is directly passed to a computing farm running the high level trigger software.

The HLT provides a full reconstruction of the event topologies, if needed, and allows an event-filtering according to several desired physics goals. The PC farm allows maximum flexibility in terms of the optimisation of selection algorithms or trigger thresholds as well

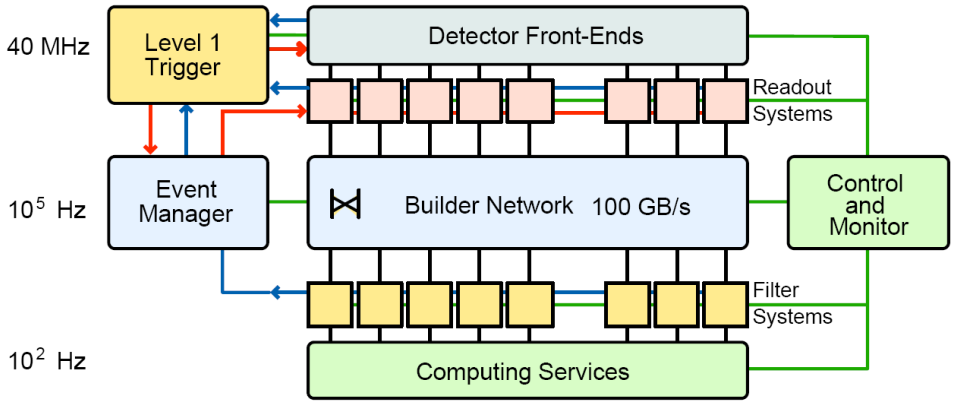


Figure 3.9: Schematic view of the CMS trigger and data acquisition system [14]

as profiting from the fast evolution of computing technology during the years of operation. Due to this last step, a manageable data rate of about 150 Hz is reached and the events are recorded on a central storage system based at CERN. More detailed information about the DAQ and trigger system can be found in [20] and [21].

Although the event rate can be decreased by five orders of magnitude, the overall amount of data will reach several PB per year. The requirements on the analysis software, computing power and storage capacity are immense. The needed resources could not easily be provided by a few or even single institutions. The chosen approach based on globally distributed computing and storage resources is presented in chapter 4. An overview to the CMS analysis software framework is given in the next section of this chapter.

In summer 2006, a first test of the complete data taking procedure was performed during the Magnet Test and Cosmic Challenge (MTCC). For this purpose, only a slice of the CMS detector was installed. Goal of this challenge was the detection of cosmic rays in the muon chambers which were used to trigger the readout of the remaining detector parts installed in CMS. During this test, the superconducting solenoid magnet system was operated at its nominal field of 4 T. The results obtained at the MTCC are summarised in [22] and [23].

3.4 The CMS Software Framework

Between the raw data obtained by the read-out electronics of the different detector components and a final analysis of a scientist lie many steps that have to be performed. In contradiction to former approaches with different software components for each of these steps, the CMS collaboration decided about two years ago to change the former software setup and start over with a new software framework simply called CMS Software

(CMSSW) [24].

The new framework is completely written in C++ and provides the possibility to use various external software packages and libraries written in other programming languages as well. It combines Monte Carlo (MC) simulation, HLT, reconstruction, filtering and analysis software in a modular manner. CMSSW is completely assessable by a configuration file and therefore the software behaviour, cuts on several observables or used algorithm types and their behaviour can be easily adjusted or exchanged. For illustration purposes, a sample CMSSW configuration file and a short description of its main parts can be found in appendix A.1.

3.4.1 Software Components

- **ROOT:**
  ROOT [25] is an object oriented framework written in C++. Its emphasis is on the analysis of large amounts of data and it is widely used in high energy physics. An advantage of ROOT is the usage of specially designed storage methods which allow a direct access to single desired objects stored in a file without touching the rest of the data. ROOT has also the ability to create various types of histograms or to provide arbitrary fitting methods among several other things.

  The framework can be used in three different ways: a compiled macro with the needed libraries linked to it, a command line C++ interpreter called CINT and a graphical user interface which can be interactively used in order to remove several time consuming compile and link cycles during the phase of analysis development. In order to benefit from large computing clusters and multi-processor machines, a parallelisation approach for the ROOT framework has been developed [26].

- **Event Data Model:**
  A basic part of CMSSW is the so-called Event Data Model (EDM). The EDM stores all information that can be obtained from a physics collision event and behaves as a container for all kinds of C++ objects like the raw read-out data from the different detector subsystems, reconstructed physics objects like particle tracks or jets as well as calculated observables derived from kinematical or topological event properties. All events and their information are stored in ROOT files with a treelike structure. They can be accessed by using the CMS software framework or even directly by ROOT when linking the needed libraries in order to access the desired C++ objects.

  Many physics analyses do not rely directly on all the event data stored in the ROOT trees and therefore, CMS provides several subsets of the recorded data samples in order to reduce needed storage space and computing time of the analyses significantly. The RECO data format contains all reconstructed objects whereas the Analysis Object Data (AOD) only contains a subset of physical objects for specific analyses. In

addition to these, scientists have the possibility to produce own data samples by individually selecting and storing parts of the whole event.

- **Framework Services and Event Setup:**
  During data taking, several detector conditions and parameters may change. In order to reflect these changes properly in the accomplished analyses, the CMS software framework has to store such meta information in the event data, too. This includes, among others, most basic information like run and event number as well as more sophisticated calibration data of the different detector subsystems, their alignment position and the condition of the magnetic field. The meta data is not linked to single events but considered as valid for a specific period of time (IOV). Figure 3.10 depicts how the event set-up can be determined for the collision events recorded during an IOV.

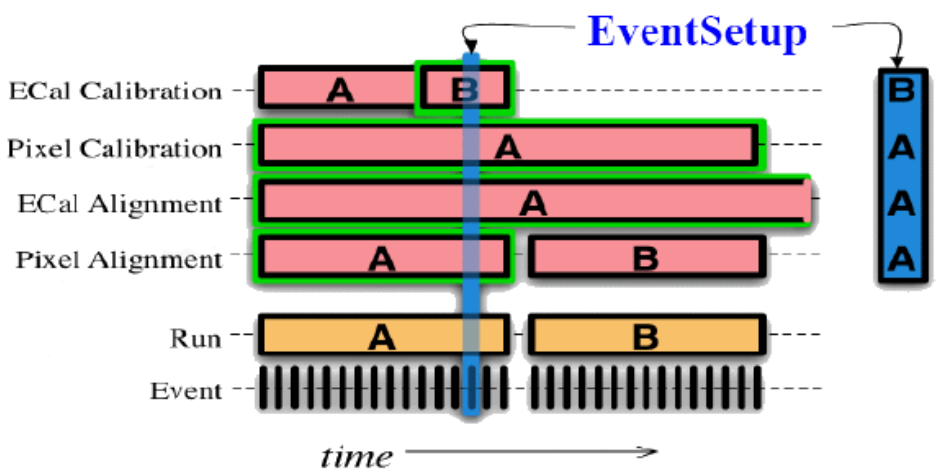


Figure 3.10: Illustration of the CMS event set-up. Each collision event is based on a specific set of calibration information, alignment position of the detector subsystems or other parameters stored in the event meta data [24]

3.4.2 Event Generation and Detector Simulation

Since there is no real physics data available until the start of the LHC in summer 2008, the written software and the algorithms have to be tested and verified on simulated Monte Carlo events in order to ensure a proper operation and well understood algorithm behaviour as soon as the data taking period begins. Monte Carlo data is used as well for comparison with real data and gives estimates on the number of expected background events in case of specific studies and physics searches. The production of realistic Monte Carlo events is done in four steps:

1. A high-energy-physics event is produced by generating a set of outgoing particles coming from the interaction between two incident particles e.g. protons in case of the LHC. CMSSW interfaces to various event generators like PYTHIA [27], HERWIG [28] or SHERPA [29]. The events used in this study have been obtained by using the default CMS Monte Carlo event generator PYTHIA in version 6.4. It is a general-purpose event generator written in FORTRAN 77 which includes several processes like hard and soft interactions, initial- and final-state parton showers, hadronisation and underlying event activity among other things. In addition, many processes beyond the Standard Model have been included. PYTHIA was developed in 1978 and is therefore well tested and accepted in the high-energy physics community.

   The initial interactions of the incoming partons of the two protons are calculated in leading order precision. The outgoing partons from the hard interaction are then passed to the parton shower model together with the remnants of the protons. Radiation effects and higher order calculations are applied and the so-called partonic final state is formed. Next step is the hadronisation by using the Lund-String fragmentation model where outgoing hadrons are produced. They act as input for the upcoming detector simulation step described below. A schematic view of the PYTHIA workflow is presented in figure 3.11. More details about the different generation steps can be found in [27].

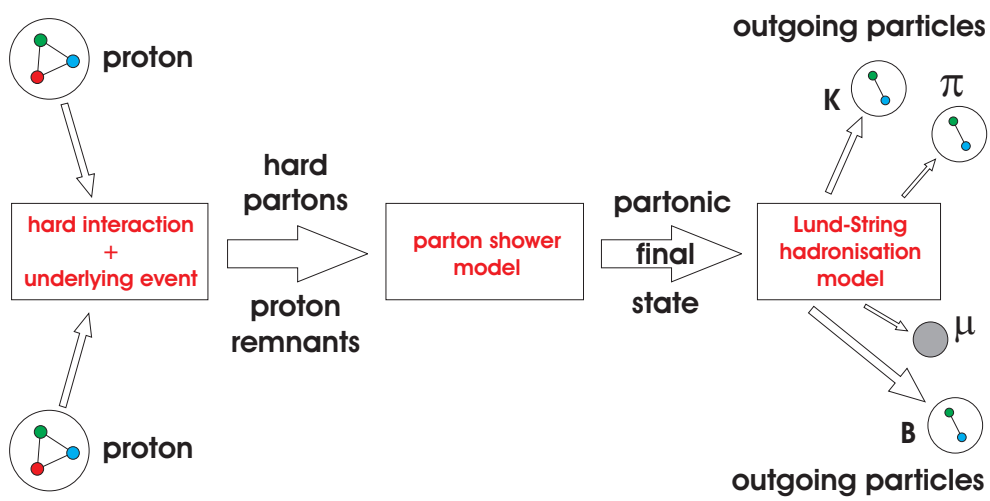


Figure 3.11: Schematic view of the PYTHIA event generation workflow. The hard interaction between two partons of the incident protons is calculated. The outgoing hard partons and the proton remnants are then passed to the parton shower model and the partonic final state is formed. Afterwards, a hadronisation procedure is performed and outgoing hadrons according to the Lund-String model are produced

2. Each LHC proton and therefore the interaction region has a spatial spread. Thus, the location of the primary collision point varies with respect to the centre of the

detector. In order to reflect this as well in the simulated events, the primary vertex has to be smeared out by several available smearing algorithms, e.g. flat or Gaussian shaped, based on a beta-function or a test beam profile.

3. The generated event is then passed to the detector simulation of CMS. It is based on the Geant4 [30] toolkit together with a detailed description of the detector volume and material, which has knowledge about the sensitive detector parts with read-out electronics attached as well as the rest of the detector like support structures or cabling material. Geant4 simulates the propagation of the decay particles through the detector by taking e.g. the magnetic field, interactions with the detector material and thus created secondary particles into account. The location of each particle's interaction point with a detector module is recorded as so-called hit in the event. The needed computing time for simulating events with Geant4 is quite large an can take up to several minutes per event depending on the simulated decay products present in the event. CMS therefore provides an additional approach called Fast Simulation [31] to be able to produce much larger numbers of events on smaller time scales which could be necessary for specific analyses or background estimates. Several simplifying assumptions, parametrisations and optimised algorithms lead to much lower computing times for producing events with the Fast Simulation but in turn to a lower accuracy with respect to the full Geant4 simulation.

4. Last part is the so-called digitisation step. Here, the electronics response of the detector modules which have been hit by a traversing particle is simulated. Afterwards, a Monte Carlo event has exactly the same structure as a real event recorded during data taking and therefore the following processing and reconstruction steps can be performed with the same software and algorithm set-up.

3.4.3 Reconstruction of Physics Objects

The raw event data, either obtained from Monte Carlo simulation or the real data taking procedure, is stored in an EDM file. Physics analyses however rely on physics objects like charged particle tracks, energy deposits in the calorimetry or particle jets and therefore the raw data has to be processed and a reconstruction step has to be performed. Since CMSSW provides a multiplicity of different reconstruction algorithms for all kinds of physical objects, some standard reconstruction sequences have been defined which were run in the processing step directly after the digitisation. The sequences link all needed reconstruction steps together with the most common algorithms and their optimal settings and provide a common basis for most of the analyses performed at CMS. Due to the modularity of the framework, an exchange of single algorithms or the complete reconstruction sequence in a future reprocessing step is possible at any time. The following section shortly describes the reconstruction of some physics objects which are important for the analyses presented in this thesis:

- **Reconstruction of Charged Particle Tracks:**
  Charged particles traversing the silicon tracking system leave a trace in form of measurable hits. At first, the digitised hits are transformed into reconstructed hits in the local coordinate system of the silicon tracking device. Therefore a "Clustering" and "RecHitConversion" algorithm is used providing clusters of adjacent detector modules hit by a particle above a certain threshold and applying an error to the reconstructed hit position, respectively. Second step is the so-called "Seed Finding" where a starting point for a possible track candidate and its initial trajectory parameters are found. There are two kinds of seeding algorithms implemented in the software framework, the default one uses information of the silicon pixel and strip detector, whereas the other one only uses information given by the silicon strip modules. The seeding information is then passed to the "CombinatorialTrackFinder" (CTF) [32] in the first case or to the "RoadSearch" (RS) [33] algorithm in case of the pixel-less approach. This part of the track finding is called "Pattern Recognition" where the CTF iteratively proceeds from layer to layer of the silicon tracking device using Kalman Filter techniques [34] whereas the RS algorithm collects all measurements along so-called "roads" opened up around the seed trajectories. Both algorithms produce track candidates which are then transformed to particle tracks in a final fit and a subsequent smoothing procedure. More details about the track reconstruction efficiency and resolution in case of the CTF and the default settings for the "Patter Recognition" step are presented in appendix D.1.

- **Reconstruction of Particle Jets:**
  The constituents of hadrons, namely quarks and gluons, are called partons. In high energetic proton-proton collisions, the initial partons are separated which rises the interaction energy due to the confinement of colour charged particles. At some point, the production of quark-anti-quark pairs is favourable with respect to the energy balance and thus, streams of collimated hadrons are produced along the flight direction of the primary parton. These particle streams are called jets which are defined by their energy deposit E_{T} measured in the ECAL and HCAL as well as their momentum vector $\mathbf{p}_{\mathrm{T}}$, both in transverse direction. In order to measure the jet energy as well as the direction of the jet, all electromagnetic and hadronic calorimeter cells above a certain signal to noise ratio are grouped into so-called towers which act as basis for the different jet reconstruction algorithms provided by CMS [35]. In principal, there are two algorithm types included in CMSSW. Cone-based algorithms like the MidPoint [36], IterativeCone [37] and SISCone [38] group the calorimeter towers in a cone around a high energetic seed tower whereas clustering algorithms like the k_{T}-algorithm [39] group calorimeter objects based on their distance relative to each other in the four-momentum space, iteratively.

  The b-tagging studies presented in this thesis are based on jets reconstructed with the IterativeCone algorithm. It is well tested, offers a good reconstruction performance and is used as default jet reconstruction algorithm in many b-tagging related analyses at CMS. Thus, a comparison between the results of former analyses can be ensured.

The IterativeCone algorithm is described here briefly:

A list of calorimeter towers ordered by transverse energy acts as input for the IterativeCone algorithm. The tower with the highest E_{T} acts as seed and a cone with a configurable size of $\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}$ is cast around it. All remaining towers inside the cone are then taken to form a so-called "proto jet" and its energy and direction is determined. This direction is the basis to form a new "proto jet" and the algorithm iteratively continues until the energy and direction of the resulting jet does not change by more than 1% and $\Delta R < 0.01$, respectively. The jet is then added to a list of reconstructed jets and the used calorimeter towers are removed. This procedure is repeated, until all calorimeter towers above a certain energy threshold are assigned to one reconstructed particle jet.

Jet reconstruction algorithms should provide a so-called infra-red and collinear safe behaviour. Infra-red safe algorithms may not merge two stable cones close to each other into one single particle jet, if an additional low energetic particle e.g. from *pile-up* is found in between them. The IterativeCone and MidPoint algorithms in contrast to the SISCone algorithm are not infra-red safe at all. In addition, collinear radiation may not change the algorithm output as well. Two collinear particles produced due to a splitting in the parton shower share the energy which would be necessary for acting as a seed. A collinear safe algorithm would accept these split particles in order to reconstruct the jet.

- **Reconstruction of Muons:**
  In order to reconstruct muons at CMS, primarily a "standalone" approach is made only using information derived from the muon system itself. The trajectory seeding starts with track segments obtained by measuring the drift times of the traversing particles in the innermost muon chambers. A Kalman filter method subsequently uses additional measurements from proximate layers and continues to the outside of the detector, taking energy loss in the detector material, multiple scattering and the inhomogeneous magnetic field into account. The "standalone" muon tracks are then propagated back towards the interaction point. Another more time consuming but more accurate muon tracking procedure is implemented in the CMS software framework as well. Starting point for this "global" tracking approach are "standalone" muon tracks. In addition, the extrapolated muon trajectory towards the interaction point is then compared with hits in the silicon tracking device and a final track fit which combines the information obtained in the tracking detector and the muon system is made in order to produce so-called "global" muon tracks. More detailed information can be obtained from [40].

- **Reconstruction of Electrons:**
  The signature of an electron in the detector consists of a single track in the silicon tracking system and in addition an energy deposit in the electromagnetic calorimetry. Since electrons radiate a large fraction of their energy as bremsstrahlung due to the bending of the trajectory by the magnetic field, the energy deposit in the calorimeter

cells is spread in ϕ direction. Thus, a set of adjoining calorimeter clusters, called "supercluster", has to be identified [41], [42]. Starting from this "supercluster", two hits in the tracking device have to be found acting as seed for the electron trajectory. In case of electrons with high transverse momenta, a Kalman filter technique is used to obtain the electron tracks. For lower energetic electrons with 5 GeV/$c < p_T <$ 30 GeV/c, a more accurate and nonlinear filter approach which takes the energy loss of the electrons into account is implemented as well [43].

Chapter 4

The CMS Computing Model

The LHC experiments require enormous amounts of storage capacity and computing power during the operation of the LHC. In addition to the collected raw collision data, there are several millions of simulated Monte Carlo events and various reprocessed or skimmed datasets. The total amount of data adds up to more than 15 Petabytes per year which has to be stored and made accessible for further analyses. These requirements can not be afforded by single institutions or even single countries.

Therefore, an approach with distributed computing resources all around the world is used. It is called the Worldwide LHC Computing Grid (WLCG) [44]. The organisation of the WLCG is presented in the next section. It provides a hierarchical tier structure with one main Tier-0 centre based at CERN and several Tier-1, Tier-2 and Tier-3 sites connected to it. Section 4.2 presents the efforts which were made at the Tier-1 centre GridKa, located at the Forschungszentrum Karlsruhe, during the Computing, Software and Analysis Challenge 2006 (CSA06). In the last section of this chapter, improvements made due to this and succeeding challenges are summarised and the current status is depicted.

4.1 The Worldwide LHC Computing Grid

The Worldwide LHC Computing Grid project is developed at CERN. In cooperation with the EGEE project (Enabling Grids for E-sciencE) [45] , whose main goal is the provision of access for scientists to geographically distributed computing Grid services, large parts of their computing infrastructures are shared. The information presented in this chapter are obtained within the scope of the CMS experiment which was introduced above.

The CMS offline computing system has to provide the possibility to store, to transfer and to process the recorded data during the whole lifetime of the experiment. Therefore, a modular and thus easy extendable software and hardware structure has to be guaranteed

which could keep up with the increasing amount of recorded physics collisions as well as with changing or evolving objectives presented by the LHC.

4.1.1 WLCG Tier Structure

The CMS computing model [13] is based on the hardware infrastructure and various Grid related services offered by the WLCG and EGEE project. It is based on four hierarchically-ordered layers of participating Grid sites distributed around the globe. Figure 4.1 schematically presents the tier structure of the WLCG.

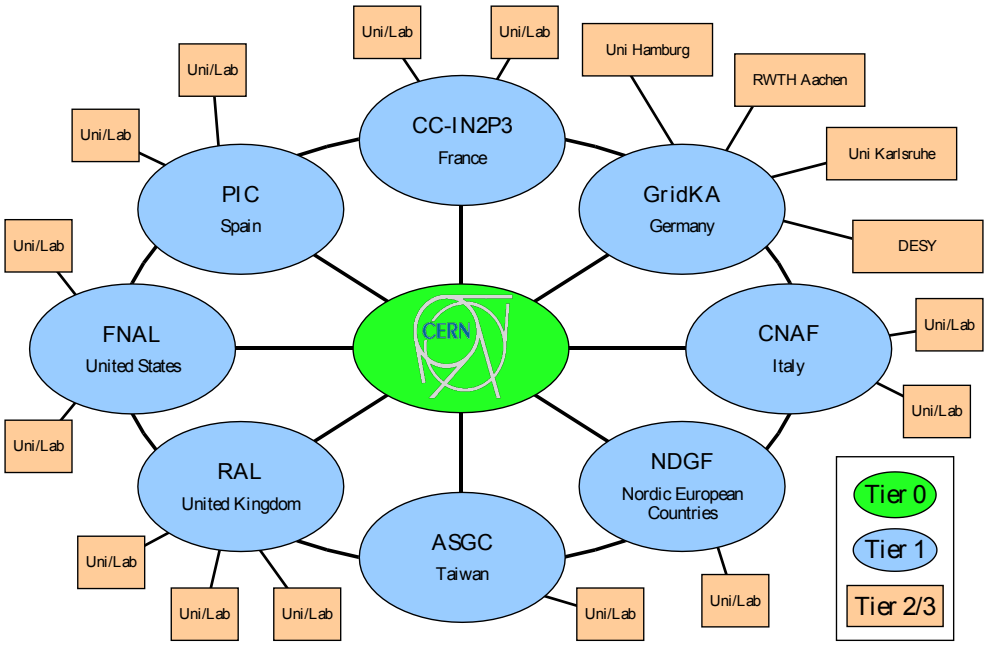


Figure 4.1: Schematic overview of the WLCG/CMS tier structure including the central Tier-0 at CERN (green), the eight Tier-1 centres (blue) and several corresponding Tier-2 and Tier-3 sites (salmon)

Between data taking at the CMS detector and analyses preformed by physicists around the world, several steps have to be passed in order to compete for new discoveries possibly being made at the LHC.

- **Tier-0:**
  The CMS central Grid centre, called Tier-0, is located at CERN. It collects all raw collision data which was selected by the trigger system of CMS and performs a first reconstruction of desired physics objects as described in chapter 3.4.3. The reconstruction is done at a rate of up to 150 events per second and the processed

data is permanently stored on a magnetic tape system as well as temporarily on disk-pools for further distribution. The storage system at CERN is called CASTOR (CERN Advanced STORage Manager) [46].

- **Tier-1:**
  A second copy of all the RAW[1] and RECO[2] datasets stored at CERN is distributed among eight Tier-1 centres in Europe, Northern America and Asia. The Tier-1 centres provide fewer computing resources and storage capacities than the Tier-0, according to the Memorandum of Understanding (MoU) [47]. Thus, only a selection of several data samples are stored permanently on the affiliated tape systems for further reprocessing. The Tier-1s perform a re-reconstruction of already existing datasets with e.g. newer software versions due to improved or exchanged reconstruction algorithms as well as calibration and alignment jobs in order to improve the accuracy of the different CMS subdetector components. Another task which is done at the Tier-1 sites is the production of smaller data samples which match certain selection criteria, called skims.

- **Tier-2:**
  Tier-2 centres are hosted at several CMS institutes all around the world. A typical Tier-2 divides its resources between local institute members and the CMS community as a whole. In general, they provide storage capacities for some smaller skims or AOD[3] datasets used by physicists for their final analyses on a flexible and easy to access computing farm. Like in case of the Tier-1s, a Tier-2 has to support offline calibration and alignment tasks. Another duty is the production of Monte Carlo datasets and the subsequent transfer to their associated Tier-1 for long term storage.

- **Tier-3:**
  The Tier-3 sites are not explicitly stated in the CMS computing model. They provide resources for the local user community which are used basically for interactive analyses, software development, testing or the final interpretation of received analysis results. In addition, they offer access for the local users to the different Grid services described below.

4.1.2 The Grid Middleware

A computational Grid should provide common interfaces in order to access the different underlying computing hardware and storage structures of its various participating sites. There are quite homogeneous hardware set-ups mostly found at larger Grid sites, but in

[1] direct output from the CMS detector, e.g. electronic response of the various detector modules
[2] stored data of all reconstructed physics objects, e.g. charged particle tracks
[3] a subset of physics objects of a RECO dataset for specific analyses

the majority of the cases, a grown structure of different computing hardware and storage solutions can be found. Since it is not adequate for the physicists to rely on such a heterogeneous structure, a common software layer, called Grid middleware, is introduced. The Grid middleware provides access to all hardware related services on a Grid site with common tools in order to form a steady environment on all centres around the world. The WLCG and EGEE projects are both based on the gLite 3.1 middleware which is described in [48].

gLite is not capable of directly talking to the hardware offered at the various sites. It is designed as a layer on top of the operating system of the different machines. Currently, the only supported operating systems are Scientific Linux 3 (32bit architecture) and Scientific Linux 4 (32 and 64bit architecture) [49]. However, it may run on other systems as well if some manual work is applied.

4.1.3 WLCG Services

The Grid middleware provides the infrastructure to access the distributed hardware components in an homogeneous environment. In order to get access to, manipulate, reprocess or analyse the huge amount of data collected by CMS, this infrastructure has to provide several services which are essential for a proper operation of the Grid. Since the WLCG/EGEE Grid covers more than 200 participating sites, and therefore thousands of CPUs and Petabytes of storage capacity, an adequate monitoring of the current hardware and software status is indispensable. This role is challenged by the Grid Operations Centre (GOC) [50]. Each Grid user has the possibility to access the current status of the different Grid sites by checking their offered monitoring pages in realtime.

All participating sites allocate their computing resources to a huge user community and make them accessible all over the internet. Therefore, another main aspect of the WLCG is a high security scheme which is required to protect the various members, the computing hardware and the experiment's data from unauthorised access and thus potential damage and violation. The following list presents an overview of the main services offered by gLite and the WLCG:

- **Security Aspects:**
  The security concept of the WLCG is based on public key encryption, X.509 certificates and the Secure Sockets Layer (SSL) communication protocol. Each Grid user has to be a member of a so-called Virtual Organisation (VO). As the name suggests, VOs are virtual, abstract entities which are used to group users, institutions or even resources into a distinct domain with well defined access rights, e.g. users who belong to the CMS experiment are a member of a VO called "CMS" and therefore have access to all CMS related services in the Grid. In addition, there are more specialised and privileged VOs for e.g. administrative applications. All Grid users are identified by their Grid certificate which is used for authentication or accounting purposes.

- **User Interface:**
  The User Interface (UI) is the basic access point in order to reach the different services available in the Grid. A UI is typically installed on computers where the various Grid users have local login accounts and their Grid certificates are installed. These machines provide numerous command line tools in order to perform all needed Grid operations, including submission of jobs to remote grid sites for execution, job management issues like output retrieval, monitoring or troubleshooting, access to Grid information services or data management.

- **Computing Element:**
  The Computing Element (CE) is a sort of a common gateway to various kinds of computing resources provided by a participating Grid site. In general, it provides access to a local cluster or batch system, a collection of various worker nodes, on which user jobs were executed according to some load sharing, job management and queueing system like PBS, LSF or Condor, to name just a few. The worker nodes have basically the same software environment installed as the UI in order to allow access to the different Grid services by the user jobs themselves as well.

- **Storage Element:**
  The Storage Element (SE) offers uniform access for the Grid users to the underlying storage systems of an individual Grid site. The SE has the ability to control simple hard-disk based storage solutions as well as large disk arrays or mass storage tape systems. The storage systems are controlled by the Storage Resource Manager (SRM) service which offers various interfaces to common used storage solutions present at High Energy Physics related Grid centres like the Disk Pool Manager (DPM), CASTOR or dCache [51].

- **Information Service:**
  The Information Service (IS) collects data about the resources provided by the different participating Grid sites. In addition, it supervises the status of each Grid resource and therefore can be used for monitoring and accounting purposes. The IS plays an important role in the discovery of free disk space or worker nodes within the WLCG. It consists of two different components, the Globus Monitoring and Discovery Service (MDS) [52] and the Relational Grid Monitoring Architecture (R-GMA) [53] which are used for resource discovery and status publication as well as accounting and monitoring, respectively.

- **Workload Management:**
  A Workload Management System (WMS) receives users jobs, assigns them to an appropriate CE, tracks the current job status and retrieves the job output from the executing site. The WMS is run on a machine called Resource Broker (RB). Each Grid job is described by using the Job Description Language (JDL). The JDL configuration file carries information about e.g. needed resources or a required software component which has to be installed at the remote site. The RB discovers an

adequate CE which satisfies all the requirements requested in the JDL and finally submits the job for execution.

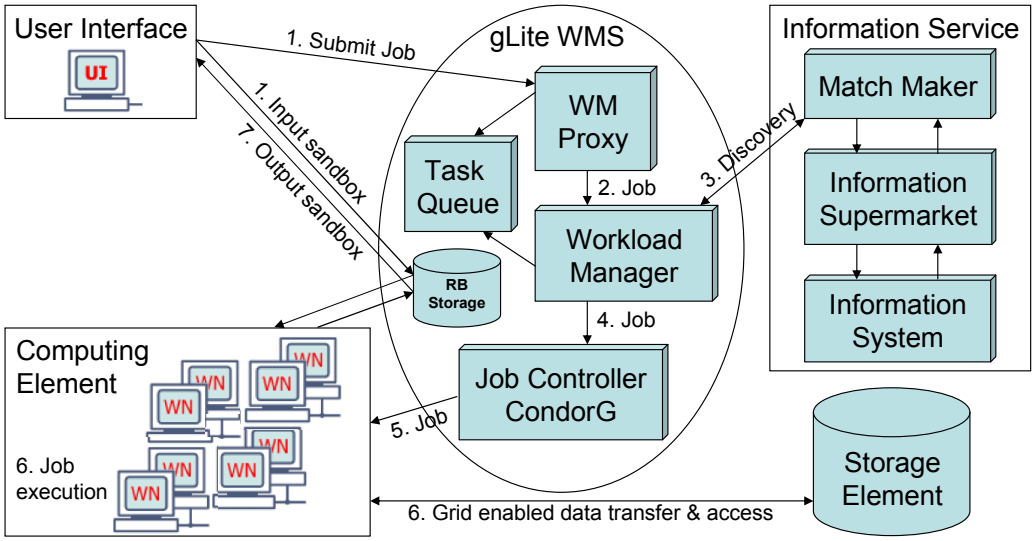


Figure 4.2: Schematic workflow of a typical WLCG Grid job

Figure 4.2 schematically shows the typical workflow of a Grid job from its submission at the UI up to the execution at a Grid site and the subsequent retrieval of the job output by the user. Each job runs through the following steps:

1. Submission from the user interface to the gLite WMS. At the same time, various input files (input sanbox), e.g. software configuration files or specific user code, are copied to the Resource Broker storage system.

2. The user job is then passed to the Workload Manager itself.

3. The WM starts to discover appropriate sites depending on the requirements described in the JDL. The best matching CE is selected for further processing of the job.

4. The Job Controller supervises the assignment of jobs to matching CEs, depending on e.g. free resources on the different remote sites.

5. Finally, the job is sent to the selected Computing Element for execution on one of the attached worker nodes.

6. While running, the job has the possibility to read or write data to Grid Storage Elements in order to access a larger storage space as the Resource Broker and therefore the size-limited input sandbox could provide.

7. The job output (output sandbox) is copied back to the RB storage until it is retrieved by the user to a UI.

4.2 Operation of the Tier1 Centre GridKa

In preparation for the enormous amount of data which are expected to arise during operation of the LHC, several so-called service challenges were performed during the last years. Goal of these challenges was to test the functionality of the whole processing chain which will be run after data acquisition at CMS as well as of all involved software components. The following steps have been performed during these tests: data processing at the CERN Tier-0, data distribution to the Tier-1 and Tier-2 sites, organised re-processing of stored RECO datasets as well as various physics analyses by single users.

4.2.1 Computing, Software and Analysis Challenge 2006

During autumn 2006, CMS performed one of these mentioned service challenges for a period of four weeks. The following section summarises the efforts which were made during this Computing, Software and Analysis Challenge (CSA06). The main focus, however, is on the participating German CMS Grid sites at DESY Hamburg, the university of Aachen and especially GridKa, located at Forschungszentrum Karlsruhe. The information presented in this chapter is summarised in [54] published at the German eScience Conference 2007. More details beyond what is given here can be found in the CSA06 Summary Report [55].

CMS Production and Grid Tools

The following Grid and CMS production tools were used during CSA06 as well as in subsequent challenges described later in this chapter.

- **Monte Carlo Production Agent:**
  The CMS Monte Carlo Production Agent (ProdAgent) [56] is a tool to supervise a large-scale and Grid-based generation and simulation of Monte Carlo events. The purpose of this tool is, e.g. to submit jobs to remote sites, to verify the job output or to merge and store the retrieved output files on an appropriate storage space.

- **CMS Remote Analysis Builder:**
  The CMS Remote Analysis Builder (CRAB) [57] is a tool for physicists in order to access and analyse remotely stored data via the Grid infrastructure. The CMS data management services DBS and DLS are queried in order to discover adequate remote sites where a desired dataset is accessible. CRAB automatically populates the job

input sandbox with user code, e.g. additional C++ libraries or configuration files which were required for a CMSSW analysis job. Finally, it takes care about the complete interaction with the gLite middleware starting from job submission, job splitting, job status monitoring up to job output retrieval.

- **Dataset Bookkeeping System:**
  Each CMS dataset is divided up into several data blocks and each of these blocks again into several hundreds of files. The Dataset Bookkeeping System (DBS) [58] ensures a proper mapping between files, blocks and datasets and provides in addition some processing related information. These are for example the number of events or the size of a single file, the data block or the complete dataset. Each file is represented by a Logical File Name (LFN).

- **Data Location Service:**
  The Data Location Service (DLS) [59] is used to relate a data block and therefore all included files to Grid sites which are physically hosting them. The DLS is based on the Logical File Catalog (LFC) infrastructure developed by the EGEE project.

- **Trivial File Catalog:**
  DBS and DLS provide the location of data blocks in the Grid and the LFNs of all associated files, respectively. The Trivial File Catalog (TFC) is responsible for mapping the logical filename to a physical filename which corresponds to a unique location on each sites storage system. The TFC is a xml-based mapping file which has to be installed on each tier site individually matching the set-up of the local storage system.

- **Physics Experiment Data Export:**
  The CMS data placement tool PhEDEx (Physics Experiment Data Export) [60] is used for the data transfer between different Grid sites. So-called agents, modular software components, have to be installed and run on the various Grid sites according to, e.g. the hardware set-up of the local storage system or the type of tier site they are running at. In addition, several tools for monitoring and management tasks are available.

  - The Transfer Management Database (TMDB) stores all transfer related information. These are for example requests for data transfers to remote sites, available data on a remote site or the status of current or completed file transfers.
  - Transfer agents are responsible for the movement of files between sites as well as for the migration of these files to local mass storage systems.
  - Management agents, for example, assign files to a remote site for transfer according to data subscriptions stored in the TMDB or provide information about the file transfer routing topology used by the File Transfer Service (FTS).
  - Management tools are available for all kinds of administrative purposes.

  - Local agents are used for registering files in file catalogues or the TMDB after a transfer was successfully finished.
  - Several web tools provide monitoring information about e.g. file transfers or the current agent status.

Motivation and Goals

The CSA06 was planned to test the computing, storage, software and network capabilities at a rate of 25% with respect to the requirements for the LHC startup in 2008. In order to consider the CSA06 as a success, pre-defined metrics had to be fulfilled by the participating tier sites. The following list presents the goals of CSA06 (parenthesised numbers indicate a threshold for being considered as a success):

- Automatic data transfer from the Tier-0 to all participating Tier-1 centres as well as subsequently to the Tier-2s via PhEDEx.
- Read and write access to the offline database at CERN via a database caching service Frontier [61] and Squid [62].
- Running of re-reconstruction jobs based on updated database information, e.g. new calibrations or alignment constants.
- Submission of user jobs via CRAB and the dataset bookkeeping and location services DBS and DLS.
- Number of participating Tier-1 centres: 7 (5) with an uptime of more than 90%. The Tier-1 centre NDGF was not participating in this challenge.
- Number of participating Tier-2 centres: 20 (15)
- Weeks of running at the required rates: 4 (2).
- Tier-0 efficiency during the best two weeks: 80% (30%).
- Number of running Grid jobs per day: 60 000 (30 000). Each job runs about 2 hours on average and the numbers are accumulated for Tier-1 and Tier-2 sites.
- Grid job efficiency: 90% (70%).
- Data serving capability for the jobs per execution slot: 1 MB/sec. This results in about 400 MB/sec and 100 MB/sec in case of a Tier-1 and Tier-2, respectively.
- Data transfer rates from the Tier-0 to the Tier-1, depending on the individual Tier-1: 10 - 50 MB/sec (5 - 25 MB/sec).

- Data transfer rates to the Tier-2s: 20 MB/sec (5 MB/sec).

A summary of the nominal and achieved goals of CSA06 as well as of subsequent challenges performed in 2007 and 2008 is presented in figure 4.9 at the end of this chapter.

Participating Sites and Supplied Resources

The core centre of the CMS computing model, the Tier-0 based at CERN, provided two CASTOR mass storage pools with 64 TB and 155 TB of disk space for incoming and outgoing file transfers, respectively. 700 dual-CPU worker nodes with a system clock of 2.8 GHz and 2 GB of memory provided the required computing power for processing the Monte Carlo events at a rate of about 50 Hz during the initial phase. This rate was increased to 160 Hz towards the end of the exercise.

The 7 participating Tier-1 sites provided about 25% of the resources described in the MoU. In order to succeed in CSA06, an estimate for about 300 CPUs and 70 TB of storage capacity was given in case of a nominal Tier-1. However, these values slightly varied depending on the individual Tier-1. The number of Tier-2 sites which participated in CSA06 amounts to 27. The estimated requirements were about 100 CPUs and 10 TB of disk space. In contradiction to the Tier-1s, the variance between the Tier-2 sites was rather wide.

The German Tier-1 centre GridKa provided the requested 70 TB of storage space and about 250 CPUs during CSA06. Four dCache pool nodes were dedicated to CMS. In order to achieve the required transfer rates, fast wide area network connections have been established between the different participating sites. The network connection between GridKa and CERN consisted of a 10 GBit link via optical fibre. GridKa and the other Tier-1 or Tier-2 sites were connected via the standard Internet. The PhEDEx software and the required agents were installed and executed on a dedicated machine, called VO-box.

The German Tier-2 centre is a federation of Grid installations at DESY Hamburg and RWTH Aachen. DESY provided about 100 CPUs and 35 TB of dCache storage space, whereas Aachen contributes with another 90 CPUs and 5 TB of disk space.

Accomplishment

In preparation for CSA06, more than 67 million Monte Carlo events had to be simulated beforehand on several Tier-1 and Tier-2 sites. The samples were then transferred to the Tier-0 at CERN to act as input for the reconstruction procedure performed during CSA06.

- **Data Processing:**
  The complete processing chain, including prompt reconstruction of collision events

at the Tier-0, creation of various skims for further alignment, calibration or analysis jobs, file merging to larger output files, injection into the Dataset Bookkeeping System and the PhEDEx TMDB and access to the alignment and calibration database via Frontier, was exercised during CSA06. The Tier-0 ran almost unattended and stably during the four weeks of the challenge. However, only due to scheduled manual operator interventions like fixing trivial bugs in the Tier-0 framework and the offline database or changing the reconstruction software version, all proposed metrics of CSA06 could be achieved.

- **Data Distribution:**
  During CSA06, the data export to the Tier-1 and Tier-2 centres was expected at transfer rates of about 25% of the requirements during data taking in 2008. The target and threshold values as well as achieved transfer rates from CERN to the Tier-1 sites are summarised in table 4.1. In addition, network speed and provided storage capacity are given, too. It can be seen that all sites exceeded the transfer rate goals and thresholds if the 15-day average is concerned. However, the threshold values were reached during the whole period of operation.

  The cumulative amount of data which was transferred to all participating sites during the four weeks of CSA06 adds up to more than 1 Petabyte and is shown in figure 4.3 depending on each destination site.

Table 4.1: Intended and achieved transfer rates during CSA06 from the Tier-0, CERN to all participating Tier-1 centres. Given are a 15- and 30-day average as well as the provided network and storage capacity at the different sites. The parenthesised values indicate installed components after the begin of the challenge

Tier1 Site	Goal [MB/sec]	Threshold [MB/sec]	30-day avg. [MB/sec]	15-day avg. [MB/sec]	Storage [TB]	Network [Gb/sec]
ASGC	15	7.5	17	23	48+(36)	2
CNAF	25	12.5	26	37	40	10+(10)
FNAL	50	25	68	98	700	11
GridKa	25	12.5	23	28	40+(30)	10
IN2P3	25	12.5	23	34	70	10
PIC	10	5	22	22	50	1
RAL	10	5	23	33	60+(20)	11

- **Tier-0 to Tier-1 Operations:**
  At the beginning of CSA06, the transfer rate between CERN and the seven participating Tier-1 centres could not reach the desired rate of 150 MB/sec due to a

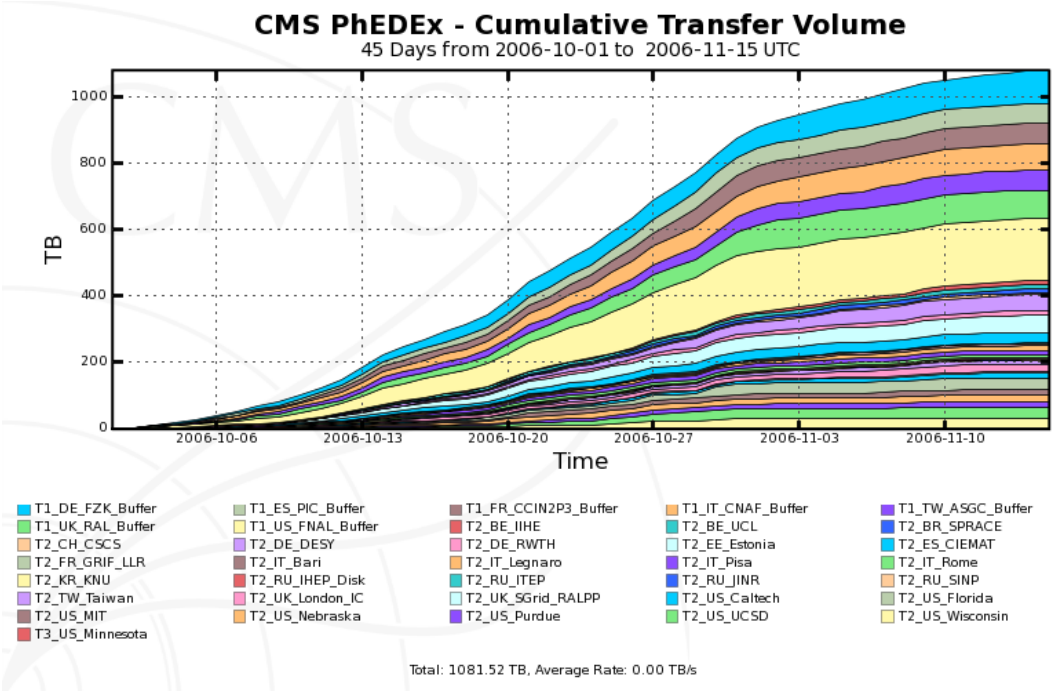


Figure 4.3: Cumulative amount of data transferred during the four weeks of CSA06 operation. Transfers from Tier-0 to Tier-1, Tier-1 to Tier-1, Tier-1 to Tier-2 and Tier-2 to Tier-1 are taken into account

relatively low reconstruction rate of about 50 Hz. Towards the end of the challenge, the reconstruction rate was successively increased to about 160 Hz. Thus, the outgoing data rates were increased as well and could easily reach 300 MB/sec. The transfer rates between CERN and the Tier-1 centres are given in figure 4.4. The large fluctuations were a result of problems with the offline database at CERN. After manual intervention, the accumulated data was dissipated and transferred to the Tier-1s within a few days running at more than 350 MB/sec. All shown transfer rates are averaged over 24 hours of operation. Thus, hourly transfer rates even exceeded 650 MB/sec.

Another important issue of CSA06 was to monitor the status of the file transfers between the different sites. Most of the Tier-1 centres performed at a very high level of transfer quality as shown in figure 4.5. The transfer quality is measured as the number of successfully transferred files with respect to all transfer attempts within a certain period of time. There are several reasons for a file transfer to abort, e.g. problems with the export or import of data due to failures of the software or hardware of the mass storage systems at the different Grid sites, connection or routing problems at network level, problems with central database systems at CERN or problems due to a overload of the VO-box hardware or the various PhEDEx agents.

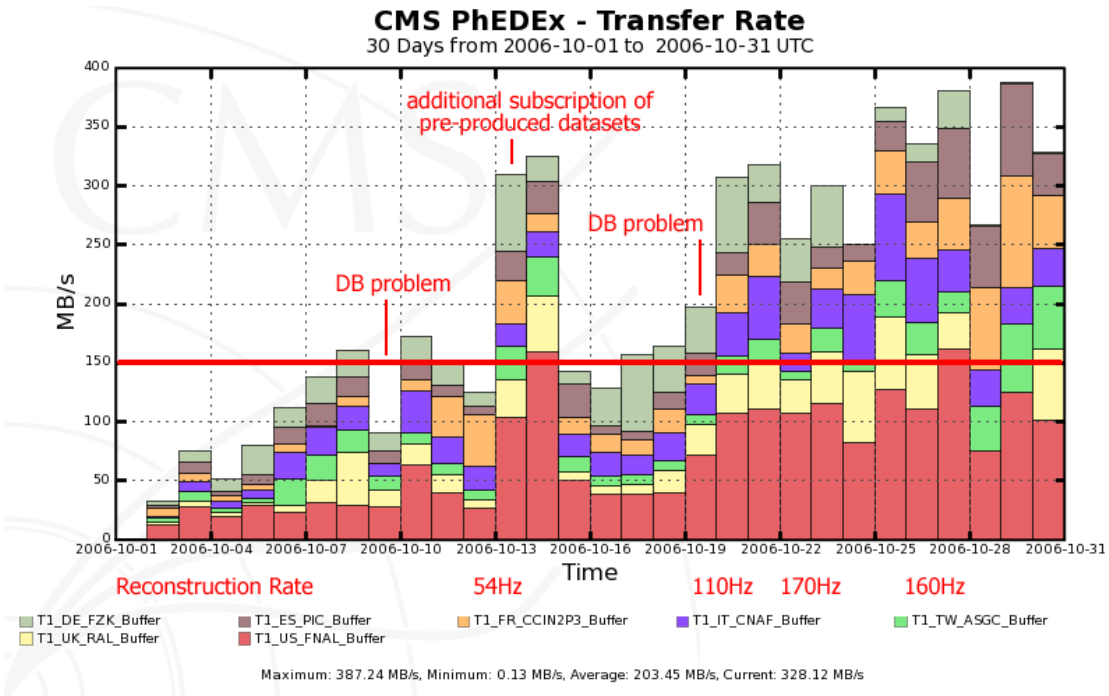


Figure 4.4: Data transfer rates between CERN and the seven participating Tier-1 sites during 30 days of CSA06 operation. The red line indicates the target rate. The prompt reconstruction rate at the Tier-0 is denoted at the bottom of the plot

In case of GridKa, severe problems with the initial version of the dCache storage software, which required frequent manual interventions, made a scheduled software update inescapable. The subsequent reconfiguration and tuning of the storage system led to large inefficiencies during the last week of CSA06. Anyway, GridKa was able to successfully finish the challenge with only a few days of delay. The achieved data transfer rates exceeded 400 MB/sec for more than 8 hours.

- **Tier-1 to Tier-2 Operations:**
  Figures 4.6 and 4.7 summarise the achieved transfer rate and transfer quality between the Tier-1s and all participating Tier-2 sites, respectively. The goals of CSA06 were reached by 20 out of the 27 participating Tier-2 centres whereas one of the remaining sites could just achieve the thresholds. Transfers to Tier-2 sites are expected to occur in bursts, since subscribed data blocks are only transferred as a whole and not file-based, thus the transfer quality and data rate is expected to be not that continuous as in case of the Tier-1s.

  Figure 4.8 presents the transfer quality to the federated German Tier-2 centre at DESY and Aachen and an example of transfer rates to the Tier-2 centre DESY. In comparison to many other participating Tier-2 sites they performed extremely well during the whole period of CSA06 with an uptime of almost 100%. The goal transfer

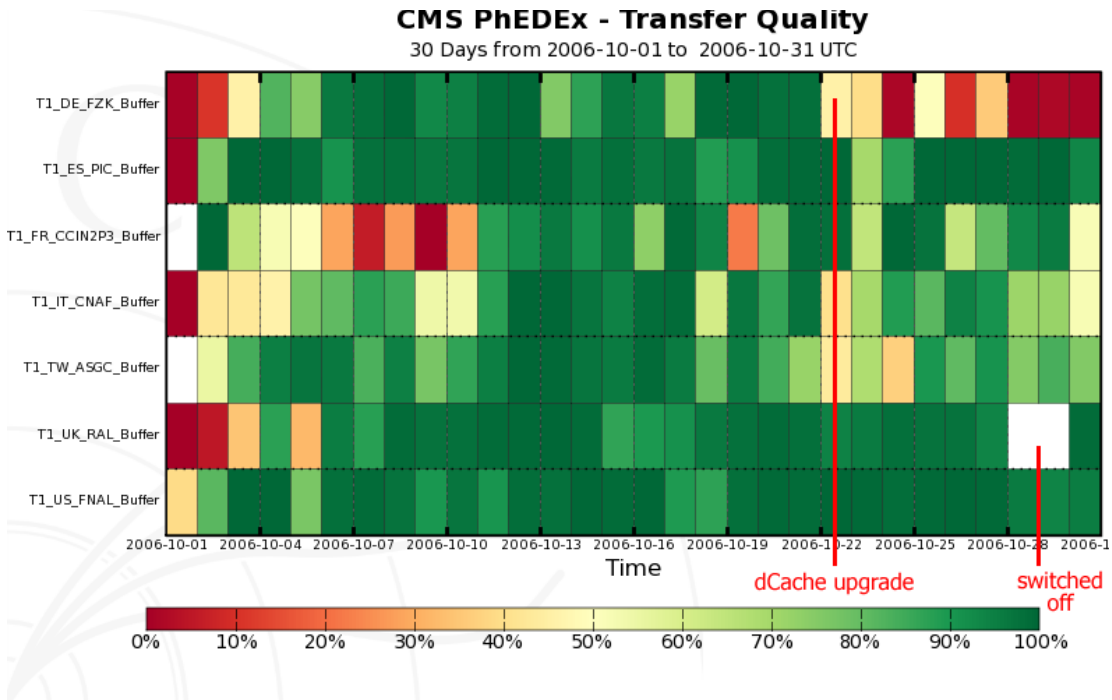


Figure 4.5: Data transfer quality between CERN and the seven participating Tier-1 sites during 30 days of CSA06 operation

rates of about 20 MB/sec for a nominal Tier-2 were occasionally exceeded by more than a factor of five.

- **Data Re-Processing and Analysis:**
  Another goal was the re-reconstruction of existing datasets with new calibration constants derived from the central offline database at CERN. Therefore, CMSSW jobs for about 100 000 events were intend to be performed at each Tier-1. However, due to problems in the reconstruction software and with the size of the input files, not all Tier-1 centres achieved this goal in the first place. After skimming the input to smaller files and running the reconstruction sequence on these smaller samples, all Tier-1 centres succeeded in this exercise.

  The data which was distributed during CSA06 was not only meant for pure testing purpose, but also for various physics analyses of individual users which were performed by using the Grid infrastructure. In order to generate a constant number of Grid analysis jobs during CSA06, a so-called "job robot" was used in order to submit jobs at a constant level to the participating sites. It was discovered, that the performance of the Workload Management System was not sufficient to pass enough jobs to available worker nodes in order to keep all CPUs busy.

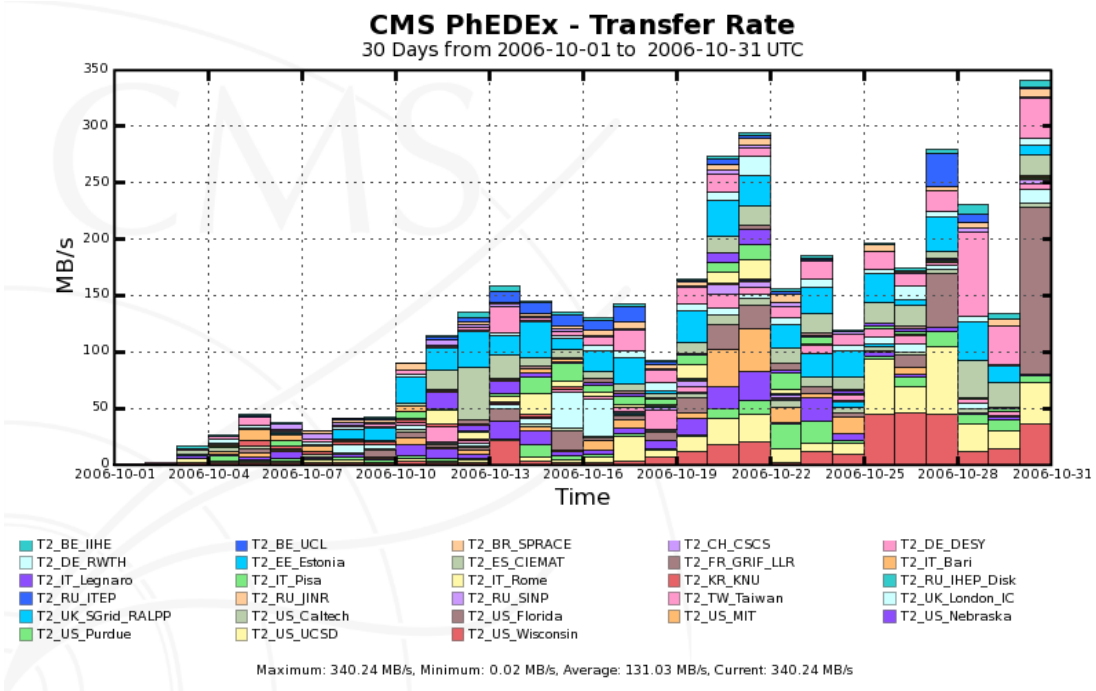


Figure 4.6: Data transfer rate between the Tier-1s and all participating Tier-2 sites during 30 days of CSA06 operation

4.2.2 Lessons Learned and Current Status

The Computing, Software and Analysis Challange performed in 2006 demonstrated the base functionality of the WLCG Grid infrastructure and services with respect to the CMS experiment at a level of 25% of the requirements for the LHC startup in summer 2008. In addition to the reconstruction jobs at CERN, the CMS offline software framework was tested by Grid-based re-reconstruction and analysis jobs run remotely at the different sites as well. It was observed that especially the Grid storage elements and Resource Brokers showed some fragility under heavy load. Many problems could only be solved by manual interventions and substantial efforts on the operational side. Due to CSA06, performance bottlenecks and instabilities of the Grid infrastructure were identified in order to be solved during one of the future challenges. Therefore, another two Computing, Software and Analysis challenges were performed in the following two years, 2007 and 2008, at levels of 50% and 100% of the needs for LHC startup, respectively. The intended and achieved goals of CSA06, CSA07 and CSA08 are summarised in figure 4.9.

Although CSA07 was estimated to test the functionality of the whole reconstruction, data distribution, re-reconstruction and analysis chain of the CMS experiment in the first place, it was more or less focused on testing the Tier-0 stability at the higher rates in the long run. The Tier-1 centres were only served in bursts and therefore most sites could not reach

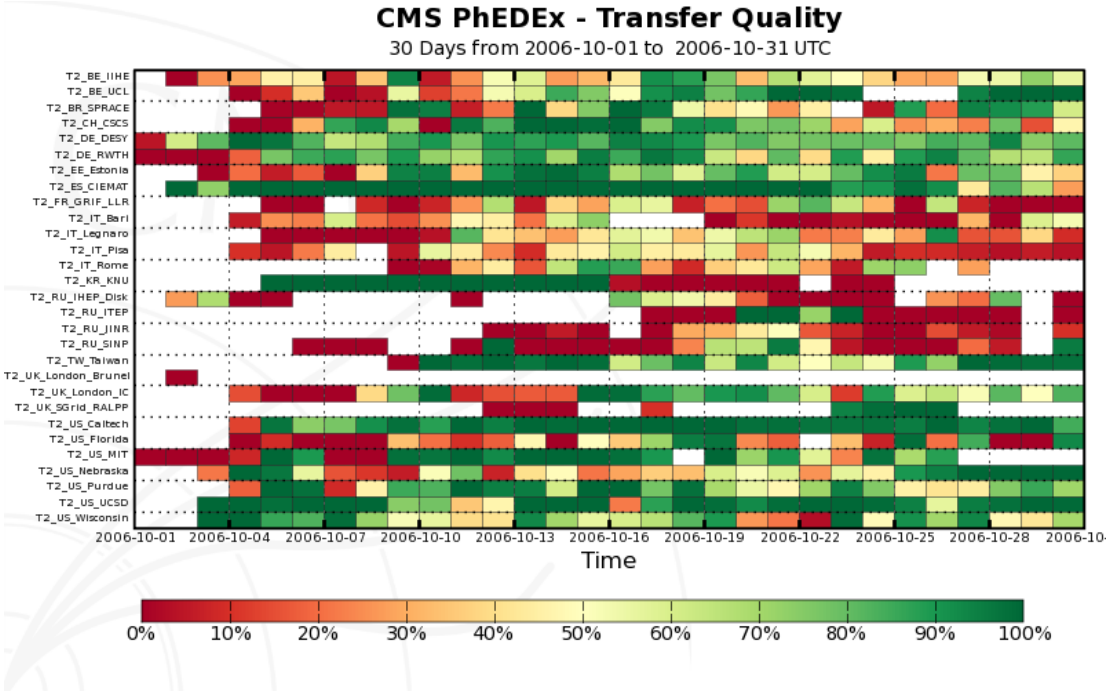


Figure 4.7: Data transfer quality between all participating Tier-1 and Tier-2 sites during 30 days of CSA06 operation

the pre-defined metrics during the running period of the exercise.

For CSA08, GridKa allocated about 300 TB of disk space for performing the arranged tests. The identified storage problems during former tests led to further improvements of the software set-up and tuning parameters of the involved dCache storage system in order to optimise software and hardware stability under heavy load. CSA08, in contradiction to CSA06 and CSA07, was a multi-VO test where the data distribution and the processing chain of all four main LHC experiments was exercised at the same time. Figure 4.10 presents incoming and outgoing transfer rates depending on the different experiments, as well as the CMS data rate for read and write access to and from the tape storage system at GridKa. It can be seen that the dCache disk and tape systems operated at very high rates during CSA08, even under heavy load. However, the activities of the other VOs were rather low during this challenge. Therefore, another exercise will be performed within the next few months, where larger data rates of the other experiments are expected as well.

In summary, the performed service challenges greatly helped to improve the overall stability of the involved WLCG Grid infrastructure as well as the experiment's offline software framework. From today's point of view, the approaching startup of the LHC should not yield any problems with respect to data taking and data distribution for the CMS experiment.

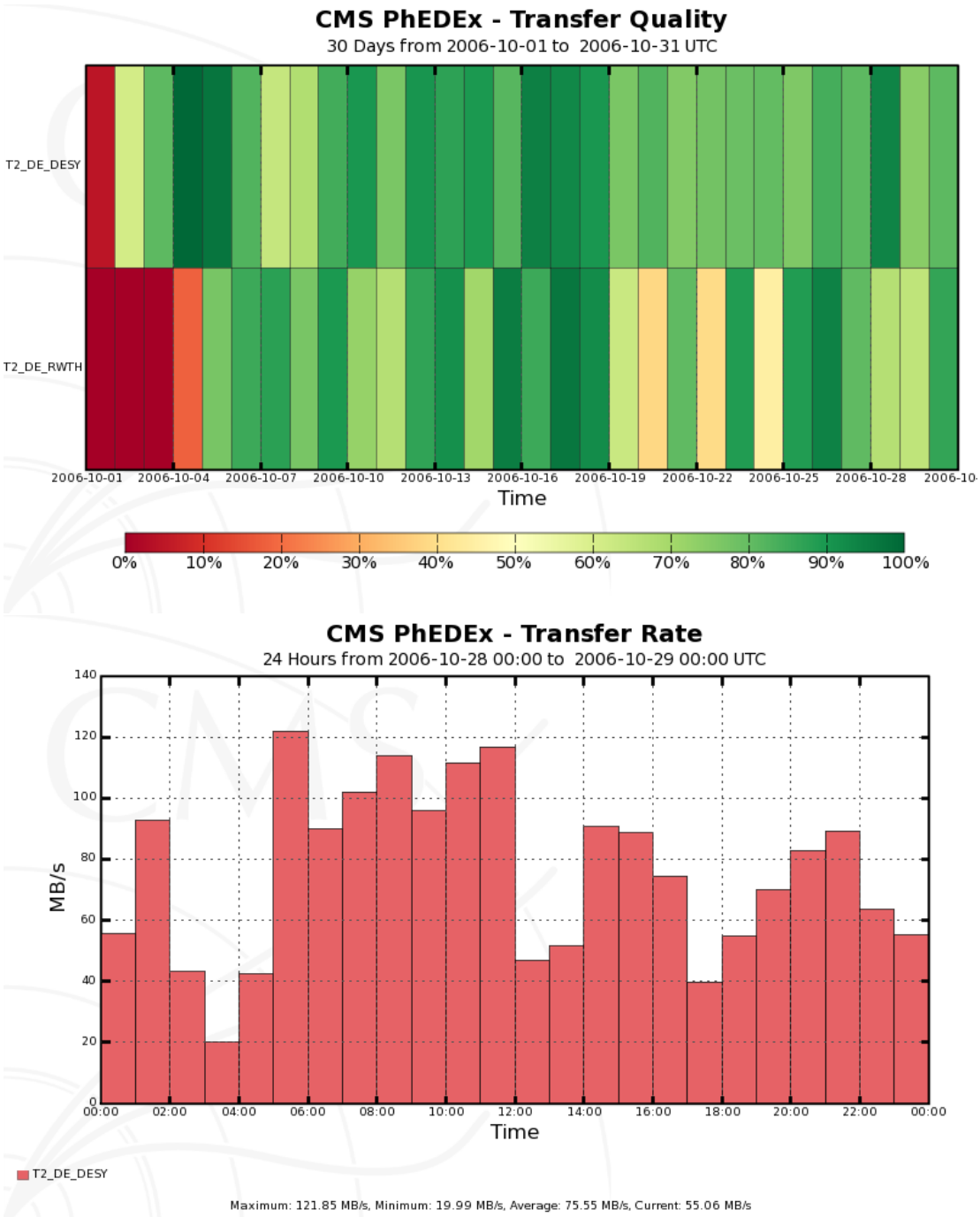


Figure 4.8: Data transfer quality for the federated German Tier-2 at DESY and Aachen during 30 days of CSA06 operation (top). Example transfer rate to the Tier-2 centre DESY (bottom)

Service	Goal 2008	Status 2008	Goal 2007	Status 2007	Goal 2006	Status 2006
Tier-0 Reco Rate	150-300 Hz	Achieved	100 Hz	Only at bursts	50 Hz	Achieved
Tier-0 → Tier-1 Transfer Rate	600 MB/sec	Achieved partially	300 MB/sec	Only at bursts	150 MB/sec	Achieved
Tier-1 → Tier-2 Transfer Rate	50-500 MB/sec	Achieved	20-200 MB/sec	Achieved partially	10-100 MB/sec	Achieved
Tier-1 → Tier-1 Transfer Rate	100 MB/sec	Achieved	50 MB/sec	Achieved partially	N/A	-
Tier-1 Job Submission	50 000 jobs/day	Achieved	25 000 jobs/day	Achieved	12 000 jobs/day	3 000 jobs/day
Tier-2 Job Submission	150 000 jobs/day	Achieved	75 000 jobs/day	20 000 jobs/day	48 000 jobs/day	Achieved
Monte Carlo Simulation	1.5×10^9 events/year	Achieved	50×10^6 events/month	Achieved	N/A	-

Figure 4.9: Overview of the intended and achieved goals of CSA06 and its two successors performed in 2007 and 2008

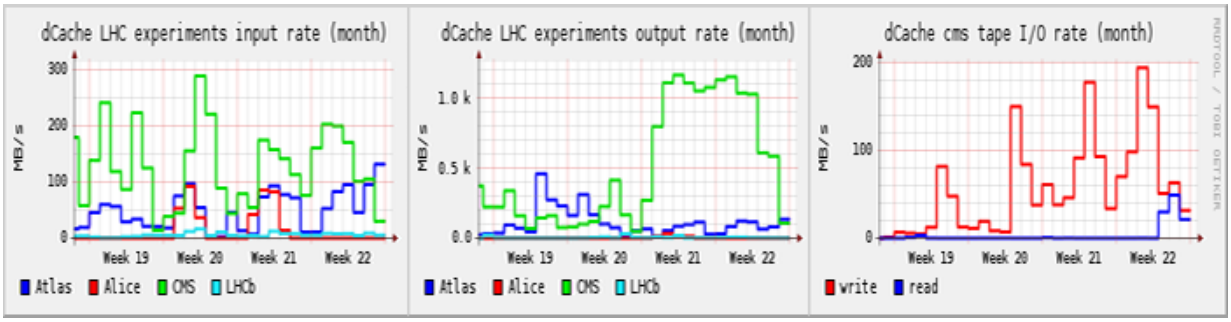


Figure 4.10: dCache disk import (left) and export (middle) transfer rates during CAS08 as well as the data rates for read and write access to and from the dCache tape system (right) at GridKa

Chapter 5

b-Quark Jet Identification and Alignment Studies

A proper identification of b-quark jets (b-tagging) is a crucial part for many physics analyses performed at today's hadron colliders. As introduced in chapter 2, b-tagging plays an important role in reaching various physics goals of the LHC. Among them are searches for the Higgs boson, top physics or discoveries beyond the Standard Model, where jets originating from the hadronisation of b-quarks are present in the final state. Since b-tagging relies on several well measured lifetime and kinematical based properties of the production and the decay of b-hadrons, a good knowledge of the involved detector subsystems has to be ensured.

Precisely reconstructed tracks and secondary decay vertices are the basis for the various available lifetime-based b-tagging algorithms at CMS. Thus, their performance strongly depends on the spatial resolution provided by the silicon tracking system. Its resolution, in turn, directly depends on the alignment of the device. Alignment in this case means the knowledge of the location and orientation of each of the more than 16 000 modules and substructures of the silicon strip and pixel detector. During the initial operation phase of CMS, the assumed alignment precision is limited by the module mounting accuracy obtained from survey measurements during construction, measurements of the laser alignment system and muon tracks originating from cosmic rays. In order to improve the knowledge of the alignment position, a track based and thus data volume dependent alignment procedure has to be performed, starting with data taking in summer 2008.

This chapter presents b-tagging methods used at CMS and gives an overview of the expected performance of all algorithms in case of an ideally aligned detector. These studies have been performed for the first time at CMS within the new software environment CMSSW and thus present a basis for comparison with results obtained in former analyses published in the experiment's physics technical design report [12]. Furthermore, the algorithm performance is studied for a misaligned tracking detector corresponding to first

periods of LHC collisions at integrated luminosities of 10 and 100 pb^{-1}.

At the end of this chapter, the development of a simple and easy-to-understand b-tagging algorithm is presented which is now part of the official CMS software framework and integrated in the default reconstruction sequence during the official CMS reconstruction procedure. It is shown that this algorithm is the most insensitive to misalignment and its performance is nevertheless comparable with the one obtained from other b-tagging algorithms.

5.1 b-Quark Jet Identification

In order to distinguish b-jets from jets of other flavours, it is necessary to exploit several properties of the production and the weak decay of b-hadrons. The most significant property is the comparatively long lifetime of about 1.5 ps ($c\tau \approx 450$ µm) in case of a B^0-meson [3] which leads to a measurable flight distance and thus to a displaced secondary decay vertex (SV). The charged particle tracks associated to the SV are incompatible to the primary vertex (PV) of the proton-proton collision. A typical b-hadron decay is presented in figure 5.1.

The b-hadron decay leads to a secondary vertex in jet direction with tracks incompatible to the primary interaction point of the event. In the most cases a subsequent decay to a c-hadron, which forms a displaced tertiary vertex due to a non-negligible lifetime of about 1 ps ($c\tau = 312$ µm) for $D^{\pm}$-mesons [3], occurs. The lifetime of c-hadrons improves the identification of b-quark jets due to an increased number of tracks incompatible to the PV. On the other hand, collisions with c-jets in the final state have similar event topologies and therefore make the reduction of background events much more difficult due to a real separated vertex.

Jets of light flavour hadrons or gluons basically contain no displaced vertex and therefore a reduction of the light flavour background is much easier to achieve. Other properties which make b-hadrons distinguishable are their large mass and their large multiplicity of charged particle tracks in the final state. They decay to about 5 charged particles on average and carry a large fraction of the total jet energy due to the hard b-fragmentation function.

Another attribute which is very important for b-tagging is a property that can be determined for each particle track associated to the jet. This impact parameter (IP) is a measure for the displacement of the track with respect to the primary vertex of the event. It is measured between the point of closest approach of the extrapolated particle trajectory and the primary vertex. Tracks originating from a displaced vertex, and therefore from a b- or c-hadron decay, have measurable impact parameters whereas tracks of light flavour jets have no impact parameter at all. In order to reject e.g. mismeasured tracks, a so-called lifetime sign to the IP is introduced. In case the point of closest approach lies in jet direction, the impact parameters of the associated tracks are signed positively, otherwise

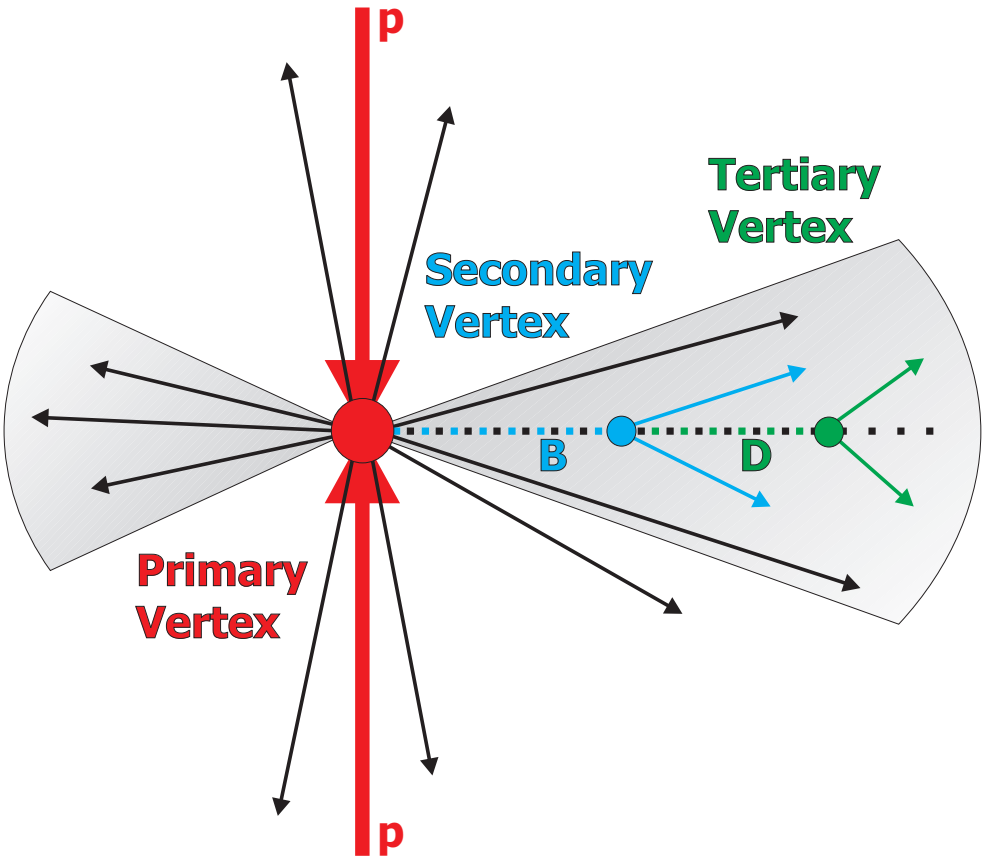


Figure 5.1: Schematic view of a typical b-hadron decay. The dotted blue line indicates the flight direction of the b-hadron with a displaced secondary vertex and a possible subsequent decay of a c-hadron forming a tertiary vertex inside the b-jet cone (dotted green line)

negatively. Figure 5.2 schematically gives an impression on the measurement of lifetime signed track impact parameters which can be calculated in the transverse plane or in three dimensions.

Another approach in order to identify b-jets benefits from the fact that about 30% of the b- and c-hadrons decay leptonically and therefore have electrons, muons or taus in the final state. This leads to about 20% of semi-leptonically decaying b-hadrons per lepton flavour, if cascade decays of the c-hadrons are taken into account as well. Since CMS has an excellent muon system and calorimetry, the identification of muons and electrons inside the jets is exploited as well, although the overall performance of the soft lepton tagging algorithms is much lower due to the semi-leptonic branching fraction.

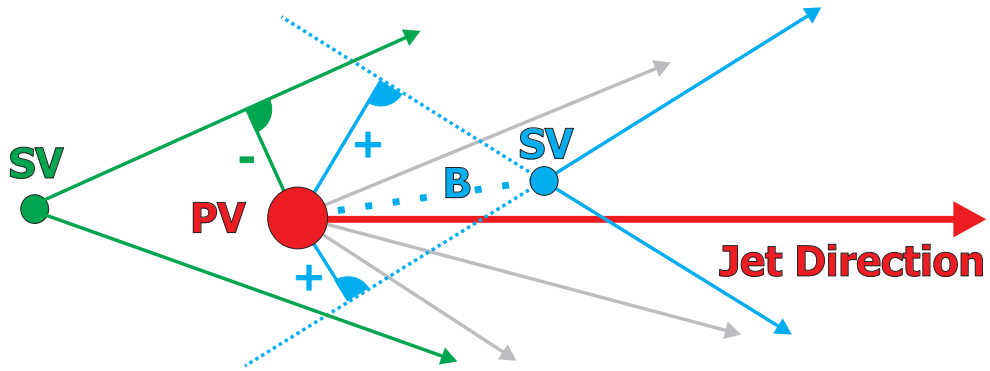


Figure 5.2: A positive signed impact parameter is assigned to tracks originating from a decay which lies in jet direction (blue) or a negative one otherwise (green)

5.2 b-Tagging at CMS

The CMS offline software framework provides several different b-tagging algorithms in a widely modular manner. Some of them rely on impact parameter based observables, whereas others combine them with secondary vertex properties or exploit the presence of leptons in the b-hadron decay chain. The results presented in this chapter are summarised in a CMS internal analysis note [63] and published in the corresponding physics analysis summary [64].

5.2.1 Available Algorithms

Two b-tagging algorithms rely solely on track impact parameters [65] from reconstructed charged particle tracks associated to decay products inside the jet cone, whereas another one is based on the reconstruction of a secondary decay vertex [66] in addition. Two muon and electron based algorithms [67] are implemented in CMSSW as well. The output of each b-tagging algorithm is given in form of a discriminating variable which is computed for each jet from input observables. A cut on the value of the discriminator is used to distinguish b-jets from jets of other quark flavours or gluons. An overview of the used input variables and the expected algorithm performance is presented later on in this chapter.

- **TrackCounting Algorithm:**
  The TrackCounting b-tagging algorithm is the most basic algorithm available at CMS. It relies only on the measurement of impact parameters for the selected tracks inside a jet cone. Tracks selected due to the quality criteria presented in the next

section are ordered by their signed impact parameter significance, defined as

$$IP_{sig} = IP_{sign} \cdot \frac{IP_{value}}{IP_{error}} \tag{5.1}$$

The discriminator is defined as the impact parameter significance of the n-th track, where n is an algorithm parameter which can be chosen according to desired performance needs. High values of n increase the purity of the selected jets, since more tracks with high impact parameters have to be present, whereas low values of n are a good choice in case of a high efficiency for tagging b-jets is needed. Two default TrackCounting algorithms are defined in CMSSW. One is called "high efficiency" and the other "high purity" algorithm, taking the second and the third track, respectively. The first track could be used to design a further algorithm which works best in the very high b-efficiency range close to 100%, but practically there is no use case, since the mistagging rate would be very high as well.

- **JetProbability Algorithm:**
  This algorithm is also based only on the measurement of track impact parameters. In contrast to the TrackCounting method, which just cuts on a single variable, a more sophisticated approach is used here. For each selected track, a "track probability" to originate from the primary vertex is computed and afterwards all tracks are combined to a total probability for the jet. Thus, a low probability to come from the PV is an indication for a b-jet. The JetProbability algorithm requires a training of the track probability distributions by introducing several track categories and their corresponding histogrammed resolution functions, $R(IP_{sig})$. The parameters that define the categories are: the number of hits in the pixel detector, $|\eta|$ of the track, p of the track and normalised χ^2 of the track. Based on these categories, the track probabilities $P_{tr}(IP_{sig})$ are computed and then combined to a probability P_{jet} for a jet with N tracks, following the three equations

$$P_{jet} = \Pi \cdot \sum_{j=0}^{N-1} \frac{(-\ln \Pi)^j}{j!} \tag{5.2}$$

,

$$\Pi = \prod_{i=1}^{N} \tilde{P}_{tr,i}(IP_{sig}) \tag{5.3}$$

and

$$P_{tr}(IP_{sig}) = sign(IP_{sig}) \cdot \int_{|IP_{sig}|}^{\infty} R(IP_{sig}) dIP_{sig} \tag{5.4}$$

$\tilde{P}_{tr,i}(IP_{sig})$ is the redefined track probability in order to get only positive values for the track probabilities: $\tilde{P}_{tr,i}(IP_{sig}) = P_{tr,i}(IP_{sig})/2$ for $P_{tr,i}(IP_{sig}) > 0$ and $\tilde{P}_{tr,i}(IP_{sig}) = 1 + P_{tr,i}(IP_{sig})/2$ for $P_{tr,i}(IP_{sig}) < 0$. The resulting discriminator value for this algorithm is defined as $-\log P_{jet}$ which is small in case of light quark jets and large for jets originating from the decay of b-hadrons.

- **CombinedSecondaryVertex Algorithm:**
  The CombinedSecondaryVertex algorithm combines impact parameter significances and different topological and kinematical variables derived from the reconstruction of a secondary decay vertex inside the examined jet. In order to reconstruct secondary vertices, a two step procedure is applied. First step is the vertex finding followed by the vertex fitting procedure. Two different algorithms namely the "Trimmed Kalman Vertex Finder" (TKVF) with the "Trimmed Kalman Fitter" [68] and the "Adaptive Vertex Finder" (AVF) with the "Adaptive Vertex Fitter" [69] are used in this thesis. The latter one was chosen as default for the studies presented here, but a comparison of these two and their impact on the b-tagging performance is discussed in 5.4.3. The selection cuts for the vertex reconstruction are presented in the next section.

  Based on the results of the secondary vertex finding, three different categories are defined:

  1. **RecoVertex:** At least one secondary vertex candidate was found, reconstructed and passed the SV selection criteria. The tracks associated to all thus found vertices are used for the computation of the vertex related variables if more than one accepted secondary vertex is present.
  2. **PseudoVertex:** If no reconstructed secondary vertex candidate can be found, a so-called PseudoVertex is created from the charged particle tracks which are not compatible with the primary vertex. There have to be at least two tracks with a signed impact parameter significance greater than 2 in the jet in order to form a PseudoVertex. Discriminating variables like the invariant mass can be determined, but no flight distance is computed since no actual vertex fit is performed.
  3. **NoVertex:** This category is selected, if neither a RecoVertex nor a PseudoVertex can be determined.

  The final discriminator of the algorithm is computed by combining all impact parameter significances of the accepted tracks with a set of variables obtained from the reconstruction of a secondary vertex. In case where a secondary vertex could be reconstructed, the following variables are used in addition. In the PseudoVertex or NoVertex case, only a subset of these observables are used.

  - The invariant mass of charged particles associated to the secondary vertex.
  - The track multiplicity of charged particles associated to the secondary vertex.

- The transverse flight distance significance: distance between PV and SV in the transverse plane divided by its error.
- The energy fraction at the secondary vertex: the energy of charged particles associated to the secondary vertex divided by the energy of all charged particles associated to the jet. Due to the hard b-fragmentation function, most of the jet energy is associated to the b-hadron.
- The rapidities of charged particle tracks; this variable is used n-times, where n is the secondary vertex track multiplicity.
- The track impact parameter significances of all tracks associated to the jet in the transverse plane.
- The track impact parameter significance of the first track exceeding a so-called charm threshold. All tracks are sorted by their signed impact parameter significances in decreasing order and an invariant mass is calculated iteratively starting from track 1 to j, where j is the number of tracks associated to the jet. If this mass exceeds a specific threshold directly related to the mass of c-hadrons, the impact parameter significance of the last used track is added to the calculation of the b-tagging discriminator. The threshold value is set to be 1.5 GeV/c^2 and lies below a typical charm hadron mass in order to account for not reconstructed neutral or charged decay particles.

The single discriminating variable d is calculated by using a Likelihood ratio method. It is defined by the following two equations

$$\mathcal{L}^{b,c,q} = f^{b,c,q}(\alpha) \times \prod_i f_\alpha^{b,c,q}(x_i) \tag{5.5}$$

and

$$d = f_{BG}(c) \times \frac{\mathcal{L}^b}{\mathcal{L}^b + \mathcal{L}^c} + f_{BG}(q) \times \frac{\mathcal{L}^b}{\mathcal{L}^b + \mathcal{L}^q} \tag{5.6}$$

with α being the secondary vertex category, x_i the different input variables, q denotes u-, d-, s-quark and gluon jets. The expected prior for c- and q-content in non-b jets is $f_{BG}(c, q)$ with $f_{BG}(c) + f_{BG}(q) = 1$. $f^{b,c,q}(\alpha)$ is the probability for flavour b, c and q to fall in category α and $f_\alpha^{b,c,q}(x_i)$ are the corresponding probability density functions for observable x_i in case of category α and flavour b,c and q. They are extracted from simulated Monte Carlo events for various bins of jet energy in transverse direction or the pseudorapidity η of the jet.

- **SoftLepton Algorithms:**
  The presence of electrons or muons in the final state of semi-leptonically decaying b-hadrons is exploited in the so-called SoftLepton algorithms. In contradiction to the

official lepton identification algorithms which tend to isolated leptons, a dedicated lepton identification for non-isolated muons and electrons is used. The identification of lepton tracks among all reconstructed particle tracks of a jet is done in the following way:

In case of the SoftMuon algorithm, "standalone" muons are taken and a matching procedure is applied which compares the muon track to tracks reconstructed in the silicon tracking detector. A positively matched track requires 70% of the reconstructed hits of that track to be in common with the "standalone" muon track. The muon matching is limited to tracks with $p_{\mathrm{T}} > 3.5$ GeV/c due to a cut in the "standalone" muon reconstruction procedure.

Electrons used by the SoftElectron algorithm are determined by extrapolating the charged particle tracks from the silicon tracking device towards the calorimetry. The energy deposit in a region around the track is analysed and the extrapolated tracks are matched to ECAL clusters and HCAL towers. Due to this, a variety of variables can be derived and a single variable for selecting proper electron candidates is computed by using a neural network technique.

Parameters of the matched muons or electrons and parameters of the jet are then fed into a feed-forward neural network and its output is taken as discriminator for the algorithm. The following variables are used as input for the neural net:

- The transverse momentum of the lepton relative to the lepton-excluded jet axis. It is calculated as p_{T}-weighted average of the other charged tracks in the jet.
- The ratio of lepton momentum and jet energy.
- The impact parameter significance of the lepton track in three dimensions.
- The distance between lepton and lepton-excluded jet axis in the $\eta - \phi$ plane.

An additional approach with the same input variables except the impact parameter significance is implemented as well but was not studied in this thesis.

5.2.2 Jet, Track and Vertex Reconstruction

The studies presented here have been obtained from a simulated data sample of about 22 000 inclusive $t\bar{t}$-events. They were produced using the tools presented in chapter 3.4.2 More detailed information about the Monte Carlo event generator settings and software version can be found in appendix B.1. However, the achievable b-tagging performance may vary depending on the underlying event topology as discussed in [66]. QCD dijet events, for example, lead to a different overall light flavour tagging efficiency, but the relative performance change due to several misalignment scenarios is expected to be unchanged. Since the main focus of this study lies on the algorithm behaviour with respect to misalignment of the silicon tracking device and thus the derived relative performance change

of the algorithms, $t\bar{t}$-events have been studied here as a typical benchmark use case for b-tagging.

- **Jet Reconstruction:**
  The jets used as input for the b-tagging studies were reconstructed from calorimeter towers with the "IterativeCone" algorithm described briefly in chapter 3.4.3. The configurable cone size of the algorithm was chosen to be $\Delta R \leq 0.5$ and jets with a transverse momentum of $30 < p_T < 120$ GeV/c and in a pseudorapidity region of $|\eta| < 2.4$ corresponding to the silicon tracker coverage were selected.

- **Track Reconstruction:**
  The charged particle tracks are shared between all lifetime based b-tagging algorithms and were reconstructed by using the CMS standard "CombinatorialTrackFinder" algorithm described in 3.4.3. Table 5.1 presents the quality cut parameters used in order to select and reconstruct tracks which were then associated to particle jets found in the event. This is done by casting a cone around the reconstructed jet axis whereas all tracks inside this cone were selected.

  The default association cone size is set to $\Delta R \leq 0.5$ and is reduced to $\Delta R \leq 0.3$ in case of the JetProbability and CombinedSecondaryVertex algorithm. Since ΔR has quite some impact on the achievable b-tagging performance, a comparison of different cone sizes is discussed in chapter 5.4.5. One should be aware that the cuts on the transverse and longitudinal impact parameters in this case refer to the centre of the beam spot and not to the reconstructed location of the primary event vertex.

- **Primary Vertex Reconstruction:**
  Input for the primary vertex reconstruction [70] are the same tracks that passed the track quality cuts as described above. A cut of 5 on the maximum impact parameter significance with respect to the beam spot position is applied in order to reject tracks which are not compatible with the primary interaction point of the event. All remaining tracks are combined in clusters depending on the z-component of their closest approach to the beam line and a cluster separation of at least 1 mm is required. In a last step using the "Adaptive Vertex Fitter" algorithm [69], a primary vertex candidate is produced. The distance between the resulting vertices and the beam spot must not exceed 0.2 mm and the fit probability has to be at least 1%. For each vertex candidate, the p_{T}^2 sum of the associated tracks is calculated and the vertex with the highest value is selected as primary vertex of the hard proton-proton interaction. The expected efficiency for reconstructing the correct primary vertex is almost 100%.

- **Secondary Vertex Reconstruction:**
  Secondary decay vertices and thus several kinematically and topologically related variables are used by the CombinedSecondaryVertex algorithm described above. In addition, the newly developed b-tagging algorithm introduced and studied at the end

Table 5.1: Basic track selection cuts used by the "CombinatorialTrackFinder" algorithm in order to reduce the number of misreconstructed or fake tracks. The tracks which pass these cuts are used as input for all lifetime-based b-tagging algorithms. In case of the JetProbability and CombinedSecondaryVertex algorithm, the cut on ΔR is reduced to 0.3

Parameter	Cut
Number of hits in the silicon tracker	≥ 8
Number of hits in the silicon pixel layers	≥ 2
Transverse impact parameter	< 0.2 cm
Longitudinal impact parameter	< 17 cm
Transverse momentum	> 1.0 GeV/c
Normalised χ^2 of the track fit	< 5.0
ΔR to jet axis	≤ 0.5

of this chapter relies on the reconstruction of a secondary vertex as well. The reconstructed charged particle tracks associated to the jet act as input for the "Adaptive Vertex Finder" algorithm without further selection cuts. The AVF iteratively forms new vertices out of the remaining tracks left over from previous iteration steps. This procedure typically leads to a re-reconstruction of the primary vertex as well as to secondary or even tertiary decay vertices inside the jet cone. In order to remove implausible vertex candidates, like the re-reconstructed primary vertex, several distance and significance cuts with respect to the globally reconstructed primary event vertex are applied in a final step. Additionally, the secondary vertex based b-tagging algorithms provide a simple rejection of separated decay vertices originating from lighter particles such as the neutral K_s^0-meson. K_s^0-mesons typically decay in a π^+- and π^--meson which form a V-shaped decay pattern in the silicon tracking device. Due to this shape and the non-detectable neutral K_s^0-mesons, the rejection of such vertices is called V0-rejection and is simply applied by comparing vertex properties like the invariant mass of all two-track vertices with the one obtained by using a K_s^0 hypothesis.

Another approach, applying cuts on the minimum impact parameter significance of the tracks used for the AVF, shows that the reconstruction efficiency of the AVF is reduced in such a case. Table 5.2 and 5.3 summarises the settings of the "Adaptive Vertex Finder" and the default cuts used for the secondary vertex reconstruction, respectively. The overall SV reconstruction efficiency of the AVF is discussed in

the following section. In case of the misalignment studies presented in this thesis, several variations of these cuts have been applied and the results are presented in chapter 5.4.2.

Table 5.2: Configuration parmameters of the "Adaptive Vertex Finder" used in the secondary vertex reconstruction procedure

Parameter	Cut
Track weight	> 0.5
Reduced distance for PV association	< 1.8
Reduced distance for SV association	< 6.0

Table 5.3: Basic quality selection cuts for secondary vertices. In addition to these, a V0-rejection is implemented in the secondary vertex based algorithms as well

Parameter	Cut
Vertex mass	< 6.5 GeV/c^2
Fraction of tracks shared with the PV	< 0.65
Track multiplicity	≥ 2
2D flight distance	0.01...2.5 cm
2D flight distance significance	> 3.0
ΔR to jet axis	< 0.5

5.2.3 Input Variables

In this section, a selection of the most important input variables for the various b-tagging algorithms is presented. A more complete list is available in appendix C.1. All plots show three distributions on top of each other representing the b-, c- and light flavour jet content whereas the latter one includes gluon-jets as well. In order to determine the true flavour of the examined jet, the CMS standard analysis tools have been used. The "algorithmic" definition evaluates the type of the jet and derives the true flavour from the Monte Carlo truth information in the event. b-jets originating from gluon splitting, for example, are

labelled as b-jets and not as gluon jets as it would be by using the "physics" definition of the algorithm. Thus, the "algorithmic" definition leads to gluon tagging efficiencies which are similar to the ones obtained in case of light flavour jets.

Track Parameter Based Variables

Track parameter based variables act as input for most of the b-tagging algorithms available at CMS. The TrackCounting and JetProbability algorithms rely solely on them whereas the CombinedSecondaryVertex and SoftLepton algorithms use them in addition to secondary vertex or lepton based variables. Especially, the significance of the track impact parameter turns out to discriminate quite well between tracks associated to b-jets and tracks originating from the decay of other flavours.

Figure 5.3 shows the track impact parameter value as well as its significance for the three jet flavour categories b, c and udsg (udsg abbreviates up, down, strange and gluon jets) in the transverse plane. Tracks originating from b-jets have significant higher values with respect to tracks from charm or light flavour background, which is represented by the large tail of the distribution. However, the shape for c-jets does not differ as much from b-jets as a result of the already mentioned measurable lifetime of c-hadrons and the therefore existing real impact parameters of the associated tracks. The distribution of the impact parameter measurement error as well as of the impact parameter significance calculated in three dimensions is shown in figure C.1 and C.2.

The track multiplicity of charged particle tracks associated to the jet is presented in figure C.3. b-jets consist of slightly more tracks passing the track selection cuts than c- or udsg-jets.

Secondary Vertex Based Variables

The CombinedSecondaryVertex as well as the SimpleSecondaryVertex algorithm described later on in this chapter rely on input variables derived from the reconstruction of a secondary decay vertex inside a jet. First of all, the efficiency[1] of finding and reconstructing a secondary vertex in jets of a particular flavour is of interest. This issue is discussed in [71]. The reconstruction efficiency sets a limit on the maximum b-tagging efficiency which can be achieved by using information related to a secondary vertex and is defined as

$$\epsilon_{SV} = \frac{number\ of\ taggable\ jets\ with\ a\ reconstructed\ SV}{number\ of\ all\ taggable\ jets} \tag{5.7}$$

[1] It is more or less the fraction of reconstructed secondary vertices in a taggable jet of particular flavour than an efficiency in the classical sense. Taggable jets are jets with at least 3 tracks

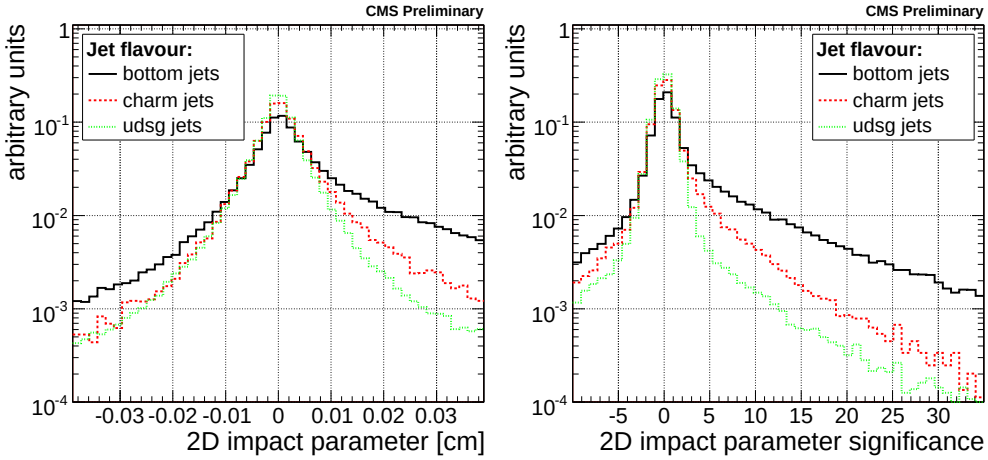
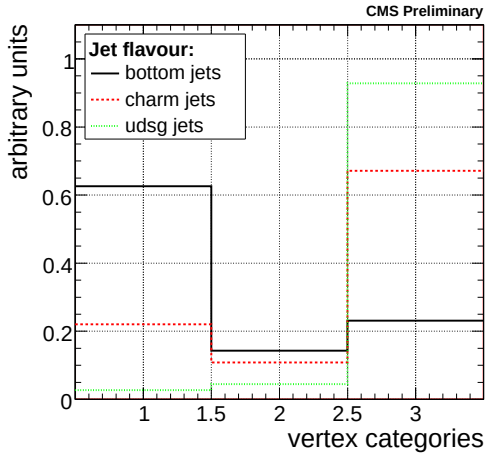


Figure 5.3: Impact parameter value (left) and significance (right), both calculated in the transverse plane. The black line shows the distribution in case of b-jets whereas red and green depicts c- and light flavour jets including gluons, respectively

A taggable jet is a jet with at least 3 associated tracks. In case of a perfectly aligned tracking detector, the efficiency of finding a secondary vertex inside a b-jet is about $\epsilon_{SV} = 63\%$. Figure 5.4a shows the distribution of the three vertex categories introduced above. They are shown for b-, c- and udsg-jets and the values are summarised in table 5.4b. The first row marked as RecoVertex corresponds to the algorithm efficiency ϵ_{SV} to reconstruct a real secondary vertex inside a certain particle jet. The presence of a secondary vertex is already a very powerful discriminator in order to distinguish b-jets from non-b background jets.

The most important secondary vertex variables used as input for b-tagging are related to the separation between primary and secondary vertex. This separation is denoted as flight distance of the reconstructed secondary vertex and can be calculated in the transverse plane and in three dimensions. Like it is in case of impact parameters, the flight distance significance carries more information in order to correctly tag a b-jet since it takes the measurement error of the flight distance into account as well.

The flight distance significance is defined as ratio between flight distance value and measurement error and is shown in the left plot of figure 5.5 for the three dimensional case. Light flavour and gluon jets have no real secondary vertex, except for some negligible amount of V0-decays mostly rejected by the algorithm, and the distribution is strongly peaked around zero. Decay vertices from charm and especially bottom hadrons have larger flight distance values due to their non-negligible lifetime which is represented by the large tails towards higher values in the flight distance significance distribution. Figures C.4, C.5 and C.6 presented in appendix C.1 show the distributions of the flight distance value, error and significance in both two and three dimensions.



(a) Distribution of the vertex categories for b-, c- and udsg-jets. Category 1 corresponds to a real reconstructed vertex, whereas category 2 and 3 refer to PseudoVertex and NoVertex, respectively

Vertex category	SV reconstruction efficiency [%] b-jets	c-jets	udsg-jets
RecoVertex	62.6	22.0	2.7
PseudoVertex	14.3	10.9	4.5
NoVertex	23.1	67.1	92.8

(b) Secondary vertex category efficiency for b-, c- and udsg-jets. The first row marked as RecoVertex corresponds to the secondary vertex reconstruction efficiency ϵ_{SV}

Figure 5.4: Secondary vertex category efficiencies for the different jet flavours

Another variable is the invariant mass of charged particles associated to a secondary vertex which is shown in the right plot of figure 5.5. Secondary vertices originating from the decay of heavy b-hadrons have higher reconstructed invariant masses than it is in case of the lighter charm hadrons, as expected. The same is true for misrecostructed secondary or non-rejected V0 vertices inside a light flavour or gluon jet.

Additional secondary vertex variables used as input for the CombinedSecondaryVertex algorithm like the energy fraction of charged particles associated to the secondary vertex with respect to all charged particles associated to the jet and the multiplicity of charged particle tracks at the SV are shown in figure C.7 and C.8, respectively.

5.2.4 Algorithm Performance

Each CMS b-tagging algorithm calculates a discriminator value for each examined jet, depending on the particular input variables described above. Figure 5.6 shows the discriminator distribution for the TrackCounting (high efficiency) algorithm using the second track ordered by impact parameter significance as well as for the CombinedSecondaryVertex algorithm. A complete set of all distributions for the TrackCounting, JetProbability, CombinedSecondaryVertex, SoftMuon and SoftElectron algorithm are presented in appendix C.2. Basically, all distributions show a similar behaviour which allows to distinguish b-jets from non-b background jets. In all cases, b-jets possess higher discriminator values followed by charm jets and then by jets of lighter flavours and gluons.

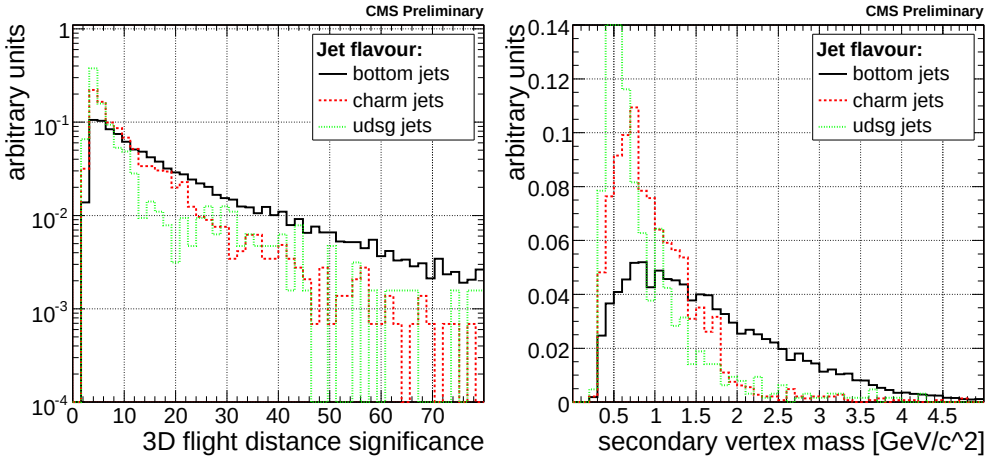


Figure 5.5: Secondary vertex flight distance significance in three dimensions (left) and invariant mass of charged particles associated to the secondary vertex (right). The black line shows the distribution in case of b-jets whereas red and green depicts c- and light flavour jets including gluons, respectively

Typically, the performance of a b-tagging algorithm is represented by a histogram, where the misidentification rate or misidentification efficiency is plotted against the efficiency to correctly tag a b-jet. These histograms are produced by calculating so-called flavour tagging efficiencies based on a specific cut on the discriminator. The flavour tagging efficiency ϵ_q labelled as b-jet efficiency in case of b-jets and non b-jet efficiency in case of charm, light flavour or gluon jets in all performance plots presented in this thesis is defined as

$$\epsilon_q = \frac{number\ of\ jets\ of\ flavour\ q\ tagged\ as\ b-jet}{number\ of\ jets\ of\ flavour\ q} \tag{5.8}$$

By scanning the whole discriminator range, a flavour tagging efficiency can be calculated for each possible cut on the discriminator output. The left plot in figure 5.7 presents as an example the flavour efficiencies for b-, c- and udsg-jets with respect to a cut performed on the b-tagging discriminator of the CombinedSecondaryVertex algorithm. It becomes clear that a given b-jet tagging efficiency always corresponds to one specific mistagging rate for charm and udsg background. The algorithm performance then is described by a performance plot where the flavour mistagging efficiency is plotted against the b-tagging efficiency as presented in the right plot of figure 5.7. A complete set of tagging efficiency and performance plots for all algorithms is given in appendix C.3.

The performance of the different b-tagging algorithms widely varies due to the various methods implemented in the algorithms which exploit several aspects of weakly decaying b-hadrons. Depending on the use case, it is upon the physicist to choose the right algorithm. An appropriate cut on the b-tagging discriminator according to the needs of the

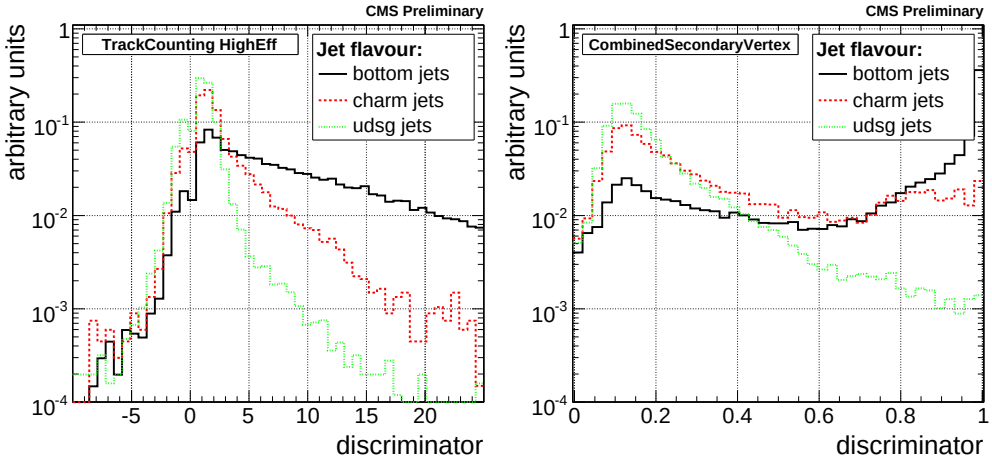


Figure 5.6: Discriminator output for the TrackCounting (high efficiency) algorithm (left) and for the CombinedSecondaryVertex algorithm (right). The black line shows the distribution in case of b-jets whereas red and green depicts c- and light flavour jets including gluons, respectively

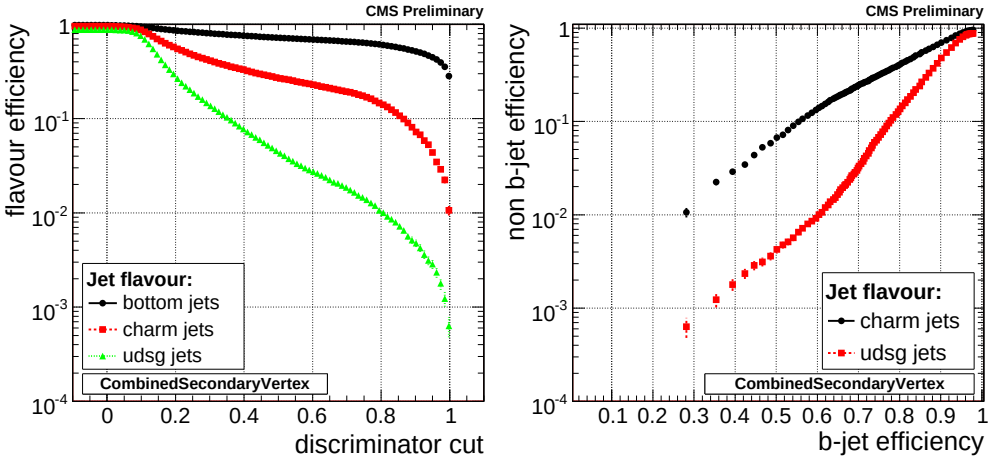


Figure 5.7: The left plot shows the tagging efficiencies for the different jet flavours with respect to the discriminator cut and the right plot the b-tagging performance, both for the CombinedSecondaryVertex algorithm.

desired analysis goal has to be found. For some studies, rather pure samples with a high reduction of background jets, but as well fewer selected b-jets, are preferable, whereas in other cases a high b-tagging efficiency is of prime importance. Therefore, three pre-defined discriminator cuts, so-called working points are determined for each algorithm which provide fixed mistagging rates for the different flavours. The three working points are called loose, medium and tight and correspond to a non-b mistagging rate of 10%, 1% and 0.1%,

respectively. Since the determination of such working points only makes sense on fully trained and, in terms of software development, stable algorithms, they have not been determined yet.

A comparison of the performance of different b-tagging algorithms is given in figure 5.8. The left plot shows the expected b-tagging performance of all taggers in case of light flavour jets including jets from gluons and the right plot corresponds to jets from the decay of a charm hadron. As can be seen, at 50% b-tagging efficiency, the curves of the two TrackCounting algorithms cross each other. As the name suggests, the TrackCounting high efficiency algorithm provides better performance in the high b-efficiency tagging region whereas the high purity algorithm scores otherwise. The performance of the two SoftLepton algorithms is limited to a convolution of lepton reconstruction efficiencies and the rather low decay probability of b-hadrons into either muons or electrons. The CombinedSecondaryVertex and JetProbability algorithms perform comparably in case of charm jets over the whole b-efficiency range, whereas the latter one gets slightly better in the high efficiency region in case of udsg-jets.

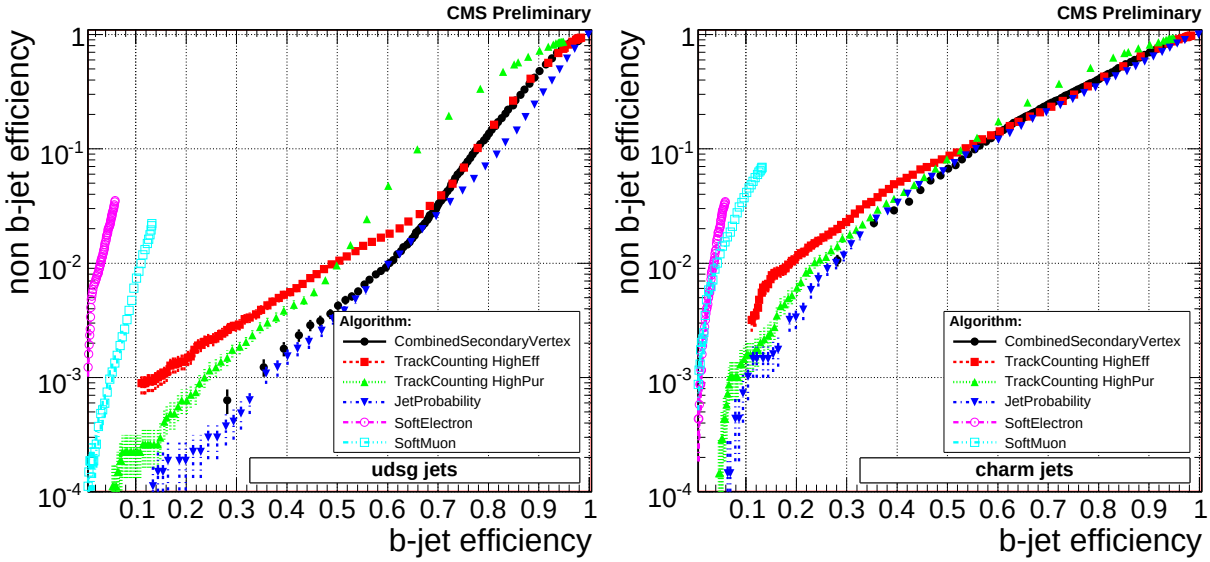


Figure 5.8: Comparison of the performance of all default CMS b-tagging algorithms for light flavour and gluon jets on the left, and jets originating from charm hadrons on the right

Since the b-tagging performances have been studied for the first time within the current software framework CMSSW, a short discussion and comparison with former results is presented in appendix E.1.

5.3 Alignment Studies

The studies presented up to now have all been performed with a perfectly aligned CMS detector. Actually, the measurement accuracy of each detector component involved in the b-tagging procedure has an impact on the achievable performance of the different b-tagging algorithms introduced above. The detector performance will be limited mainly by the misalignment of the silicon tracking device and the muon chambers as well as the miscalibration of the electromagnetic and hadronic calorimetry.

Due to the required high resolution on charged particle track impact parameters and reconstructed secondary vertices, the most important impact on the b-tagging performance stems from the silicon tracking device of CMS, which is studied here. Calorimeter miscalibration may affect the resolution, direction and energy scale of jets as well as the identification of electrons in case of the SoftElectron algorithm, whereas the misalignment of the muon system on the other hand may affect the SoftMuon algorithm. These effects remain as topics for future studies and were not considered here.

The accuracy of the tracking system relies strongly on the knowledge of the exact location and orientation of each subdetector component. This knowledge is denoted as alignment position and has to be determined for all of the about 16 000 silicon detector modules during data taking. The high intrinsic spatial resolution of 10 to 50 μm, depending on the detector module, requires a knowledge of the alignment parameters to similar precision in order to not compromise the performance of the detector. In order to improve these parameters, a track based alignment procedure is applied during the first phase of data taking. More information about the track based alignment can be found in [72], [73], [74] and [75]. Basically, hits of a charged particle track are spread over different strip and pixel detector modules and their position relative to each other helps to determine the real alignment position of each detector module. Therefore, an increasing number of tracks improves the alignment position measurement, which in turn relies on the integrated luminosity of the collected LHC data. The improved knowledge of the alignment position leads to reduced uncertainties on hit measurements and therefore on better measured tracks as well as all other reconstructed objects which depend on them.

5.3.1 Misalignment Scenarios

In order to study misalignment effects and their impact on the silicon tracking device, the CMS Tracker Alignment Group has defined various misalignment scenarios which provide a set of alignment parameters corresponding to specific amounts of collected collision data.

Each silicon detector module as well as each module compound provides six degrees of freedom, three translational and three rotational, which are adjusted according to the selected misalignment scenario. (See figure 3.5 for more details about the detector subcomponents.) Each scenario consists of a predefined set of fixed shifts and rotations in all directions x,

y and z as well as random ones following a Gaussian distribution. They are applied to all subdetector components like rods, petals, rings, strings and layers, down to the module-level of the device. The used scenarios are briefly introduced below and are described more precisely in [76], [77] and [78].

- **Startup Scenario:**
  The tracker alignment precision, which is assumed in this scenario, is expected to reflect the detector conditions at first collisions in CMS. The knowledge of the alignment properties is only based on information derived from survey measurements during construction of the detector, the CMS Laser Alignment System and reconstructed tracks, mostly muons, originating from the decay of incident cosmic particles. In this case, uncertainties on the alignment position and therefore on the hit measurement which are introduced by a limited mounting precision and movements of the detector components after cooling and commissioning of the magnetic field preponderate.

- **10 pb^{-1} Scenario:**
  This scenario corresponds to a supposed tracker alignment precision which is improved due to the track based alignment procedure for an integrated luminosity of 10 pb^{-1}. It is assumed that the silicon tracking detector can be aligned using tracks from cosmic muons, isolated hadrons in minimum bias events and muons originating from the decay of low mass resonances like the J/Ψ- and Υ-mesons. In this scenario, the alignment precision of the larger silicon pixel detector structures is improved by a factor of five with respect to the startup case, while there is no expected improvement for the alignment on module-level. However, the strip tracker can be aligned to an accuracy of about 100 μm.

- **100 pb^{-1} Scenario:**
  With about 100 pb^{-1} of data, the track based alignment can be further improved. In addition to the information mentioned above, a significant amount of muons with high transverse momenta originating from the decay of Z and W bosons are available. The expected alignment precision of the silicon pixel and strip detector is supposed to be of the order of 20 μm and 30 to 50 μm, respectively.

In addition to these three, a further misalignment scenario corresponding to 1000 pb^{-1} of collected data was studied as well. The results were almost similar to the ideal case, where no silicon tracker misalignment was introduced. Therefore, the impact on the b-tagging performance was negligible and the obtained results are excluded in order to maintain the clarity of the presented distributions.

During the initial phase of the LHC, the complete or partial deactivation of pixel detector components due to safety reasons can be a realistic scenario as well. Therefore, some modifications of the misalignment scenarios had to be implemented. This was done for all scenarios described above, but only the distributions for the 10 pb^{-1} misalignment scenario

are presented here. It simulates a deactivation of the innermost layer of the pixel detector which lies closest to the beam line. A deactivation of the second and third pixel layer was studied as well, but is excluded in this analysis due to being a more or less unrealistic use case.

5.3.2 Technical Implementation

A complete simulation of the misaligned detector with Geant4 and afterwards running the CMS standard reconstruction sequence for all the various misalignment scenarios was too expensive in terms of needed computing time and storage resources. Therefore, another approach was adopted which introduces the tracker misalignment at reconstruction level. The tracker hits are still reconstructed by assuming a perfect detector geometry and an ideally aligned silicon tracking device. After this reconstruction step, the different detector components and modules are shifted and rotated by pre-defined fixed amounts as well as by random values according to the used misalignment scenario. In order to keep a good tracking efficiency, the so called alignment position error (APE) is added in quadrature to the tracking hit error. The way the APE is chosen is somewhat done in an idealised way. More information on the APE and its impact on the b-tagging performance is presented in section 5.4.4. At the end, the charged particle tracks are reconstructed by using the CMS default reconstruction sequence of pattern recognition and track finding followed by jet-tracks association, secondary vertex reconstruction and b-tagging discriminator calculation as described in 5.2.2.

In order to simulate the deactivation of a silicon pixel layer, a pixel hit filter (PHF) was developed which removes tracker hits in the selected pixel barrel and endcap layer. Afterwards, the complete reconstruction chain as described above is run as well. This filter simulates the effect of a degraded track measurement as a result of the missing pixel layer. A mixed seeding procedure which includes the first layer of the strip detector was used in order to recover the loss in seeding efficiency due to the missing pixel layer.

The approach of only refitting the tracks after the misalignment is introduced bears different results than running the complete track reconstruction chain. This important side-result of the studies presented here is discussed in the following section. A sample configuration file for a misalignment and b-tagging job as it was used in this study is presented in appendix A.2 for documentation purpose.

5.3.3 Refitting Versus Re-Reconstruction of Tracks

There are two ways to apply the different misalignment scenarios. One approach is to just refit the tracks in a misaligned detector with hits belonging to tracks reconstructed in a perfectly aligned detector. The other approach is to re-run the complete tracking sequence, including pattern recognition, after the misalignment is introduced to the detector. As

stated in [12], these two methods should give similar results with respect to the track reconstruction and related applications which depend on such tracks. This is correct in many cases, but it turns out that the high quality demands on charged particle tracks by the different b-tagging algorithms as well as the huge amount of tracks inside jets render this statement insufficient as far as the identification of b-quark jets is concerned.

The following three effects reduce the achievable b-tagging performance in case of a full track reconstruction with respect to a track refit-only approach:

1. The track reconstruction efficiency in a misaligned detector is degraded.
2. The rate of hits wrongly assigned to tracks is increased. In cases, where there is for example only one wrongly assigned hit in the first pixel layer, the track itself is not considered to be fake, but the measurement of the track impact parameter might be severely wrong. This is the case in about 3% to 4% of the tracks associated to jets in the examined $t\bar{t}$ sample for an ideally aligned tracker.
3. The rate of fake tracks is increased. This applies for tracks, where more than 50% of the associated tracker hits do not belong to the track with respect to the simulation level.

Figure 5.9 shows the fake rate for reconstructed tracks associated to jets which pass the track quality cuts given in table 5.1 as a function of the pseudorapidity η for different misalignment scenarios. In comparison, figure C.19 gives the fake rate for all tracks in $t\bar{t}$-events with lower quality cuts applied as it is presented in [76] and [77].

It is shown that the tight track quality cuts used for b-tagging reduce the fake rate significantly. In case of a perfectly aligned detector, the fake rate always stays below 2% whereas it is increased up to 4% if all tracks in an event are considered. In case of the 10 pb^{-1} and startup scenario, the fake rate reaches about 10% and 15%, respectively, when the quality cuts are applied and 15% and 20%, respectively, without this tight track selection. A track refit-only approach would lead to a fake rate which stays always below 2% and 4% for all misalignment scenarios as it is shown in the distributions for a perfectly aligned silicon tracker.

It becomes clear that such an approach would underestimate the impact of fake tracks on the expected b-tagging performance, as shown for the high purity TrackCounting algorithm for the startup and for the 10 pb^{-1} misalignment scenario in figure 5.10. The distributions are given for light flavour and gluon jets since the discrepancy is much lower in case of jets originating from the decay of charm hadrons. A comparison of the performance of other b-tagging algorithms is given in appendix C.4.

Depending on the b-efficiency region, the performance decrease can become quite large. At a b-tagging efficiency of 30%, the mistagging rate degrades by more than a factor of two for startup tracker conditions and the TrackCounting (high purity) algorithm. Therefore,

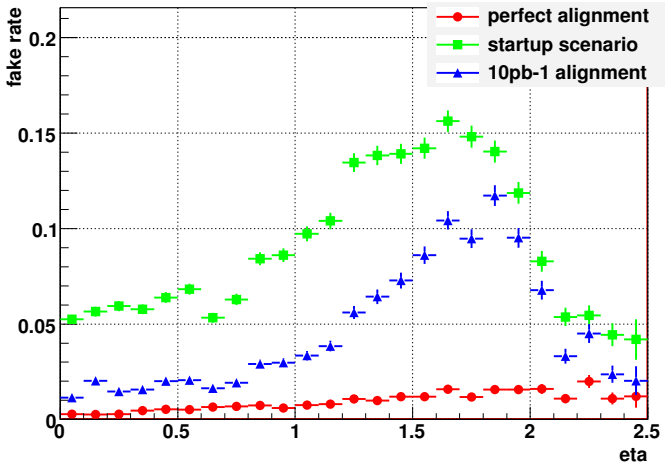


Figure 5.9: Fake rate as a function of the pseudorapidity for tracks associated to jets which pass the b-tagging selection cuts presented in table 5.1

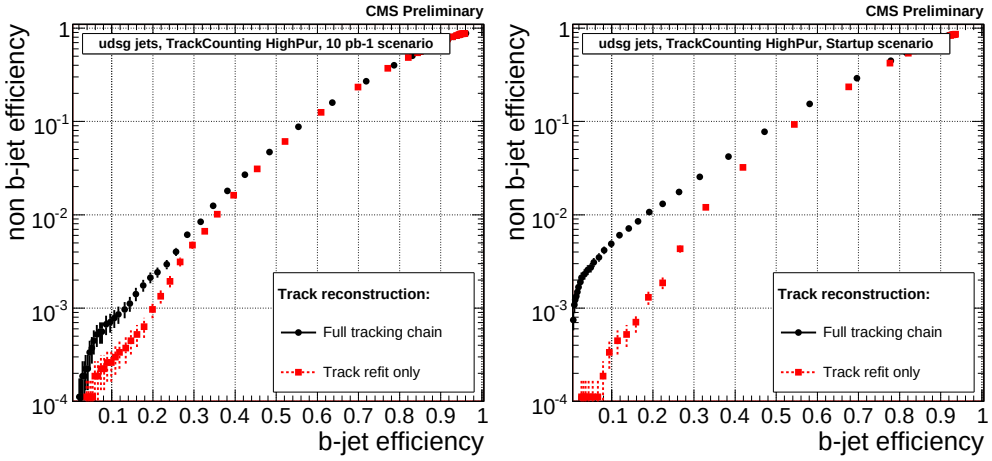


Figure 5.10: Performance of the TrackCounting (high purity) algorithm for the refit-only and the full track reconstruction approach for the 10 pb^{-1} (left) and the startup (right) misalignment scenario for light flavour and gluon jets

all figures presented in this chapter have been obtained by running the full tracking chain after the various misalignment scenarios have been applied.

5.3.4 Impact on the Input Variables

The various input variables which are used for b-tagging at CMS behave completely differently with respect to the degree of misalignment applied to the silicon tracking device. In the following section, the sensitivity to misalignment is studied for both, track impact parameter and secondary vertex based variables.

Track Parameter Based Variables

Figure 5.11 presents the track impact parameter significance with respect to the different misalignment scenarios. The distributions are given for b-, c-, light flavour and gluon jets and show the second highest track ordered by the impact parameter significance. Thus, they represent the discriminator distributions of the high efficiency TrackCounting algorithm.

It can be seen that the impact parameter significances decrease for all flavours when the misalignment parameters are increased. Tracks from b-jets are affected the most and the shape of their distribution converges towards what is observed for tracks from udsg-jets. This behaviour can be explained by looking at both the impact parameter value and its measurement error. Their distributions are presented in figures 5.12 and 5.13, respectively.

With increasing misalignment, the measurement error is increased for all flavours in a similar way but the value of the impact parameter behaves completely different. The impact parameter distributions for b-jets are not affected that much due to the fact that the impact parameters are already different from zero as a result of the lifetime of the b-hadrons. Thus, a larger error directly leads to a decreased significance. In case of light flavour and c-jets, the distributions smear out and form tails towards higher values which can not be described by a Gaussian distribution due to misreconstructed tracks. Nevertheless, the impact parameter significances in the core of the distribution for udsg-jets should remain unchanged.

A distinctive feature can be seen in the light flavour distribution of the transverse impact parameter significance in the startup and very slightly in the 10 pb^{-1} scenario. A bump is shown in the positive significance region between 3 and 15. If all tracks associated to a jet are taken into account, the bump is visible as well on the negative side, as shown in the left plot of figure C.24. This behaviour can be explained by the large number of fake reconstructed tracks in case of a misaligned silicon detector as stated above. The right plot of figure C.24 shows the same distribution for the refit-only approach. In this case, the fake rate always stays below 2% and no bump is visible on either side of the distribution.

Another important issue is the correct determination of the impact parameter errors with respect to the different misalignment conditions. This can be verified by checking the pull distributions of the d_0 parameter with respect to several misalignment scenarios which is

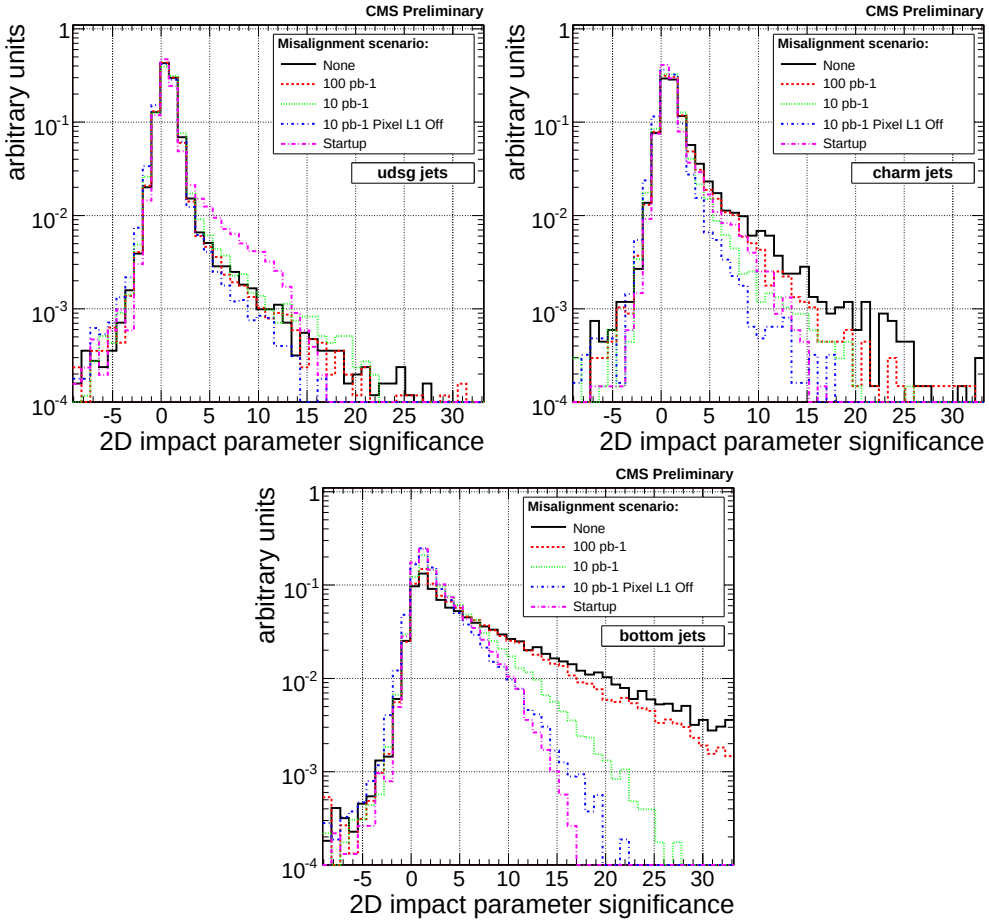


Figure 5.11: Distribution of the transverse impact parameter significance of the second highest track ordered by impact parameter significance for the different misalignment scenarios. The distributions are presented for udsg-jets (top left), charm jets (top right) and bottom jets (bottom)

given in figure 5.14. The pull of the d_0 parameter is defined as:

$$pull(d_0) = \frac{d_0^{\mathrm{rec}} - d_0^{\mathrm{sim}}}{\sigma d_0} \tag{5.9}$$

d_0^{rec} and d_0^{sim} are the reconstructed and simulated values of d_0 and σd_0 is the corresponding measurement error. The d_0 parameter is defined as the transverse impact parameter of a track with respect to the origin of the CMS coordinate system $(0, 0)$. This differs from the impact parameter definition with respect to the reconstructed primary vertex as it is used for b-tagging at CMS. Nevertheless, it is assumed that the conclusions remain the same for either definition. It can be seen that the impact parameter measurement errors are

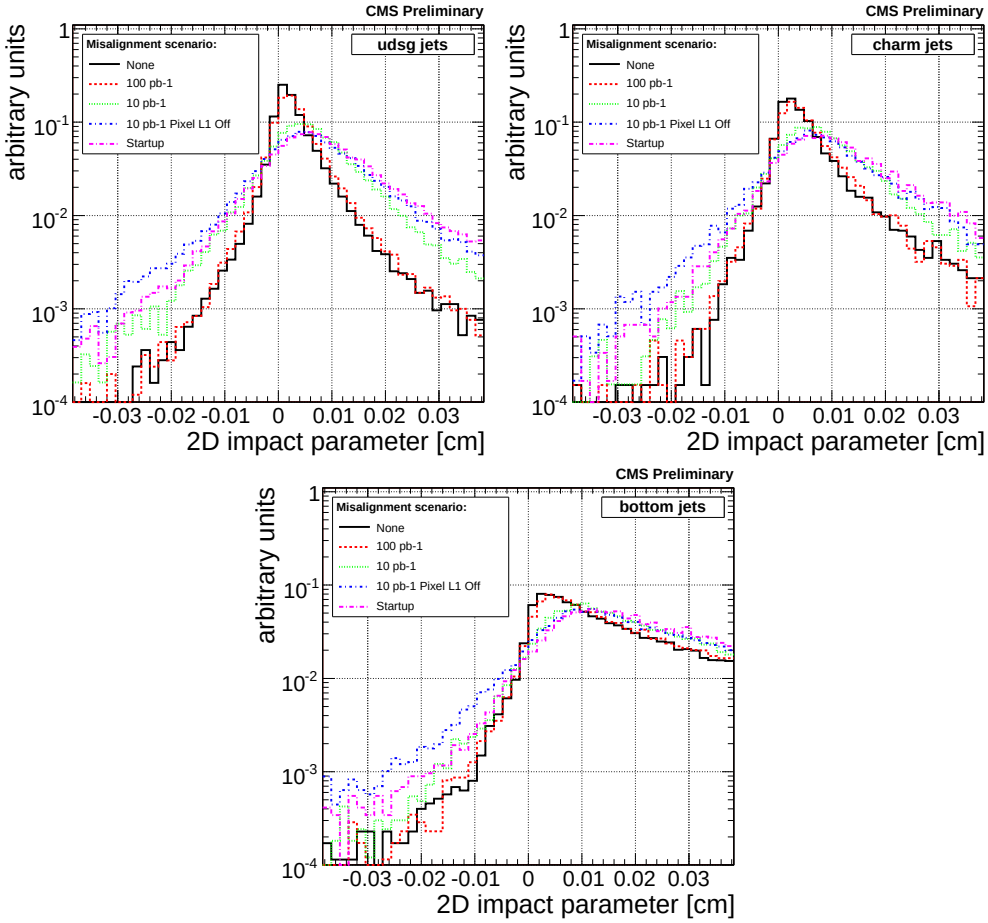


Figure 5.12: Distribution of the transverse impact parameter value of the second track ordered by impact parameter significance for the different misalignment scenarios. The distributions are presented for udsg-jets (top left), charm jets (top right) and bottom jets (bottom)

determined correctly, since the widths of the distributions do not change for the presented misalignment scenarios.

What can be seen is that misalignment has a large effect on track impact parameter based input observables and therefore its impact on the algorithm performance is expected to be rather large. The consequences are discussed in the next section.

Secondary Vertex Based Variables

As already mentioned, misalignment strongly affects the accuracy of reconstructing charged particle tracks and thus of all physics objects depending on them. Therefore, its impact

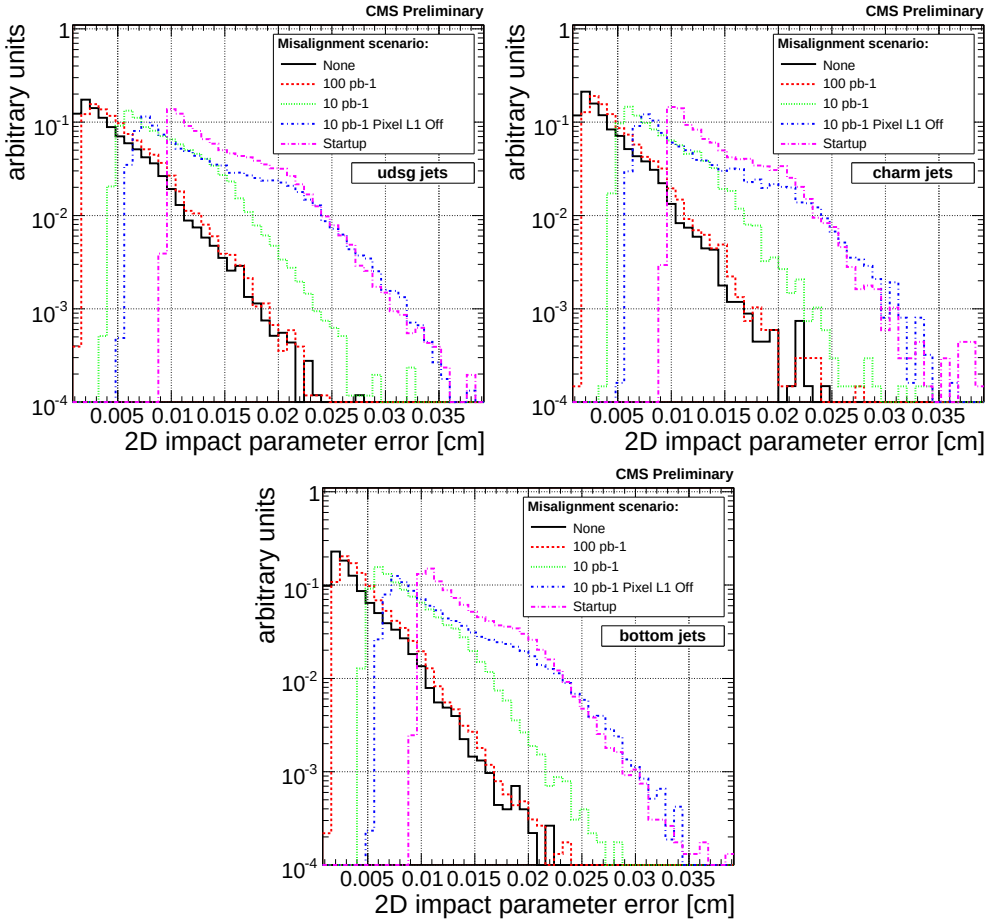


Figure 5.13: Distribution of the transverse impact parameter measurement error of the second track ordered by impact parameter significance for the different misalignment scenarios. The distributions are presented for udsg-jets (top left), charm jets (top right) and bottom jets (bottom)

on the reconstruction efficiency of secondary vertices is of substantial interest. Table 5.4 presents the efficiencies for reconstructing a secondary vertex in a particular jet flavour with respect to the different misalignment scenarios.

The efficiencies for a perfectly aligned detector and for the 100 pb^{-1} scenario are almost similar for all three jet flavours. The other scenarios are accompanied by a strong degradation which can be observed for b- and c-jets. Basically, tracks close to the primary interaction point which belong to a secondary decay vertex are more often associated wrongly to the primary vertex instead. This is due to larger track parameter errors which come along with increasing misalignment.

In addition, vertices are as well rejected by the secondary vertex finding algorithm due to

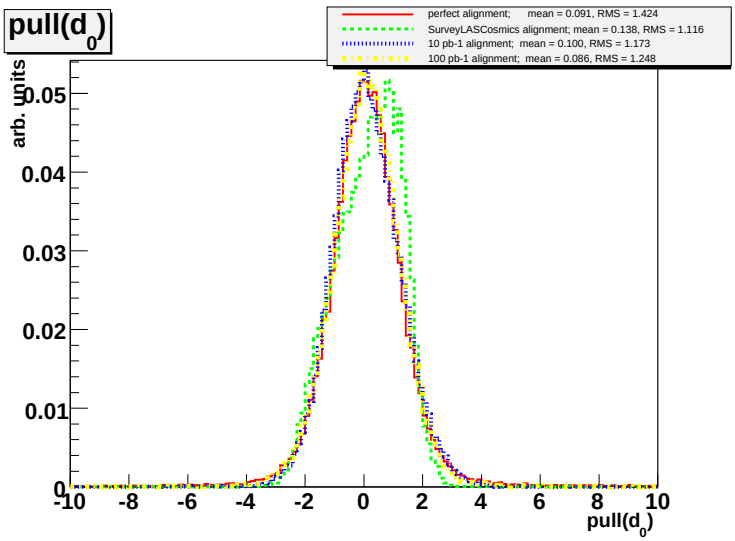


Figure 5.14: Pull distribution of the transverse track impact parameter d_0 measured with respect to the centre of the CMS coordinate system $(0, 0)$. The misalignment scenario denoted as SurveyLASCosmics corresponds to tracker startup misalignment conditions. The distributions are given for single muons with a transverse momentum of $p_{\mathrm{T}} = 100$ GeV/c

cuts on the decay length or impact parameter significances. However, a difference between the efficiency decrease in case of jets originating from the decay of charm and bottom hadrons is noticeable. The number of reconstructed secondary vertices with the startup scenario applied is only about one third of what could be achieved in case of an ideally aligned detector. This behaviour is different for b-jets, where more than 50% of the vertices can still be reconstructed due to the larger lifetime of b-hadrons.

The impact of misalignment on the most important secondary vertex observable, the flight distance significance, is presented in figure 5.15. The distributions are given for the three jet flavours b, c, and udsg, calculated in three dimensions. The corresponding values and measurement errors are shown in figures 5.16 and 5.17, respectively. It can be seen from the flight distance distributions that strongly displaced secondary vertices are more or less independent of misalignment, whereas vertices close to the primary interaction point suffer from the reasons described above. The increased measurement error, however, leads to lower significances at all.

The top left plot of figure 5.18 shows the number of tracks associated to the secondary vertex with respect to misalignment in case of b-quark jets. In addition, the invariant mass at the secondary vertex and the energy of these tracks compared to the energy of all tracks assigned to the jet are presented there as well. All three variables are only slightly sensitive to the degree of misalignment. A reconstructed secondary vertex has tracks with quite large impact parameters due to the quality cuts used in the secondary vertex finding

Table 5.4: Fraction of jets with a reconstructed secondary vertex presented for the different misalignment scenarios in case of light flavour and gluon, charm and bottom jets

	SV reconstruction efficiency [%]		
Misalignment scenario	b-jets	c-jets	udsg-jets
No Misalignment	62.6	22.0	2.7
100 pb^{-1} Misalignment	62.1	19.6	2.4
10 pb^{-1} Misalignment	53.0	12.7	2.9
10 pb^{-1} Pixel L1 Off Misal.	39.2	7.6	1.9
Startup Misalignment	37.8	7.7	3.5

algorithm. The same distributions for udsg- and charm jets are presented in appendix C.5.

These studies suggest that secondary vertex based variables and therefore b-tagging algorithms which rely only on the presence of reconstructed secondary vertices inside a jet provide a more robust behaviour with respect to misalignment, than the impact parameter based taggers. This issue was studied further and the development of such a b-tagging algorithm for the CMS experiment is described in section 5.4.

5.3.5 Expected Algorithm Performance

Finally, after introducing the impact of misalignment on the different input variables, the overall effect on the b-tagging performance itself is determined. The absolute b-tagging performance with respect to misalignment for all b-tagging algorithms is presented in appendix C.6. For a better comparison of the expected performance decrease of each algorithm in case of misalignment with respect to a perfectly aligned silicon tracker, a set of ratio plots have been produced. For each given b-tagging efficiency, the relative performance decrease defined as

$$P_{\text{dec}} = \frac{\epsilon_{c,udsg}^{\text{misalign}}}{\epsilon_{c,udsg}^{\text{ideal}}} \tag{5.10}$$

was calculated. $\epsilon_{c,udsg}^{\text{misalign}}$ corresponds to the mistagging efficiency for a misaligned detector, whereas $\epsilon_{c,udsg}^{\text{ideal}}$ presents the misidentification rate in an ideally aligned tracking device. The

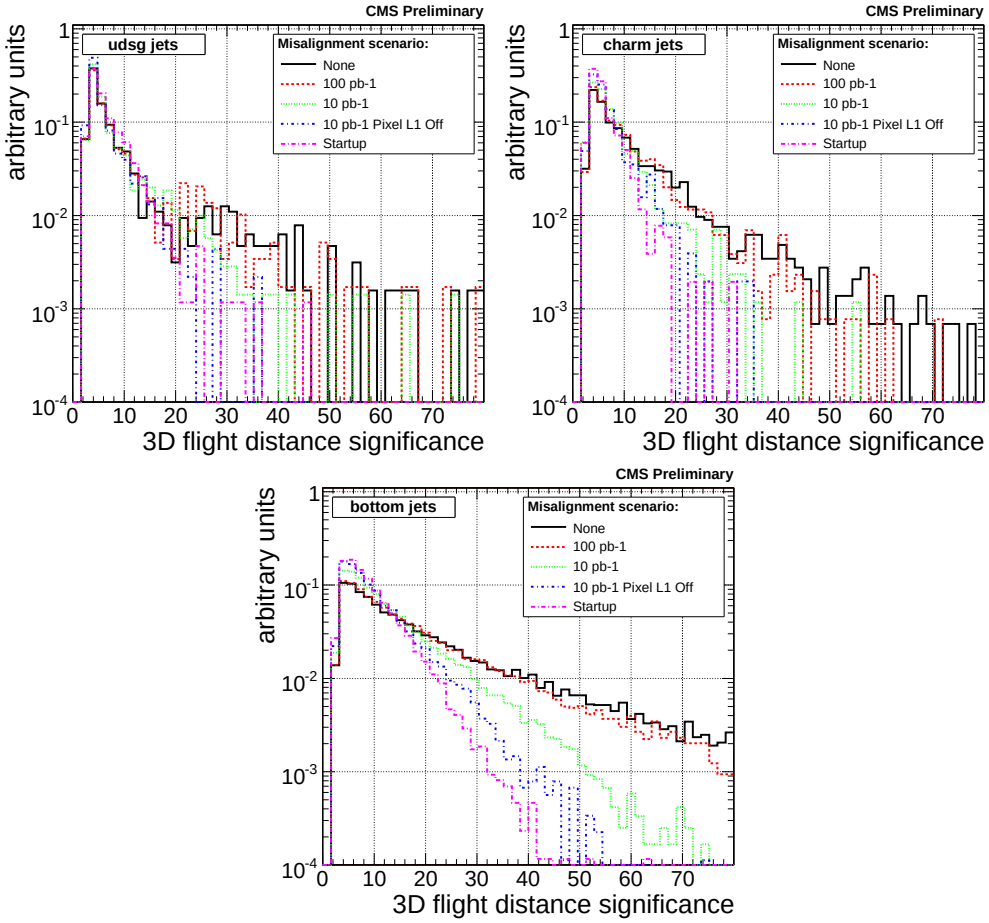


Figure 5.15: Distribution of the secondary vertex 3D flight distance significance for the different misalignment scenarios. The distributions are presented for udsg-jets (top left), charm jets (top right) and bottom jets (bottom)

relative performance decrease for all studied algorithms is given in figures 5.19 and 5.20 for c- and udsg-jets, respectively.

In general, the relative performance decrease is much more pronounced for light flavour and gluon jets. However, it can be seen that the various b-tagging algorithms react in a different way with respect to a selected misalignment scenario. The distributions show that the lepton based b-tagging algorithms only depend slightly on the silicon tracker misalignment, whereas other taggers are much more affected. The performance of the SoftElectron tagger, for example, is degraded by a factor of about two for udsg-jets when tracker startup conditions are applied. The CombinedSecondaryVertex algorithm, on the other hand, behaves much worse. A relative performance decrease of up to a factor of 35 can be observed with respect to light flavour and gluon rejection. This corresponds to a reduced b-tagging efficiency of about $20 - 40\%$ for the startup scenario, if the same mistagging

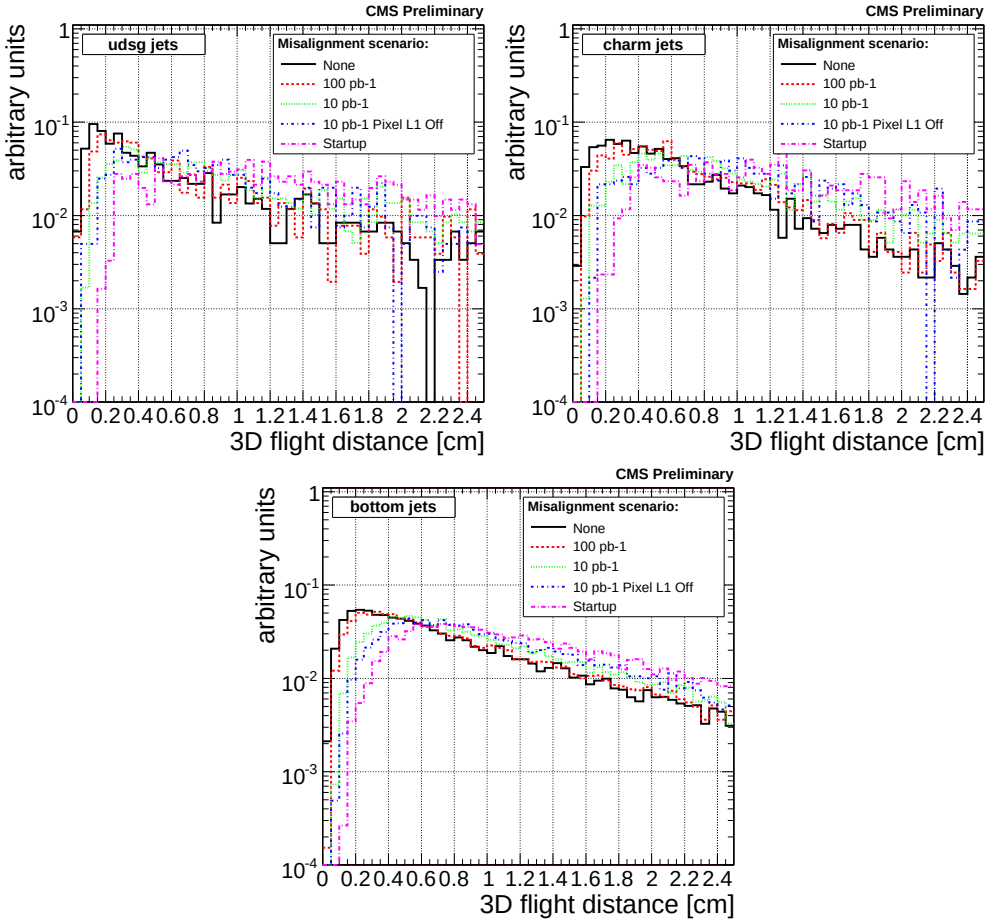


Figure 5.16: Distribution of the secondary vertex 3D flight distance value for the different misalignment scenarios. The distributions are presented for udsg-jets (top left), charm jets (top right) and bottom jets (bottom)

rate as in the ideal case is desired. The performance of the JetProbability algorithm is drastically degraded as well. The udsg mistagging rate, for example, is increased by up to a factor of 25 in comparison to a perfectly aligned silicon detector.

Both algorithms, the JetProbability and CombinedSecondaryVertex, suffer from the same fact, namely that they require a proper training in order to determine e.g. needed probability density functions (PDF) for the likelihood method which is used for discriminator calculation. The PDFs have to be determined for the different categories of the algorithms in case of background and signal events. They should be obtained from collision data collected at CMS after the LHC startup. Therefore, several methods for the measurement of b-tagging and misidentification efficiencies from data have been studied in [79] and [80].

Since it is very hard to obtain clean signal and background PDFs from real data, the

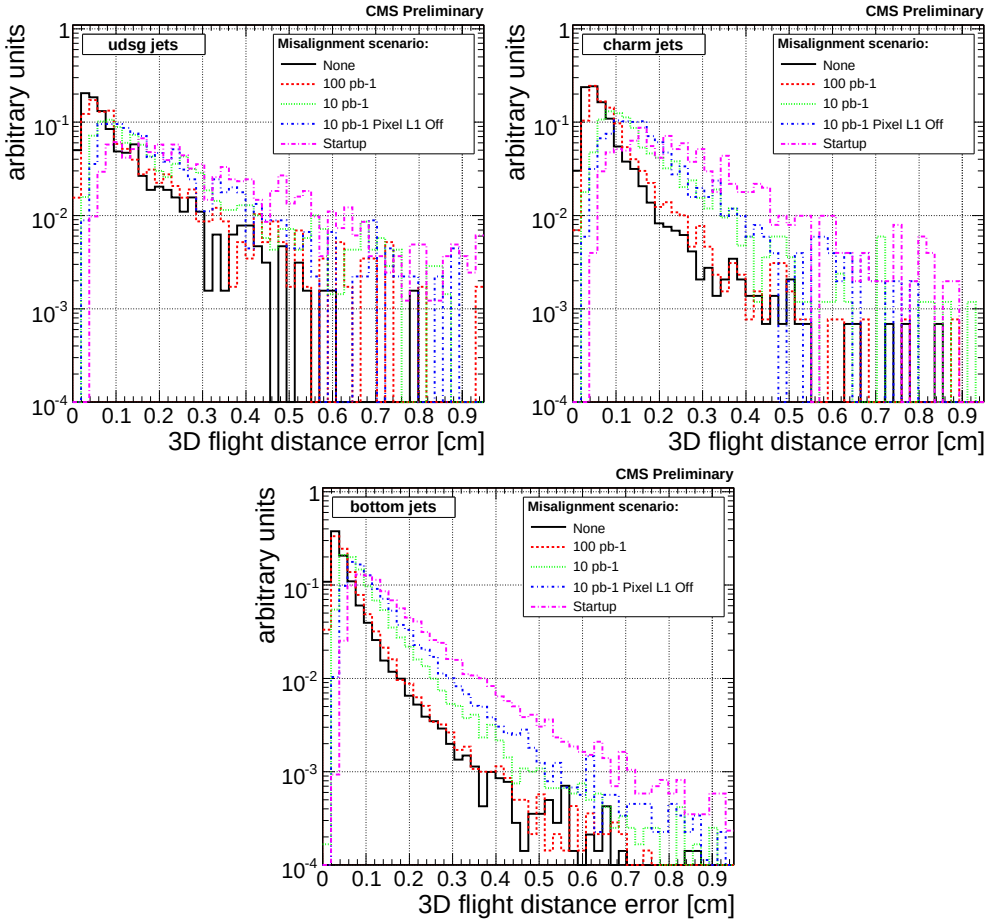


Figure 5.17: Distribution of the secondary vertex 3D flight distance measurement error for the different misalignment scenarios. The distributions are presented for udsg-jets (top left), charm jets (top right) and bottom jets (bottom)

determination can also be made on Monte Carlo samples which have been tuned for real collision events. Both possibilities require huge data samples with statistical precision in order to satisfy all available algorithm categories as well as a very good understanding of the involved detector components. Basically, at the time when a sufficient amount of data has been collected in order to train the algorithms, the misalignment has as well been determined to at least a level which corresponds to the 100 pb^{-1} misalignment scenario. Thus, the algorithm behaviour directly reflects the lack of knowledge about misalignment and of systematic uncertainties that exists during the first phase of the running experiment.

All b-tagging algorithms provide a certain amount of discrimination power to distinguish between b- and non b-jets, even with the startup misalignment scenario applied. However, the more sophisticated lifetime-based algorithms like the JetProbability and Combined-SecondaryVertex may not be used in a reasonable way at the beginning of the LHC data

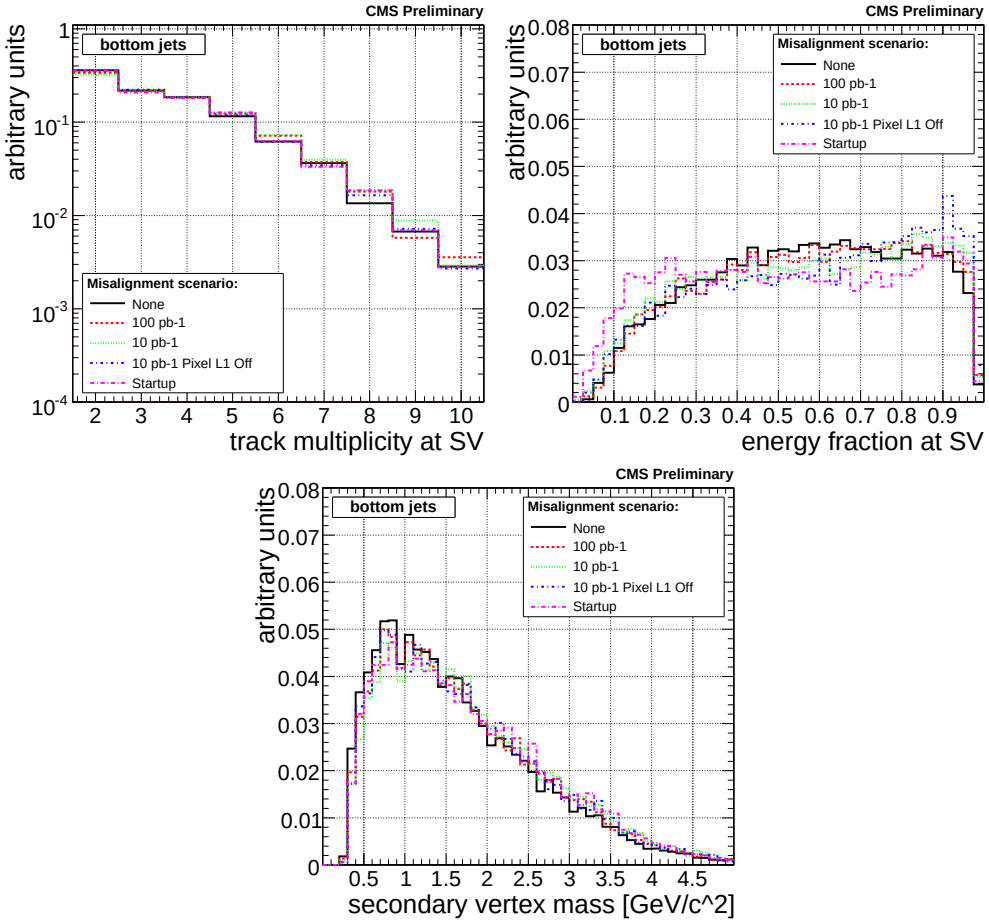


Figure 5.18: Multiplicity of charged particle tracks associated to the secondary vertex (top left) as well as energy fraction (top right) and invariant mass (bottom) of tracks at the secondary vertex in case of b-jets

taking.

5.4 Secondary Vertex Only Based Algorithm

The results presented in the last sections have shown that impact parameter and secondary vertex based variables react completely differently on the degree of misalignment which is applied to the silicon tracking device. The expected performances of the different algorithms widely vary and most of the taggers bear much worse mistagging rates during the first period of data taking at the LHC. At first sight, the secondary vertex based variables seemed less affected by misalignment. Based on this knowledge, a new b-tagging algorithm which solely depends on the presence of a reconstructed secondary vertex inside a jet was

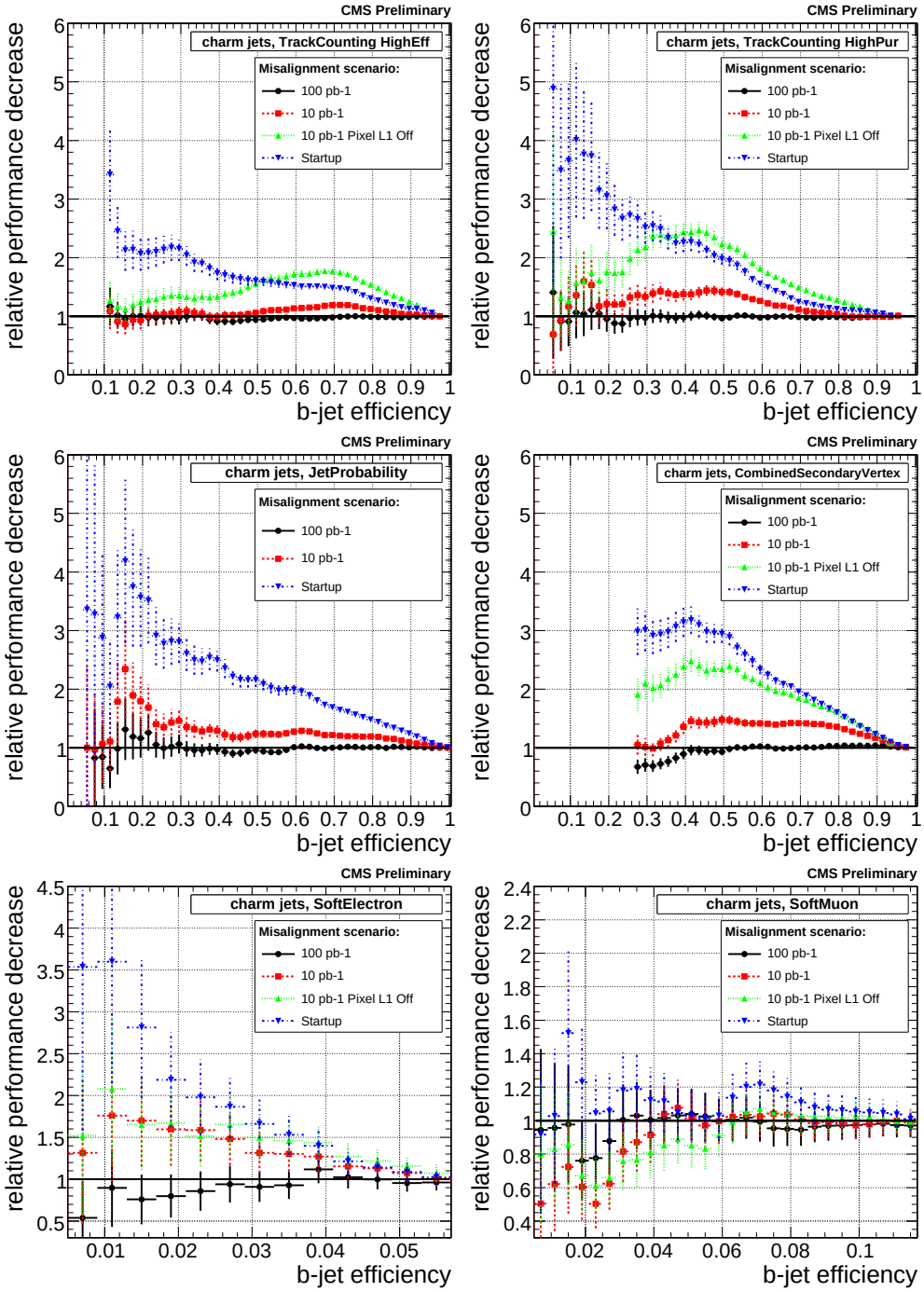


Figure 5.19: Relative performance decrease in charm mistagging rate versus b-tagging efficiency for the TrackCounting high efficiency (top left) and high purity (top right), the JetProbability (middle left), the CombinedSecondaryVertex (middle right), the SoftElectron (bottom left) and the SoftMuon (bottom right) algorithm

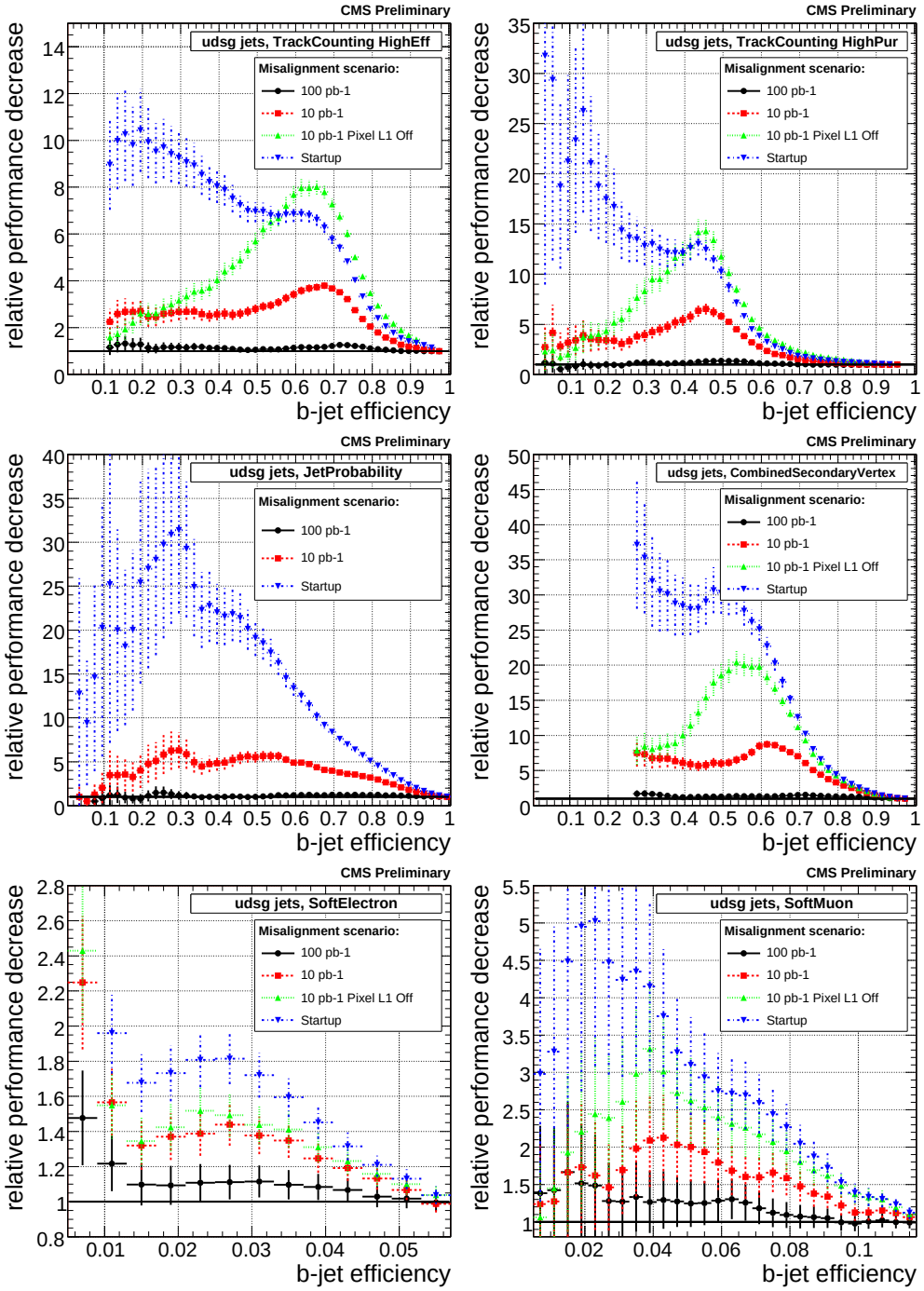


Figure 5.20: Relative performance decrease in udsg mistagging rate versus b-tagging efficiency for the TrackCounting high efficiency (top left) and high purity (top right), the JetProbability (middle left), the CombinedSecondaryVertex (middle right), the SoftElectron (bottom left) and the SoftMuon (bottom right) algorithm

developed and implemented in the CMS offline software framework CMSSW.

This algorithm was especially designed to be suitable for the early data taking period and it should provide a most robust behaviour with respect to misalignment. Robust in this sense means that its performance should not strongly depend on the degree of misalignment since the alignment constants would be almost unknown during the initial phase of CMS. Thus, they would have to be determined from data iteratively with an increasing amount of collected charged particle tracks as described above. Furthermore, the algorithm should remain as simple as possible in order to minimise various sources of systematic uncertainties as well as to avoid the usage of more complex methods for discriminator calculation. Among these methods are e.g. likelihood ratios or neural nets which combine several input variables and therefore require a careful training on large data samples that do not exist at the beginning of data taking. Hence, this simple b-tagger only takes one secondary vertex property which is directly used as discriminator of the algorithm.

5.4.1 SimpleSecondaryVertex Candidates

The maximum b-tagging efficiency of such a simple secondary vertex based algorithm is limited by the efficiency to reconstruct a secondary vertex inside a particular jet, as stated in table 5.4. More detailed studies concerning this topic, like the impact of different configuration settings for the secondary vertex finding algorithms or the use of a different algorithm itself are presented in section 5.4.2 and 5.4.3, respectively.

The presence of a successfully reconstructed secondary vertex inside a jet is already a powerful discriminator to distinguish b-jets from background jets originating from gluons or the decay of u-, d-, s- and c-hadrons. In addition to the mere reconstruction of such a displaced vertex, one variable has to be chosen for direct use as discriminator of the new algorithm. Therefore, four promising discriminator candidates were selected and the performances of the resulting algorithms were determined. These four discriminating variables are the flight distance value and significance, both calculated in two and three dimensions. The resulting performances are presented in figure 5.21 for the 10 pb^{-1} scenario in case of c- and udsg-jets. For a better comparison of their behaviour under misalignment, the corresponding ratio distributions which represent the relative decrease of the algorithm performance with respect to an ideally aligned tracking detector are given in figure 5.22. A complete set of these plots for all misalignment scenarios can be found in appendix C.7.

As expected, the flight distance significance candidates provide a better performance over the whole b-tagging efficiency range than the algorithms which use the value of the flight distance directly. All four tagger candidates behave likewise with respect to misalignment. Their mistagging rate is slightly degraded in case of udsg-jets for most of the different misalignment scenarios. Only in case of startup tracker conditions, the relative performance decrease reaches a factor of 5 when the significance candidates are considered.

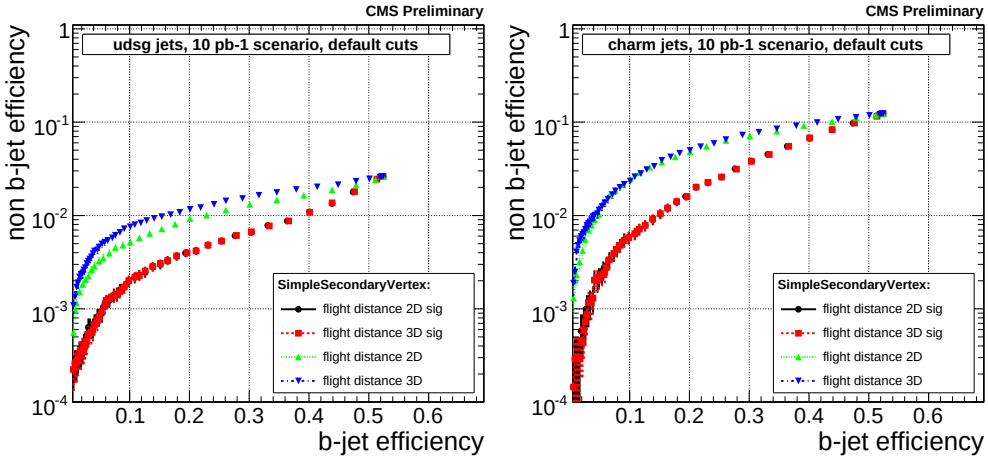


Figure 5.21: b-jet efficiency versus non b-jet efficiency for the four SimpleSecondaryVertex candidates in case of the 10 pb^{-1} misalignment scenario. The distributions are presented for light flavour and gluon (left) and charm (right) jets

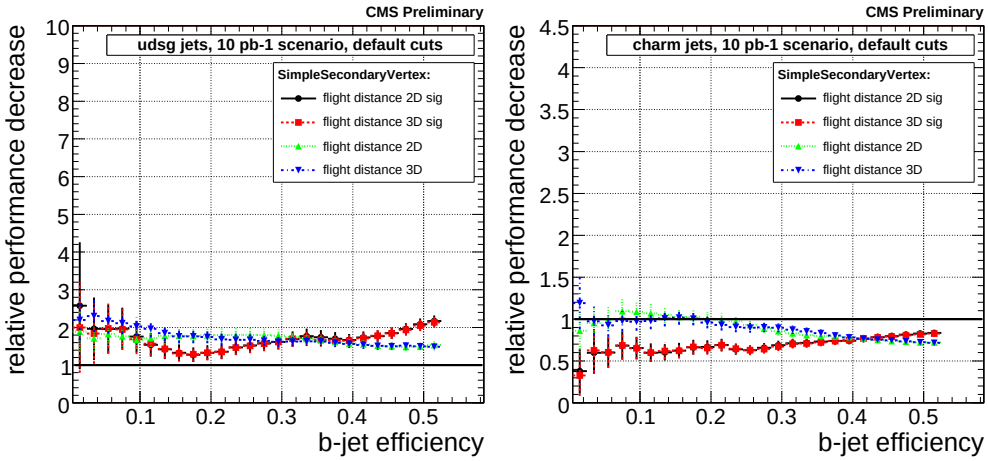


Figure 5.22: Relative performance decrease for the four SimpleSecondaryVertex candidates in case of the 10 pb^{-1} misalignment scenario. The distributions are presented for light flavour and gluon (left) and charm (right) jets

An exceptional feature appears when looking at the charm mistagging rate. It appears as if the charm rejection is improved over almost the whole b-tagging efficiency range, as soon as misalignment is applied. This circumstance seems to be non-intuitive, since the purity of the selected jets should decrease with an increasing error on the measurement. Nevertheless, this behaviour can be explained by looking at the default cuts used for the secondary vertex reconstruction, as already presented in table 5.3. Decays of charm hadrons are similar to decays of b-hadrons due to their non-negligible lifetime. This as well leads

to displaced secondary vertices, but with smaller flight distances and lower multiplicities of charged particle tracks associated to them. Due to their closer proximity to the primary interaction point, it is much harder to measure these smaller separations between primary and secondary vertex as it would be in case of b-hadrons. Therefore, misalignment has a larger impact on the vertex reconstruction efficiency in c-jets. This effect is furthermore increased when the cut of 3 on the minimal flight distance significance is applied. As a result, less secondary vertices are reconstructed and therefore the charm mistagging rate is decreased when misalignment is applied.

Figure 5.23 presents the secondary vertex reconstruction efficiency as a function of the flight distance, calculated in three dimensions for b- and c-jets. In order to cancel out effects which could arise without a cut on the maximum transverse flight distance of 2.5 cm, this acceptance limit was applied for both simulated and reconstructed secondary vertices which were used in order to calculate these efficiencies.

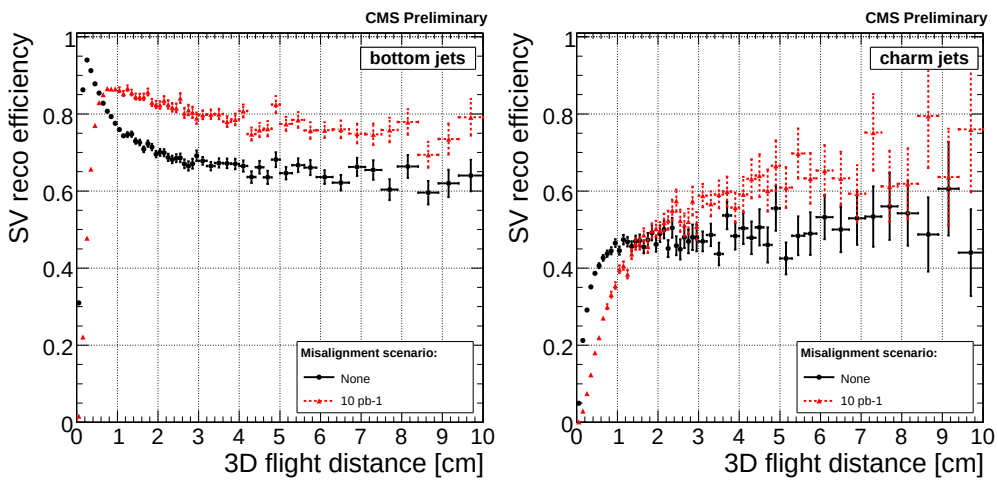


Figure 5.23: Secondary vertex reconstruction efficiency versus flight distance, calculated in three dimensions, for bottom (left) and charm (right) jets. The distributions are presented for an ideally aligned silicon detector as well as for the 10 pb^{-1} misalignment scenario

The mentioned discrepancy in reconstructing secondary vertices inside b- and c-jets can be explained by the different drop-off behaviour of the reconstruction efficiencies at small flight distance values, close to the primary interaction point:

- Secondary vertices in c-jets are harder to reconstruct close to the primary vertex. b-hadrons in contradiction, have larger track multiplicities and probably a tertiary vertex which originates from the subsequent decay of an associated charm hadron. These facts, together with the higher mass and therefore a larger opening angle between assigned tracks, lead to higher probabilities to reconstruct such a secondary vertex.

- The width of the drop-off region with respect to the efficiency plateau in case of the depicted 10 pb^{-1} scenario is about double the size when compared to a perfectly aligned tracker. The width roughly scales proportionally with the measurement errors of the track parameters.

- The efficiency plateau itself is shifted towards higher efficiency values in case of misalignment. This is due to larger errors of the track parameters and thus an increased probability of finding tracks which intersect at a common vertex.

Finally, by taking the secondary vertex reconstruction efficiency, the shorter c-hadron lifetime and the exponentially decreasing probability for having vertices at a distinct flight distance into account, the overall secondary vertex reconstruction efficiency inside c-jets is more strongly reduced than it is in case of b-jets.

It has been assumed that the locations of reconstructed and simulated secondary vertices are in good agreement. This is verified by the pull distributions of the flight distance, calculated in three dimensions, as presented in figure 5.24. The flight distance pull is defined as deviation between reconstructed and simulated flight distance, divided by the measurement error on the reconstructed value. The simulated flight distance refers to the distance between simulated primary and secondary vertex, whereas the reconstructed flight distance corresponds to their reconstructed counterparts. The pull distributions include as well misreconstructed and fake vertices which cause the asymmetric and non-gaussian tails observed. The core of the distribution, however, represents a Gaussian centred around zero with a width of approximately one. Thus, in most of the cases of the secondary vertex reconstruction, the correct vertex is found and the errors of the vertex fits are propagated correctly.

Basically, the flight distance value taggers are a bit less sensitive to misalignment. Therefore, the first choice for a most robust b-tagging algorithm during LHC startup should be made for either the flight distance value in two or three dimensions. On the other hand, their absolute performance is much worse. As a trade-off between performance and robustness, the three dimensional flight distance significance was chosen to act as the default discriminator for the newly introduced secondary vertex only based algorithm. This tagger is referred to as SimpleSecondaryVertex algorithm in the following sections as well as in the official CMSSW code. Due to the modularity of the framework, any of the four presented alternatives can be chosen, depending on the needs and desires of the analysing physicists.

The b-tagging performance of the SimpleSecondaryVertex algorithm does almost not rely on the degree of misalignment applied to the silicon tracking detector. Although secondary vertex based variables are not that sensitive to misalignment, a much stronger dependency would have been expected. An explanation for this extensive robustness is given by the behaviour of the flight distance significance distributions with respect to misalignment. The significance and therefore the discriminator is shifted in a similar way, if misalignment is applied. This can be seen by comparing the different flavour tagging efficiencies

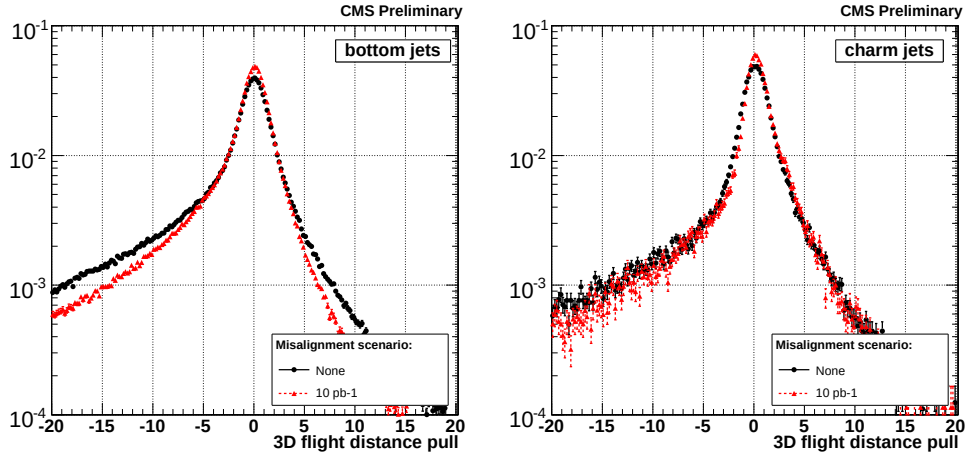


Figure 5.24: Pull distributions for the flight distance, calculated in three dimensions in case of bottom (left) and charm (right) jets. The distributions are presented for an ideally aligned silicon detector as well as for the 10 pb^{-1} misalignment scenario

with respect to the discriminator cut, as presented in figure 5.25 and figure 5.26 for the SimpleSecondaryVertex and the JetProbability algorithm, respectively.

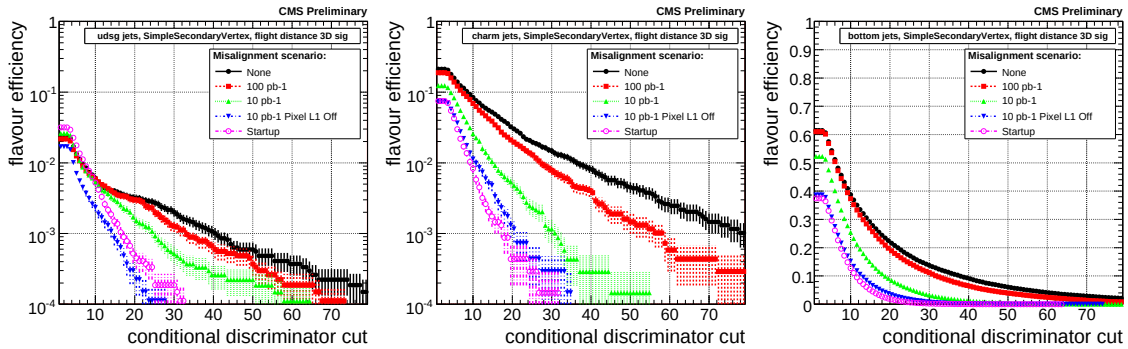


Figure 5.25: Flavour tagging efficiencies with respect to the discriminator cut for the different misalignment scenarios. The distributions are presented for usdg (left), charm (middle) and bottom (right) jets in case of the SimpleSecondaryVertex algorithm

In order to keep a certain flavour tagging efficiency with increasing misalignment, the cut on the discriminator has to be shifted accordingly. A constant b-tagging efficiency, for example, of about 20% in case of the SimpleSecondaryVertex algorithm requires a shift of the discriminator cut from 20 to about 10 when going from the ideally aligned tracker to the 10 pb^{-1} scenario. The same shift can be applied as well in case of light flavour jets to keep a constant mistagging rate of about 0.4% and in case of charm jets where it results in a mistagging rate of about 4%. The necessary shift of the discriminator cut for the

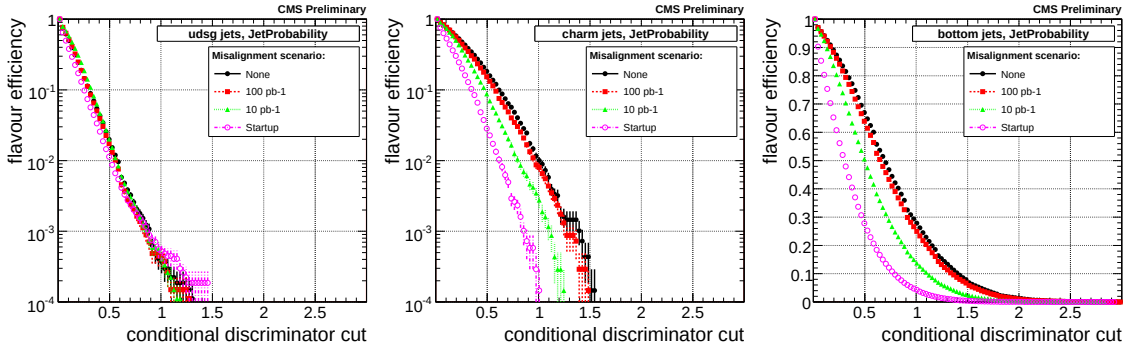


Figure 5.26: Flavour tagging efficiencies with respect to the discriminator cut for the different misalignment scenarios. The distributions are presented for usdg (left), charm (middle) and bottom (right) jets in case of the JetProbability algorithm

JetProbability algorithm for example, completely varies with respect to the different jet flavours in order to keep constant tagging efficiencies. Thus, making this algorithm more prone to misalignment. The distributions for the other b-tagging algorithms are attached in appendix C.8.

5.4.2 Secondary Vertex Reconstruction Optimisations

The secondary vertex reconstruction algorithms have several tunable parameters as already introduced in section 5.2.2. The vertex candidates have to pass several quality cuts in order to reduce the amount of fake or mismeasured vertices. In principle, all of these parameters can affect the secondary vertex reconstruction of either real or fake vertices in a different way. The default cuts which were used, represent a sensible default derived from an ideally aligned tracking detector. The tracking precision is degraded as soon as misalignment is introduced. Therefore, the associated errors have to be inflated to compensate for this issue during reconstruction. The larger errors lead to reduced significances as discussed in case of e.g. impact parameters above. These measurement errors are then propagated to the primary and secondary vertex fitting procedure and thus directly affect the reconstruction efficiencies of the primary and secondary vertex finders.

Higher track uncertainties can influence the vertex finding abilities in several ways. Two vertices in close proximity, for example, may not be distinguishable anymore. This can either result in a reduced overall reconstruction efficiency or as well lead to a merging of these two vertices. If a secondary and tertiary vertex are merged, the flight distance of the resulting vertex is pulled towards larger values. Thus, there would be no negative impact on the b-tagging performance in such a case. Otherwise, if tracks from the primary vertex are added, the resulting secondary vertex might be too close to the primary interaction point in order to be distinguished properly. In addition, the fraction of fake secondary

vertices is increased as well. Due to the increased track uncertainties, it becomes more likely that outlier tracks form such a fake vertex close to the primary interaction point of the event. These fake vertices are typically rejected due to plausibility cuts on the flight distance and flight distance significance within the used algorithms.

Misalignment might affect the reconstruction of real secondary vertices differently, as they exist in b- and c-jets, as well as of fake vertices which are present in light flavour and gluon jets. In order to optimize the robustness of the SimpleSecondaryVertex algorithm, several cuts of the secondary vertex finder algorithm have been changed and their impact on the secondary vertex reconstruction efficiency has been studied. It was expected to either find a set of cuts which would optimize the overall vertex finding efficiency whatever scenario is applied, or which would change the balance between reconstructed real and fake secondary vertices and therefore would result in a change of the performance distributions of the algorithm itself.

Several track selection and secondary vertex quality cuts have been altered. In fact, a complete parameter scan was not feasible, since the whole b-tagging chain including the time-consuming vertex finding procedure would have to be re-run for all parameter sets and misalignment scenarios. Therefore, some promising cuts have been changed in order to see a general trend in terms of algorithm performance or reconstruction efficiency (the statement in brackets corresponds to the default setting, if available):

1. The maximum allowed fraction of shared tracks between primary and secondary vertex was changed to 1.0 and therefore this cut was disabled (0.65)
2. 3D flight distance > 0.1 cm (instead of 2D > 0.01 cm)
3. 3D flight distance significance > 3.0 (instead of 2D > 3.0)
4. No cut on minimal 2D flight distance and significance
5. 2D flight distance > 0.1 cm (> 0.01 cm)
6. χ^2 of the track fit < 20 (< 5)
7. Additional track selection cut 2D IP significance > 0.0
8. Additional track selection cut 2D IP significance > 2.0
9. Additional track selection cut 3D IP significance > 2.0
10. Additional track selection cut 3D IP significance > 4.0

The impact of the different sets of cuts on the expected b-tagging performance in case of the SimpleSecondaryVertex algorithm is presented in figure 5.27 for the 10 pb^{-1} misalignment scenario in case of charm and light flavour jets. The distributions for all other scenarios

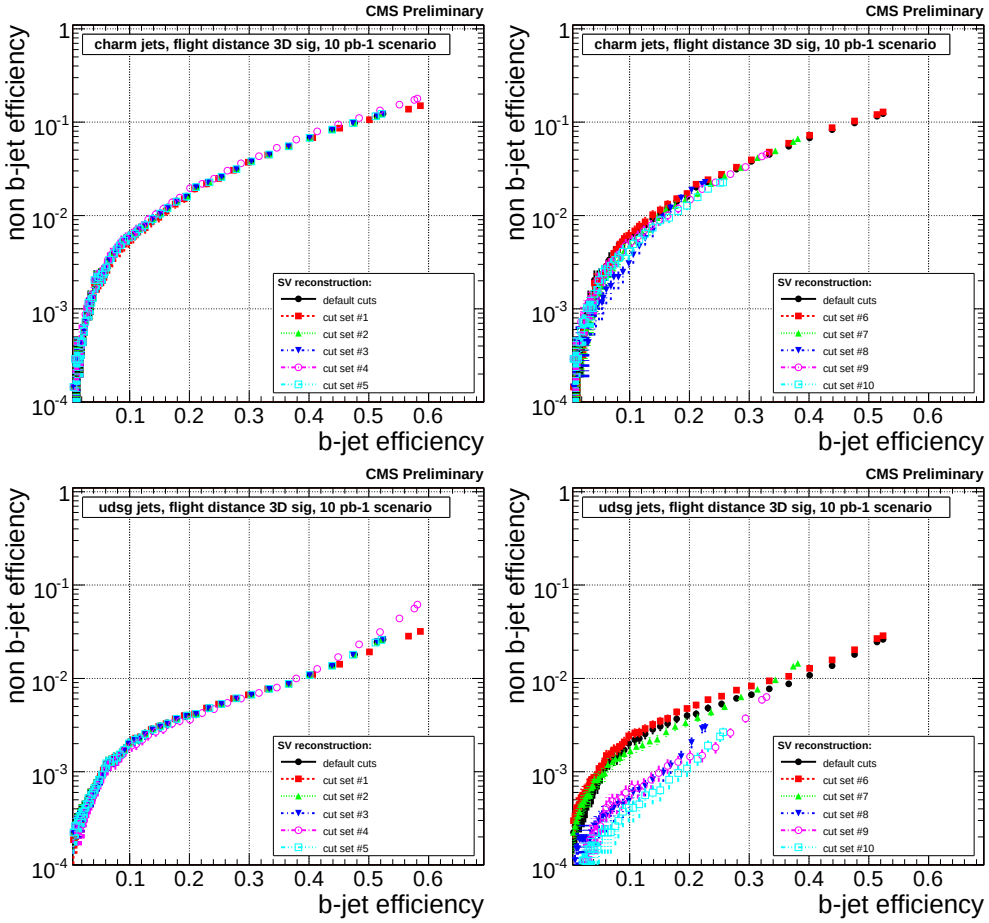


Figure 5.27: Performance distributions for the SimpleSecondaryVertex algorithm with respect to the different cut sets presented for the 10 pb^{-1} misalignment scenario in case of charm (top) and udsg (bottom) jets

are presented in appendix C.9. The results of these cut variation tests were not conclusive, especially with respect to a change in the overall secondary vertex reconstruction efficiency which could be observed in both directions by introducing the different track selection cuts. This change of reconstruction efficiency is accompanied by a change in absolute performance, as expected due to the increased or decreased rejection of fake tracks. However, an impact on the robustness of the algorithm with respect to misalignment could not be observed in any of the tests.

For every change in the secondary vertex finding efficiency in case of the perfectly aligned detector, a similar change for all the different misalignment scenarios could be observed. Thus, no evidence for a more robust behaviour with respect to misalignment was found. The c-jet rejection improvement under misalignment was observed for all the different cut sets described above, as well.

The secondary vertex reconstruction efficiency is mainly affected by the ability of the vertex finder algorithm to distinguish separated vertices. This ability in turn, directly relies on the compatibility of tracks with a displaced vertex and therefore on the resolution provided by the track reconstruction algorithm itself. Such a dependency can not reasonably be eliminated. Thus, the default cuts were already sensible even with detector misalignment introduced.

5.4.3 Secondary Vertex Finder Variations

By default, the Adaptive Vertex Finder (AVF) algorithm is used for secondary vertex reconstruction as mentioned at the beginning of this chapter. Since CMSSW features the Trimmed Kalman Vertex Finder (TKVF) as well, its impact on the b-tagging performance of the SimpleSecondaryVertex algorithm was investigated, too. However, no significant performance change could be observed regardless of what misalignment scenario was applied. This expertise is exemplarily summarised in figure 5.28 for the 10 pb^{-1} scenario.

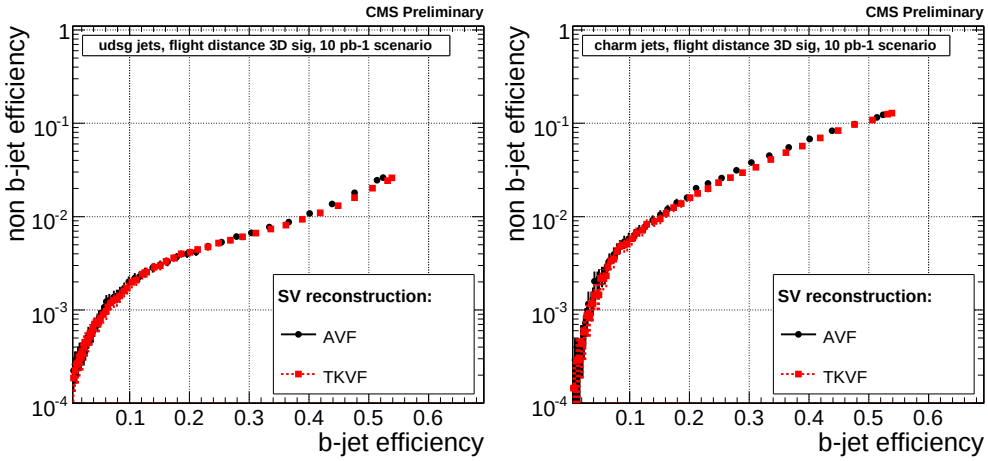


Figure 5.28: Performance of the SimpleSecondaryVertex algorithm with respect to the used vertex finder algorithm. The distributions are presented for light flavour (left) and charm (right) jets in case of the 10 pb^{-1} misalignment scenario

5.4.4 Alignment Position Error Variations

The track finding efficiency would decrease as soon as misalignment is applied to the tracking detector. In order to keep an efficiency which is comparable to the ideally aligned tracker, the alignment position error (APE) is added in quadrature to the track hit error for all the different tracker components. In case of Monte Carlo based misalignment studies, the APE is chosen in an idealised way in order to keep the tracking efficiencies at a constant

level. However, with real data, the optimal APE can not be directly determined and has to be estimated based on a "best guess".

Therefore, the uncertainty on the APE was simulated and its impact on the b-tagging performance was determined. Starting from the ideal value, the APE was scaled by a factor of 1/3, 1/2, 2 and 3 in order to reflect these possibly non-optimal estimates. Figure 5.29 shows the resulting performance distributions for the TrackCounting (high purity) and the SimpleSecondaryVertex algorithm in case of light flavour and gluon jets. The remaining distributions for all other lifetime-based b-tagging algorithms can be found in appendix C.10. They are exemplarily presented for the 10 pb^{-1} scenario, since the performance impact stays similar with respect to other degrees of misalignment.

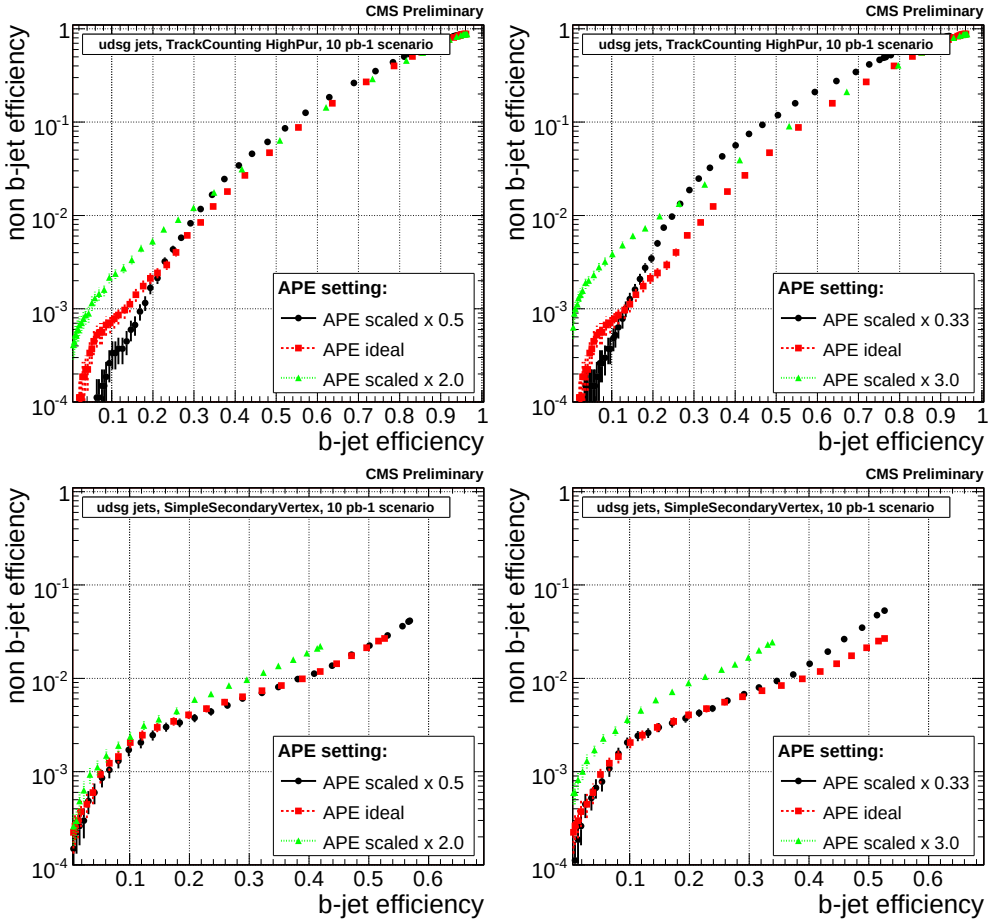


Figure 5.29: b-tagging performance for the high purity TrackCounting (top) and SimpleSecondaryVertex (bottom) algorithm with respect to scaled APEs by a factor of 1/3, 1/2, 2 and 3. The distributions are presented for light flavour jets in case of the 10 pb^{-1} misalignment scenario

Due to the misalignment, the tracking detector components are severely displaced with

respect to their ideal position. In order to pick up even hits that would otherwise be considered as incompatible during the track fit, the APE is used to inflate the hit errors which are then propagated down to the final track fit, the primary and secondary vertex positions and as well to the track impact parameters. Therefore, a change of the APE results directly in a change of impact parameter significances. Particularly, algorithms which rely on PDFs like the JetProbability and CombinedSecondaryVertex are affected the most.

The APE can be either chosen too small or too large with respect to the assumed ideal value:

- If the APE is chosen too small, the probability to pick up a hit during track finding is decreased since the hit compatibility measure with respect to the track is reduced. This leads to more rejected valid tracks due to a too large χ^2 of the track fit and therefore to a reduced tracking efficiency after all. At the same time, fake tracks are rejected more strongly as well. This results in case of the TrackCounting algorithm in a slight performance decrease in the high efficiency region due to missed tracks with high impact parameters which originate from the decay of b-hadrons. But the reduced fake rate, on the other hand, leads to a performance increase in the high purity region.

  When scaling the APE down a factor of 3, the first effect dominates and the overall light flavour rejection gets worse. A similar effect can be seen for the SimpleSecondaryVertex algorithm as well. Due to the reduction of badly measured tracks the separation between primary and secondary vertex becomes clearer and the overall vertex finding efficiency slightly increases. Since a secondary vertex can be formed from only two compatible tracks, missing a few tracks from a typically high-multiplicity b-decay does not have a large negative impact as it does for the TrackCounting algorithm.

  These results indicate that there is still room for optimisation of the track selection cuts which are used for b-tagging with respect to the misalignment which is applied. This has to be studied by using real data, since the measurement errors are effectively unknown. However, a further decrease of the APE results in a decreased b-tagging efficiency and therefore a decreased algorithm performance after all.

- If the APE is chosen too large, hits are more likely to be picked up by the seeding and pattern recognition and consequently the track fake rate increases. A too large APE leads to the opposite effects of what was discussed above. The larger fake rate leads to a light flavour mistagging rate for the TrackCounting algorithm which is increased in the high purity region. The secondary vertex finding efficiency in b-jets is decreased due to fake tracks which reduce the separation between the primary and a possible secondary vertex. Also, the probability to form a fake vertex is increased.

  These effects lead to both an increase in light flavour mistagging rate as well as a decrease in b-jet tagging efficiency. However, the vertex finding typically still picks

up the correct vertex, assuming it is found. Therefore, the efficiency drop in the high purity region is not as large as for impact parameter based algorithms. Vertices which are found in the high purity region are usually clearly displaced and thus not that sensible to single fake tracks, whereas in contrast tracks with large impact parameters can be easily caused by bad or fake track reconstruction. An additional cause besides the already mentioned ones for an increased rate of fake tracks, are tracks with badly associated hits, mainly found in the pixel detector region where the hit density becomes extremely high, especially inside jets.

The expected uncertainty on the alignment position error bears significant effects with respect to b-tagging at CMS. It is essential to give a good estimate for the APE close to its optimal value. If the discrepancy stays within a factor of two, preferably smaller, the robustness of the simple b-tagging algorithms would still be assured even during startup tracker conditions.

5.4.5 Association Cone Size Variations

In order to associate charged particle tracks to a jet, a cone of $\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}$ is cast around the jet axis and all tracks located inside this cone are then selected. The cone size itself is a configurable parameter for the different b-tagging algorithms and is set by default to $\Delta R \leq 0.3$ in case of the JetProbability and CombinedSecondaryVertex tagger. All other algorithms use an association cone of $\Delta R \leq 0.5$.

It is expected that the size of the association cone leads to an impact on the achievable b-tagging performance of the different algorithms. In order to study this dependence, several cuts on the association cone have been tested for the SimpleSecondaryVertex algorithm. Figure 5.30 shows the expected performance for the 10 pb^{-1} scenario for a ΔR of 0.3, 0.4 and 0.5. The corresponding plots for the other misalignment settings are presented in appendix C.11.

The charm mistagging rate is almost not affected by the size of the cone, regardless which misalignment scenario is applied. In case of light flavour and gluon jets, a performance increase over almost the whole b-efficiency region can be observed when using smaller ΔR values. However, the overall secondary vertex reconstruction efficiency is slightly decreased due to fewer selected tracks which are associated to the jet. A dependency between cone size and robustness with respect to misalignment could not be observed and therefore, the default cone size for the SimpleSecondaryVertex algorithm was chosen to be $\Delta R \leq 0.3$ due to the improved performance. These conclusions coincide with results obtained in former studies as presented in [12].

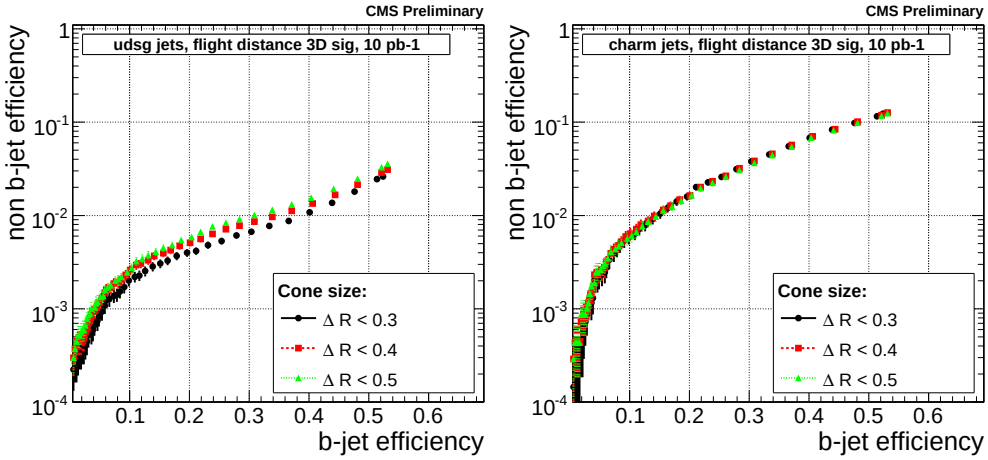


Figure 5.30: Performance of the SimpleSecondaryVertex algorithm with respect to the track association cone size. The distributions are presented for light flavour (left) and charm (right) jets in case of the 10 pb^{-1} misalignment scenario

5.4.6 Performance Comparison

The studies performed in this chapter led to the development of a simple and easy-to-understand b-tagging algorithm, which is intend to be used especially during the first period of data taking at the CMS experiment. It depends only slightly on the degree of silicon tracker misalignment although it provides a good discrimination power to distinguish bottom from charm, light flavour and gluon jets. The default secondary vertex and track acceptance parameters were already sensible with respect to the different misalignment scenarios and no evidence for a positive impact on the algorithm robustness was found. The same applies for the used association cone size, where no effect on the algorithm behaviour under misalignment could be identified.

Figure 5.31 and 5.32 presents the overall b-tagging performance and the relative performance decrease of the SimpleSecondaryVertex algorithm with respect to all other lifetime based algorithms which were studied here. The distributions are given for the 10 pb^{-1} and startup misalignment scenario which correspond to tracker conditions during the initial phase of data taking at the LHC. The remaining scenarios are presented in appendix C.12.

It can be seen that the performance of the SimpleSecondaryVertex algorithm is comparable to or even better than the results obtained with other taggers, especially with alignment conditions which correspond to first data. The overall achievable b-tagging efficiency, however, is limited by the secondary vertex reconstruction efficiency, as expected.

The alignment precision is strongly limited at the beginning of data taking, therefore the most important feature of this algorithm lies in its robustness with respect to misalignment. The SimpleSecondaryVertex algorithm therefore is best suited to be used during the initial

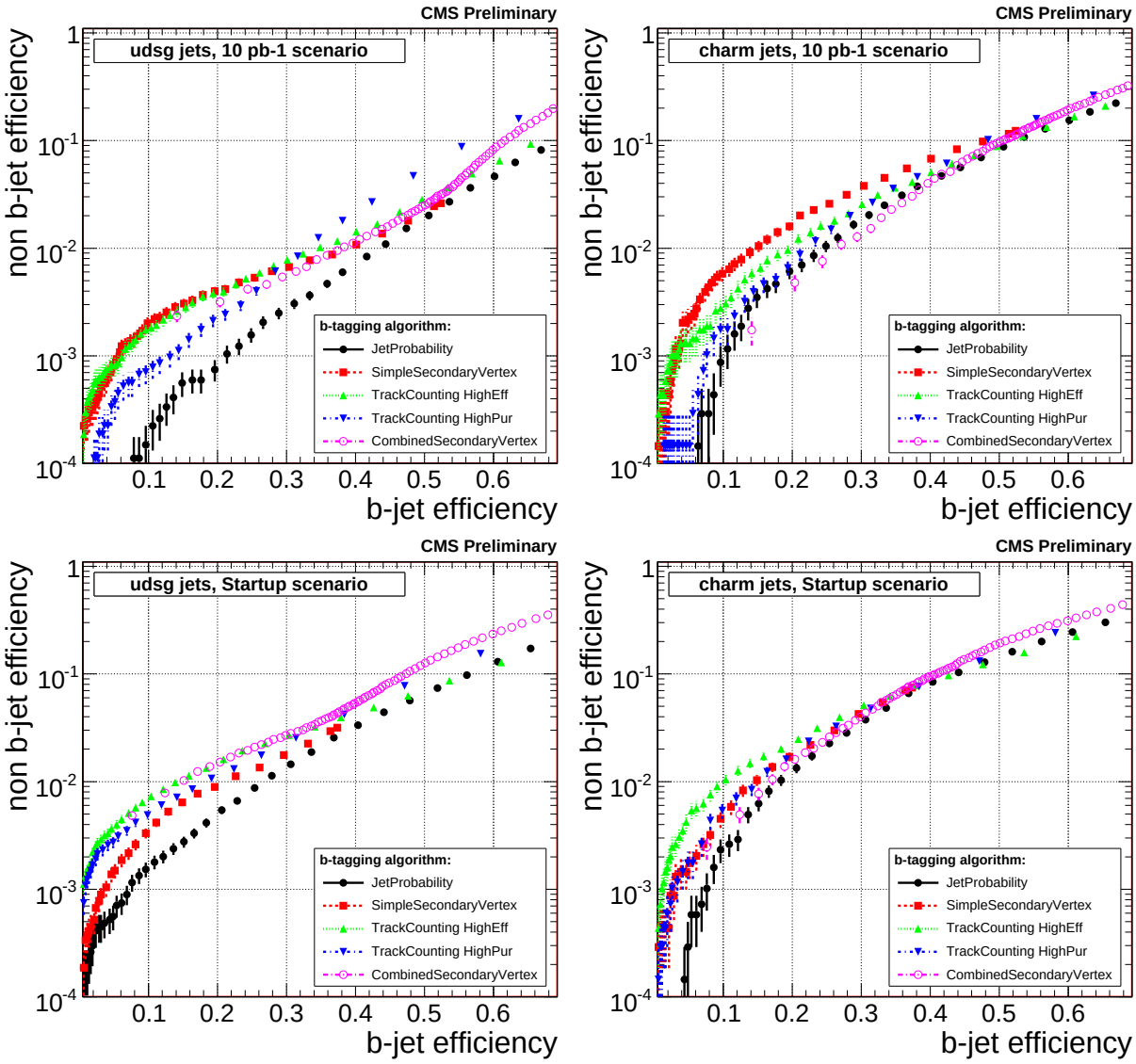


Figure 5.31: Performances of the lifetime based b-tagging algorithms presented for light flavour (left) and charm (right) jets in case of the 10 pb^{-1} (top) and startup (bottom) misalignment scenario

phase of the experiment.

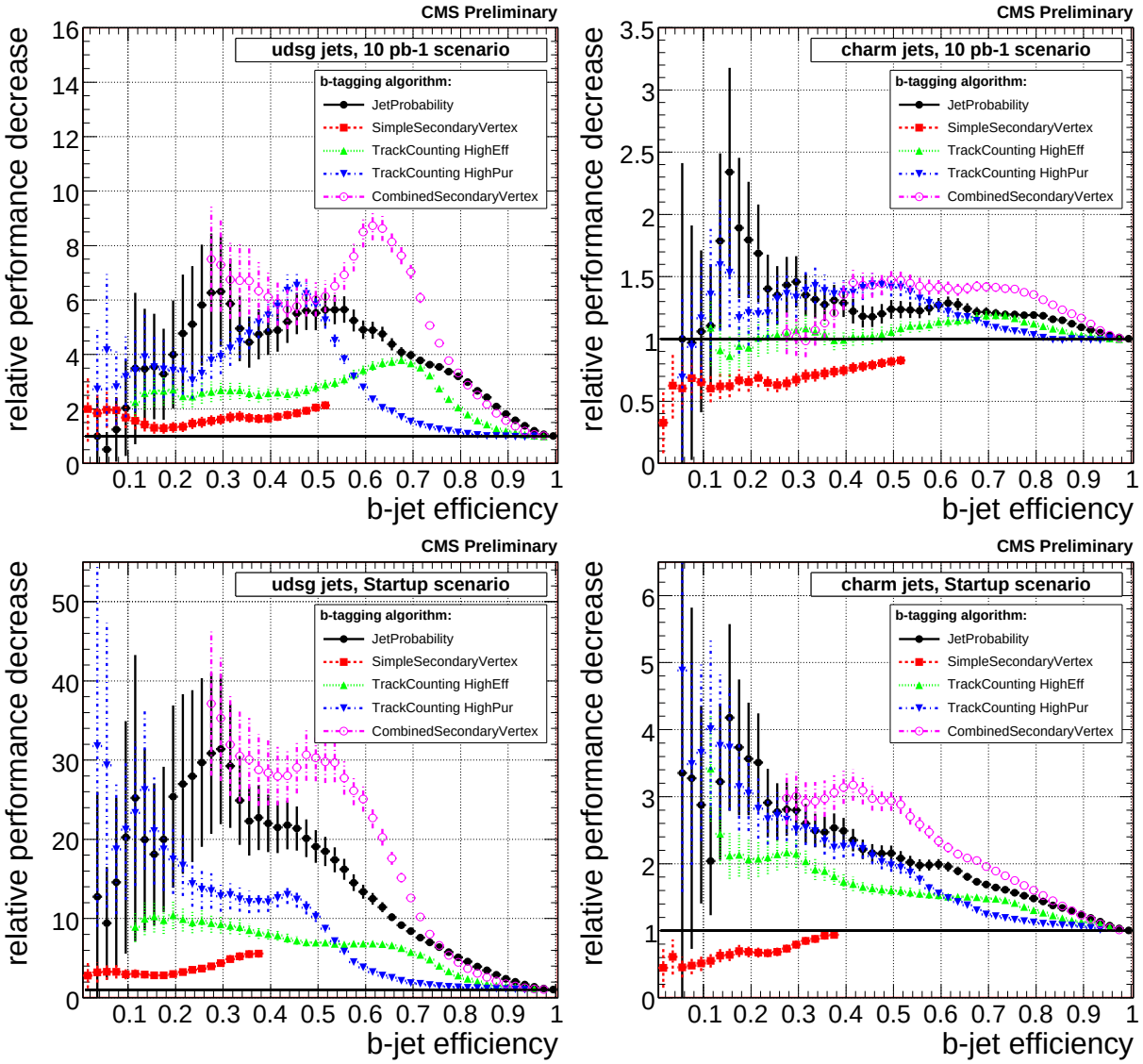


Figure 5.32: Relative performance decrease of the lifetime based b-tagging algorithms with respect to an ideally aligned tracker, presented for light flavour (left) and charm (right) jets in case of the 10 pb^{-1} (top) and startup (bottom) misalignment scenario

Chapter 6

Conclusion and Outlook

The startup of the Large Hadron Collider in summer 2008 will open a new energy range that will allow to study various undiscovered or yet unknown aspects of particle physics. Physicists hope to verify Standard Model parameters at this higher energy scale and to improve the measurement precision on already known parameters. Another defined goal of the LHC is to get evidence on electroweak symmetry breaking by discovering at least one Higgs boson or to find phenomena which are an indication on new physics beyond.

The studies presented in this thesis were performed for the CMS experiment, one of the two general-purpose detectors at the LHC. The high collision rate of the Large Hadron Collider of about 40 MHz and its centre-of-mass energy of 14 TeV lead to roughly 15 Petabyte of recordable data for all experiments combined per year. This huge amount has to be accessed and analysed by numerous physicists all around the world in order to compete for possible innovative discoveries in this most interesting new range of physics. The requirements on the detector and computing hardware, the storage capacities and the involved software components are immense.

During the last few years, various functionality tests were performed in order to proof the readiness of important CMS software components and the involved Grid infrastructure for data taking at the LHC. These so-called Computing, Software and Analysis (CSA) challenges were used to exercise the complete computing chain at rates of 25, 50 and 100% of the requirements for data taking in 2008. The efforts made at CSA06, especially with respect to the German Tier-1 centre, GridKa, were described. During this exercise, various stability problems could be identified and the knowledge gained was used to improve software and hardware stability during subsequent tests. The last challenge performed in 2008 has shown that the computing and software infrastructure of the CMS experiment will be ready for the rapidly approaching data taking phase at the LHC.

Various physics searches performed at the Large Hadron Collider will contain a large number of particle jets in the final states of the recorded collision events. Among them are several measurements of parameters related to B-hadron and top quark decays, searches

for the Higgs boson or searches for new phenomena beyond the Standard Model in e.g. the SUSY sector. Therefore, algorithms for reconstructing and identifying these jets will become an important tool in order to determine such processes and thus, obtain measurable parameters for these new fields of particle physics.

The identification of jets which contain hadrons of up, down or strange quarks as well as of gluons is hardly possible since events with such jets in the final state provide almost identical event topologies. However, weakly decaying bottom and charm hadrons on the other side provide distinct properties, like their lifetime or large mass, which lead to measurable signatures in the detector. Another aspect is the decay to a lepton and its corresponding neutrino, although the branching fraction for such decays is limited to about 10% per lepton flavour.

These features are all exploited by various algorithms implemented in the CMS offline software framework, CMSSW. All studies presented in this thesis were performed on a sample of about 22 000 inclusive $t\bar{t}$-events which were generated and simulated with PYTHIA 6.4 and the Geant4 detector simulation package, respectively.

First of all, the expected b-tagging performance in case of an ideally aligned detector was determined for all available algorithms. It was observed that algorithms which rely on impact parameters of charged particle tracks, a displaced secondary decay vertex or a lepton in the decay chain, exhibit different behaviour with respect to the expected performance and the optimal operating range. The studies presented in this thesis were performed for the first time within CMSSW and therefore, as a side result, reference the current status of the b-tagging software and the involved algorithms at CMS.

All lifetime-based algorithms rely on the reconstruction of physics objects like charged particle tracks or secondary vertices inside a jet. Since their resolution, in turn, strongly depends on the known alignment position of the involved components of the silicon tracking detector, their behaviour under distinct degrees of misalignment was studied more closely. For this purpose, a set of misalignment scenarios were defined and the alignment precision of the silicon tracker corresponding to the expected precision after specific amounts of collected collision data, was simulated, the needed physics objects were re-reconstructed and the resulting b-tagging performance with respect to misalignment was determined.

It was determined that more complex algorithms are more sensitive to misalignment and the expected performance during the initial phase of data taking is decreased by more than one order of magnitude, depending on the misalignment scenario and the algorithm. In case of the CombinedSecondaryVertex algorithm, the performance decrease reaches a factor of 35 with respect to an ideally aligned detector in case of light flavour and gluon jets. However, it turns out that some secondary vertex based b-tagging observables are less affected by the degree of alignment precision when compared to track impact parameter based variables.

Therefore, a new algorithm was implemented in the CMS offline software framework, called

SimpleSecondaryVertex. This algorithm is now part of the official CMS reconstruction sequence which is run during the first reconstruction pass at the Tier-0, directly after data acquisition at CMS. The new b-tagging method relies solely on the reconstruction of a displaced secondary decay vertex inside the examined particle jet and one additional secondary vertex observable that can be directly used as discriminator to distinguish between b-jets, c-jets and light flavour background jets.

It was observed that the SimpleSecondaryVertex algorithm is almost insensitive on the degree of misalignment, nevertheless it provides a comparable discrimination power with respect to the other methods which were studied here. Since the alignment precision is strongly limited at the beginning of data taking, the SimpleSecondaryVertex algorithm therefore will be best suited to be used during the initial phase of CMS data taking at the LHC.

It will be necessary to reproduce the conclusions of this study with first real data, for instance the actual b-tagging performance hast to be measured with techniques that work on real collision events. Such techniques are currently studied at CMS. The SimpleSecondaryVertex algorithm should be validated quickly in order to allow for example, for potential early discovery of new physics due to the measurement of certain SUSY channels. In parallel, the tracking should reach its expected high precision as quickly as possible in order to achieve the expected performances. Important parameters, such as the Alignment Position Error require special attention during the first phase of LHC operation.

Appendix A

Configuration Files

A.1 Sample CMSSW Configuration File

The CMS software is based on different modules or plug-ins which have to be specified and configured in a CMSSW configuration file, if needed. Each module performs a specific task and its functionality is dynamically loaded via libraries into the main CMSSW executable, `cmsRun`. A sample configuration file is shown here:

```
#!/usr/bin/env cmsRun
process myCMSSWProcess = {

# 1. Source Configuration
source = PoolSource {
    untracked vstring fileNames = {"file:/path/to/my/inputfile.root"}
    untracked uint32 skipEvents = 0
}

# Number of events to process, -1 for all events in the source
untracked PSet maxEvents = {untracked int32 input = -1}

# 2. Plug-in Configuration
# Define module with specific parameters
module module1 = TestModule1 {
    PSet myModule1ConfigurationSet = {
        string parameter1   = "par1"
        bool parameter2     = true
        double parameter3   = 0.3
    }
```

```
}

# Include module with default values
include "/path/to/default/definition/of/module2.cfi"
replace TestModule2.parameter1 = false

# 3. Output Configuration
module output = PoolOutputModule {
    untracked string fileName = "file:/path/to/my/outputfile.root"
    untracked vstring outputCommands = {
        "drop *",
        "keep *_*_*_myCMSSWProcess"
    }
}

# 4. Job Configuration
# Define which modules to run and in which order
path p = { module1, module2 }

# Write output specified in output module
endpath outpath = { output }
}
```

Description of the different parts of the configuration file:

1. **Source Configuration:**
   Each CMSSW job reads events from the source module. These can be either events from an already processed and stored data sample out of a ROOT file or events directly from a Monte Carlo generator or the detector's DAQ during data taking.

2. **Plug-in Configuration:**
   CMSSW provides several kinds of modules which can be defined directly in the configuration file. Another possibility is including them from the main software repository with its default configuration values in the first place. They can then be changed with `replace` statements afterwards. So-called EDProducer modules read the data from the event and calculate or produce new information out of the given input observables in order to store these additionally in the output file. EDFilter modules decide if the processing of an event is continued or skipped due to a specified abort criterion. EDAnalyzer modules are normally meant to study several properties of an event and write the output for example to ROOT histograms. They are not allowed to add new data to the event content itself. EDLooper modules are able to loop over given input data and iteratively modify the EventSetup itself. These modules are for example used in case of track based alignment procedures.

3. **Output Configuration:**
   The output module stores the event data at the end of a typical CMSSW job. In this case, the `PoolOutputModule` is used which writes the event content to a standard CMSSW ROOT file. It can be configured via `drop` and `keep` statements to reduce the amount of data written to disk. In the given example, all data is dropped in the first place and only the newly produced data in the sample process `myCMSSWProcess` is kept for further processing. Thus, the needed storage capacity can be reduced significantly if the subsequent analyses only rely on the stored information.

4. **Job Configuration:**
   The `path` and `endpath` variables define the overall job behaviour. In this example, the job reads the event from the specified input file, sequentially executes `module1` and `module2` and stores the produced data in the output file at the end of the job. Additionally, CMSSW provides the possibility to use logic operations in the `path` statement such as `( module1 & module2 )` in order to execute the two modules in parallel.

A.2 Configuration of a b-Tagging and Misalignment Job

These sample configuration files can be used in order to run silicon tracker misalignment and all further steps required for b-tagging within CMSSW. At the end, a configuration fragment for an analysis job with the default b-tagging analysis framework is presented as well. They are compatible to the software version series `CMSSW_1_5_X`. In order to get proper results for the secondary vertex based b-tagging algorithms, a backport of newer code for the `secondaryVertexTagInfos` and the algorithm itself from e.g. `CMSSW_1_8_X` is needed as well. This code has to be adjusted to fit the proper C++ `DataFormat` requirements of `CMSSW_1_5_X`.

Sample Misalignment Job

```
#!/usr/bin/env cmsRun
process misalign = {

    # Source files, e.g. ttbar Monte Carlo events
    source = PoolSource {
        include "my_configs/data/my_ttbar_input_files.cfi"
        untracked uint32 skipEvents = 0
    }

    # Number of events to process
    untracked PSet maxEvents = {
        untracked int32 input = -1 # -1 for all events
    }

    # Message logger service
    include "FWCore/MessageLogger/data/MessageLogger.cfi"

    # Get True MC jet flavour for the btag analysis
    # algorithmic definition
    module mcJetFlavour = MCJetFlavour {
        InputTag jets             = iterativeCone5CaloJets
        PSet jetFlavourIdentifier = {
            string mcSource                 = "source"
            bool fillPartons                = true
            bool fillHeavyHadrons           = false
            bool fillLeptons                = false
            double coneSizeToAssociate      = 0.3
            bool physicsDefinition          = false
```

```
        bool rejectBCSplitting          = false
        vstring vetoFlavour             = { }
    }
}

# physics definition
module mcPhysJetFlavour = MCJetFlavour {
    InputTag jets             = iterativeCone5CaloJets
    PSet jetFlavourIdentifier = {
        string mcSource                 = "source"
        bool fillPartons                = true
        bool fillHeavyHadrons           = false
        bool fillLeptons                = false
        double coneSizeToAssociate      = 0.3
        bool physicsDefinition          = true
        bool rejectBCSplitting          = false
        vstring vetoFlavour             = { }
    }
}

sequence myMCJetFlavour = { mcJetFlavour, mcPhysJetFlavour }

# Misalignment, e.g. the 10 pb-1 scenario
include "Alignment/TrackerAlignment/data/10pbScenario.cff"
es_prefer MisalignedTracker = MisalignedTrackerESProducer {}

# Transient track builder
include "TrackingTools/TransientTrack/data/
                                  TransientTrackBuilder.cfi"

# Tracking
include "RecoTracker/Configuration/data/RecoTracker.cff"

# JetTracksAssociator
include "RecoBTau/JetTracksAssociator/data/jetTracksAssociator.cff"

# Primary Vertex Producer
include "RecoVertex/PrimaryVertexProducer/data/
                           OfflinePrimaryVerticesFromCTFTracks.cfi"

# Magnetic field
include "MagneticField/Engine/data/volumeBasedMagneticField.cfi"
```

```
    # CMS Geometry
    include "Geometry/CMSCommonData/data/cmsIdealGeometryXML.cfi"
    include "Geometry/CaloEventSetup/data/CaloGeometry.cfi"
    include "Geometry/CaloEventSetup/data/CaloTopology.cfi"
    include "Geometry/MuonNumbering/data/muonNumberingInitialization.cfi"

    # Muons
    include "RecoMuon/Configuration/data/RecoMuon.cff"

    # Electrons
    include "RecoBTag/SoftLepton/data/btagSoftElectrons.cfi"

    sequence myMisalignment = { ckftracks,
                                offlinePrimaryVerticesFromCTFTracks,
                                jetTracksAssociator,
                                muonreco_plus_isolation,
                                btagSoftElectrons }

    # Output configuration
    module output = PoolOutputModule {
        untracked string fileName = 'file:my_misalignment_output.root'
        untracked vstring outputCommands = {
            "drop *",
            "keep *_iterativeCone5CaloJets_*_*",
            "keep *_globalMuons_*_*",
            "keep *_mcJetFlavour_*_*",
            "keep *_mcPhysJetFlavour_*_*",
            "keep *_jetTracksAssociator_*_*",
            "keep *_btagSoftElectrons_*_*",
            "keep *_offlinePrimaryVerticesFromCTFTracks_*_*",
            "keep *_ctfWithMaterialTracks_*_*"
        }
    }

    # Which modules to run
    path p = { myMCJetFlavour, myMisalignment }

    # Write output
    endpath outpath = { output }
}
```

Sample b-Tagging Job

```
#!/usr/bin/env cmsRun
process btag = {

    # Source files, e.g. my_misalignment_output.root
    source = PoolSource {
        include "my_configs/data/my_misaligned_input_files.cfi"
        untracked uint32 skipEvents = 0
    }

    # Number of events to process
    untracked PSet maxEvents = {
        untracked int32 input = -1 # -1 for all events
    }

    # Message logger service
    include "FWCore/MessageLogger/data/MessageLogger.cfi"

    # Transient track builder
    include "TrackingTools/TransientTrack/data/
                                    TransientTrackBuilder.cfi"

    # Magnetic field
    include "MagneticField/Engine/data/volumeBasedMagneticField.cfi"

    # CMS Geometry
    include "Geometry/CMSCommonData/data/cmsIdealGeometryXML.cfi"
    include "Geometry/CaloEventSetup/data/CaloGeometry.cfi"
    include "Geometry/CaloEventSetup/data/CaloTopology.cfi"
    include "Geometry/MuonNumbering/data/muonNumberingInitialization.cfi"

    # b-Tagging configuration for all algorithms
    # Produce the IP and SV TagInfos
    include "RecoBTag/ImpactParameter/data/impactParameter.cff"
    include "RecoBTag/SecondaryVertex/data/secondaryVertex.cff"
    include "RecoBTag/TrackProbability/data/
                                 trackProbabilityFakeCondFile.cfi"

    # Produce the SoftLepton TagInfos
    include "RecoBTag/SoftLepton/data/softElectronTagInfos.cfi"
```

```
    include "RecoBTag/SoftLepton/data/softMuonTagInfos.cfi"

    sequence myBTagging = { impactParameterTagInfos,
                            secondaryVertexTagInfos,
                            softElectronTagInfos,
                            softMuonTagInfos }

    # Output configuration
    module output = PoolOutputModule {
        untracked string fileName = 'file:my_btag_output.root'
        untracked vstring outputCommands = {
            "drop *",
            "keep *_iterativeCone5CaloJets_*_*",
            "keep *_ctfWithMaterialTracks_*_*",
            "keep *_TrackRefitter_*_*",
            "keep *_jetTracksAssociator_*_*",
            "keep *_offlinePrimaryVerticesFromCTFTracks_*_*",
            "keep *_mcJetFlavour_*_*",
            "keep *_mcPhysJetFlavour_*_*",
            "keep *_*_*_btag"
        }
    }

    # Which modules to run
    path p = { myBTagging }

    # Write output
    endpath outpath = { output }
}
```

Sample Analysis Job

```
#!/usr/bin/env cmsRun
process analysis = {

    # Source files, e.g. my_btag_output.root
    source = PoolSource {
        include "my_configs/data/my_btag_input_files.cfi"
        untracked uint32 skipEvents = 0
    }

    # Number of events to process
```

```
untracked PSet maxEvents = {
    untracked int32 input = -1 # -1 for all events
}

# Message logger service
include "FWCore/MessageLogger/data/MessageLogger.cfi"

# Transient track builder
include "TrackingTools/TransientTrack/data/
                                    TransientTrackBuilder.cfi"

# Magnetic field
include "MagneticField/Engine/data/volumeBasedMagneticField.cfi"

# CMS Geometry
include "Geometry/CMSCommonData/data/cmsIdealGeometryXML.cfi"
include "Geometry/CaloEventSetup/data/CaloGeometry.cfi"
include "Geometry/CaloEventSetup/data/CaloTopology.cfi"
include "Geometry/MuonNumbering/data/muonNumberingInitialization.cfi"

# Setup b-tagging analysis framework
# Include b-tagging algorithms to produce the JetTags
include "RecoBTag/ImpactParameter/data/impactParameter.cff"
include "RecoBTag/TrackProbability/data/
                                     trackProbabilityFakeCondFile.cfi"
include "RecoBTag/SoftLepton/data/softLepton.cff"
include "RecoBTag/SecondaryVertex/data/secondaryVertex.cff"

# Setup the conditions database for e.g. the CombinedSecondaryVertex
# This can be a online database, or a database file
include "CondCore/DBCommon/data/CondDBSetup.cfi"
es_source BTauMVAJetTagComputerRecord = PoolDBESSource {
    using CondDBSetup
    string connect = "sqlite_file:/path/to/database_file.db"
    untracked string catalog = "file:/path/to/catalog_file.xml"
    string timetype = "runnumber"
    VPSet toGet = {
        {
    string record   = "BTauGenericMVAJetTagComputerRcd"
    string tag      = "MVAJetTags_CMSSW_1_5_4"
        }
    }
}
```

```
# Define several blocks for simpler usage in the module
# definitions below, several parameters control the resulting
# analysis plot, e.g. number of bins, ranges, etc.
# Common block for all taggers
block bTagAnalysisCommon = {
    include "RecoBTag/Analysis/data/bTagAnalysis.cff"
    bool fastMC = true
    InputTag jetMCSrc = mcJetFlavour
    double effBConst = 0.5
    int32  nBinEffPur  = 400
    double startEffPur = 0.005
    double endEffPur   = 1.005
}

# Individual blocks
include "RecoBTag/SecondaryVertex/data/
                                combinedSecondaryVertexCommon.cfi"
block bTagAnalysisCombinedSV = {
    vstring jetTagModuleLabel = { "combinedSecondaryVertexBJetTags" }
    PSet combinedSV = {
        using combinedSecondaryVertexCommon
    }
    bool useTagInfo = true
    string tagInfo = "CombinedSV"
    double discriminatorStart = -0.1
    double discriminatorEnd   = 1.1
}

block bTagAnalysisSimpleSV = {
    vstring jetTagModuleLabel = { "simpleSecondaryVertexBJetTags" }
    PSet combinedSV = {
        using combinedSecondaryVertexCommon
    }
    bool useTagInfo = true
    string tagInfo = "CombinedSV"
    double discriminatorStart = -0.1
    double discriminatorEnd   = 10.0
}

block bTagAnalysisTrackCounting = {
    vstring jetTagModuleLabel = { "trackCountingHighEffBJetTags",
                                  "trackCountingHighPurBJetTags" }
```

```
    bool useTagInfo = true
    string tagInfo = "TrackIP"
    double discriminatorStart = -10.0
    double discriminatorEnd   = 25.0
}

block bTagAnalysisTrackProbability = {
    vstring jetTagModuleLabel = { "jetProbabilityBJetTags" }
    bool useTagInfo = true
    string tagInfo = "TrackIP"
    double discriminatorStart = 0.0
    double discriminatorEnd   = 3.0
}

block bTagAnalysisSoftLepton = {
    string tagInfo = "SoftLepton"
    vstring jetTagModuleLabel = { "softElectronBJetTags",
                                  "softMuonBJetTags" }
    bool useTagInfo = false
    double discriminatorStart = -0.2
    double discriminatorEnd   =  1.2
}

# Module definitions using the block statements from above
module bTagCombinedSVJetTagAnalysis = BTagPerformanceAnalyzer {
    using bTagAnalysisCommon
    using bTagAnalysisCombinedSV
    string rootfile = "bTagCombinedSVAnalysis.root"
}

module bTagSimpleSVJetTagAnalysis = BTagPerformanceAnalyzer {
    using bTagAnalysisCommon
    using bTagAnalysisSimpleSV
    string rootfile = "bTagSimpleSVAnalysis.root"
}

module bTagTrackIPAnalysisTC = BTagPerformanceAnalyzer {
    using bTagAnalysisCommon
    using bTagAnalysisTrackCounting
    string rootfile = "bTagTrackCountingAnalysis.root"
}

module bTagTrackIPAnalysisTP = BTagPerformanceAnalyzer {
```

```
        using bTagAnalysisCommon
        using bTagAnalysisTrackProbability
        string rootfile = "bTagTrackProbabilityAnalysis.root"
    }

    module bTagSoftLeptonAnalysis = BTagPerformanceAnalyzer {
        using bTagAnalysisCommon
        using bTagAnalysisSoftLepton
        string rootfile = "bTagSoftLeptonAnalysis.root"
    }

    # Sequence to produce the JetTags and run the Analysis
    # afterwards
    sequence myAnalysis = { combinedSecondaryVertexBJetTags,
                            simpleSecondaryVertexBJetTags,
                            jetProbabilityBJetTags,
                            trackCountingHighEffBJetTags,
                            trackCountingHighPurBJetTags,
                            softElectronBJetTags,
                            softMuonBJetTags,
                            bTagCombinedSVJetTagAnalysis,
                            bTagSimpleSVJetTagAnalysis,
                            bTagTrackIPAnalysisTC,
                            bTagTrackIPAnalysisTP,
                            bTagSoftLeptonAnalysis }

    # Which modules to run
    path p = { myAnalysis }

}
```

Appendix B

Used Data Sample

B.1 Inclusive $t\bar{t}$ - Events

For the studies presented in this thesis, a private sample of about 22 000 inclusive $t\bar{t}$-events has been produced. The event generation was performed using PYTHIA 6.4 as described in chapter 3.4.2 followed by the Geant4 detector simulation. The default reconstruction sequence as well as all further reprocessing steps due to the misalignment studies were run with the CMS software version `CMSSW_1_5_4`. The CombinedSecondaryVertex b-tagging algorithm was backported from the `CMSSW 1_8_X` software series. The following PYTHIA parameters have been used in the event generation step:

```
# Settings for the underlying event
MSTJ(11)=3     ! Choice of the fragmentation function
MSTJ(22)=2     ! Decay those unstable particles
PARJ(71)=10.   ! For which ctau = 10 mm
MSTP(2) =1     ! Which order running alphaS
MSTP(33)=0     ! No K factors in hard cross sections
MSTP(51)=7     ! Structure function chosen, 7 = CTEQ5L
MSTP(81)=1     ! Multiple parton interactions, 1 = Pythia default
MSTP(82)=4     ! Defines the multi-parton model
MSTU(21)=1     ! Check on possible errors during program execution
PARP(82)=1.9409! Pt cut-off for multi-parton interactions
PARP(89)=1960. ! Sqrts for which PARP82 is set
PARP(83)=0.5   ! Multiple interactions: matter distribution parameter
PARP(84)=0.4   ! Multiple interactions: matter distribution parameter
PARP(90)=0.16  ! Multiple interactions: rescaling power
PARP(67)=2.5   ! Amount of initial-state radiation
PARP(85)=1.0   ! Gluon production mechanism in MI
```

```
PARP(86)=1.0   ! Gluon production mechanism in MI
PARP(62)=1.25  ! Cut-off for space-like parton shower
PARP(64)=0.2   ! Factor for space-like parton shower
MSTP(91)=1     ! Gaussian primordial kt distribution in hadron
PARP(91)=2.1   ! Width of the kt distribution
PARP(93)=15.0  ! Upper cut-off for this kt distribution

# Settings for the ttbar process
MSEL      = 0     ! User defined processes
MSUB(81)  = 1     ! qqbar to QQbar
MSUB(82)  = 1     ! gg to QQbar
MSTP(7)   = 6     ! Flavour = top
PMAS(6,1) = 175.  ! Top quark mass
```

More information about the parameter meaning can be found in [27].

Appendix C

Additional Figures

C.1 b-Tagging Input Variables

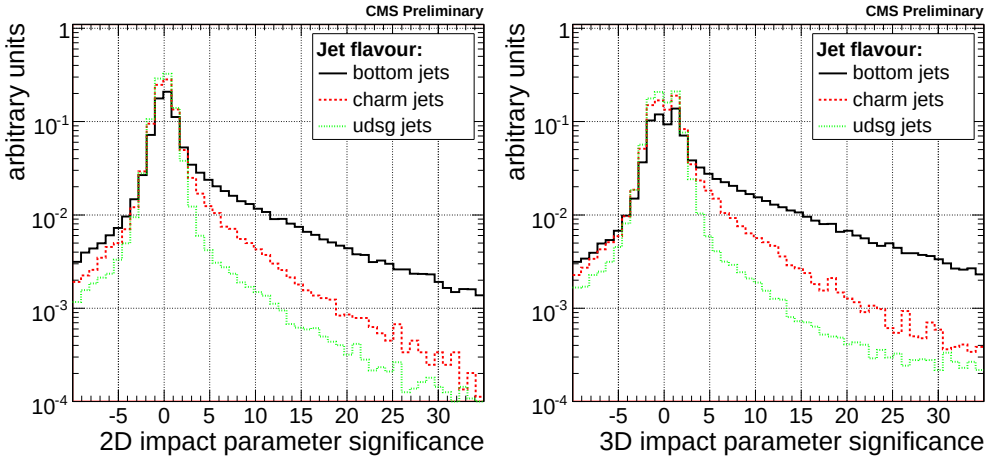


Figure C.1: Impact parameter significance calculated in the transverse plane (left) and in three dimensions (right). The black line shows the distribution in case of b-jets whereas red and green depicts c- and light flavour jets, respectively

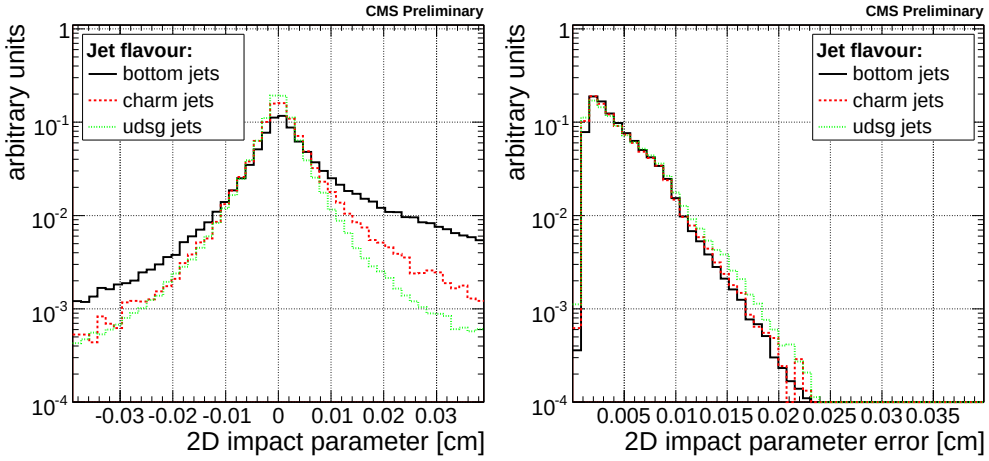


Figure C.2: Impact parameter value (left) and its measurement error (right) calculated in the transverse plane. The black line shows the distribution in case of b-jets whereas red and green depicts c- and light flavour jets, respectively

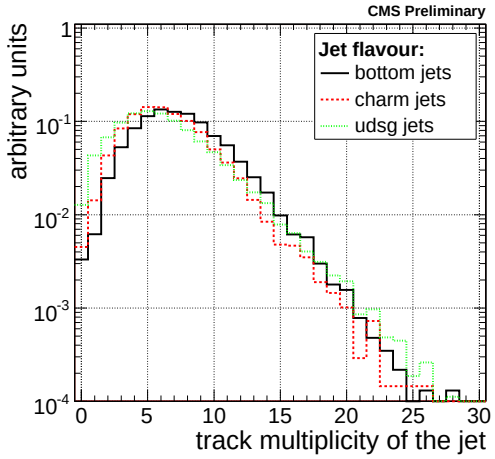


Figure C.3: Multiplicity of charged particle tracks associated to a jet. The black line shows the distribution in case of b-jets whereas red and green depicts c- and light flavour jets, respectively

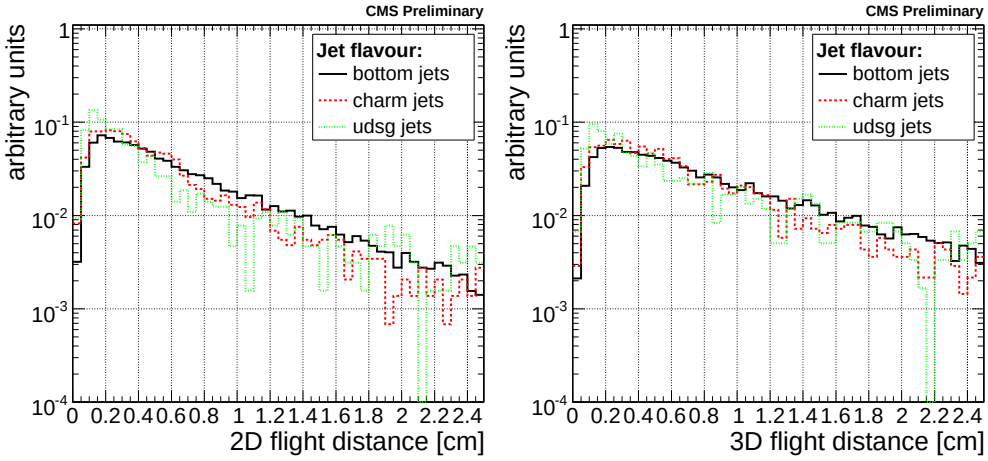


Figure C.4: Flight distance value between primary and secondary vertex in both, two (left) and three (right) dimensions. The black line shows the distribution in case of b-jets whereas red and green depicts c- and light flavour jets, respectively

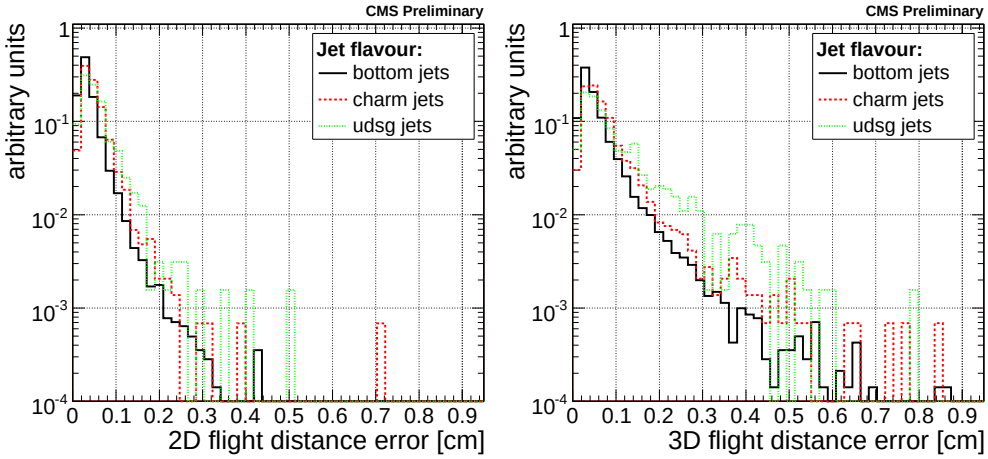


Figure C.5: Flight distance measurement error in both, two (left) and three (right) dimensions. The black line shows the distribution in case of b-jets whereas red and green depicts c- and light flavour jets, respectively

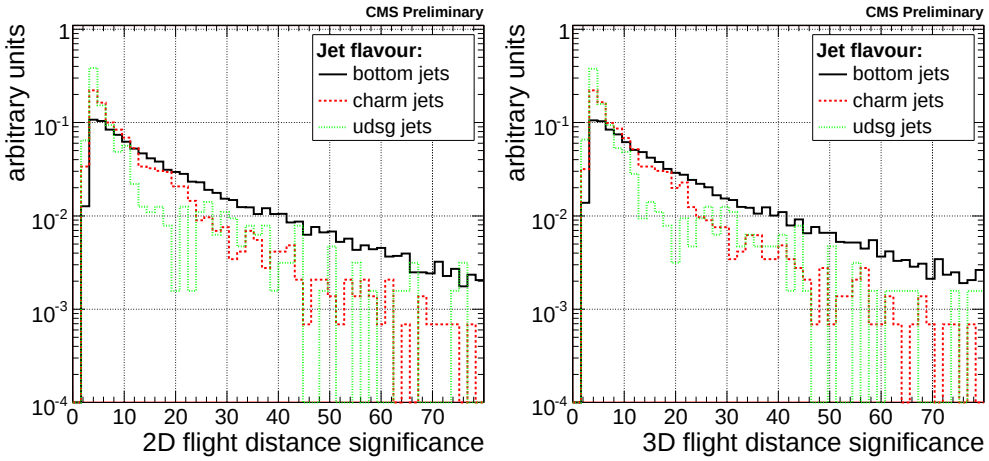


Figure C.6: Flight distance significance in both, two (left) and three (right) dimensions. The black line shows the distribution in case of b-jets whereas red and green depicts c- and light flavour jets, respectively

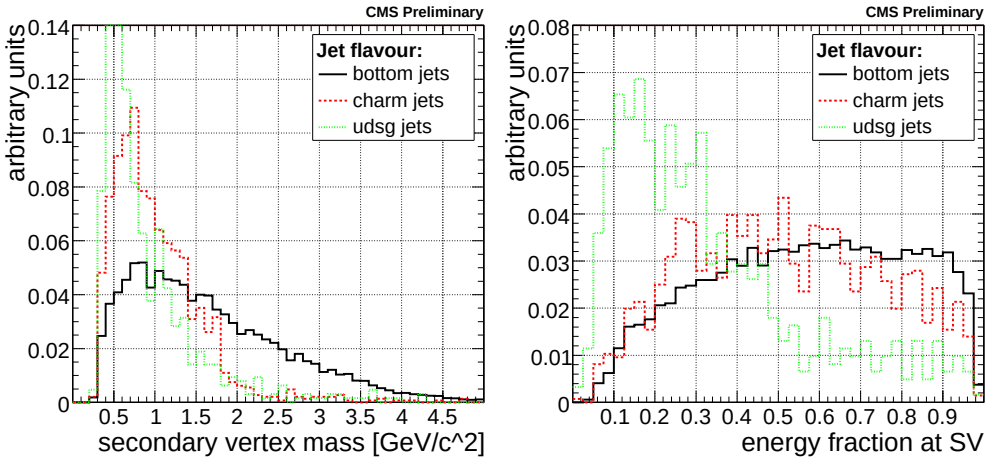


Figure C.7: Invariant mass (left) of charged particles associated to the secondary vertex. Energy fraction (right) of charged particles associated to the secondary vertex with respect to all particles associated to the jet. The black line shows the distribution in case of b-jets whereas red and green depicts c- and light flavour jets, respectively

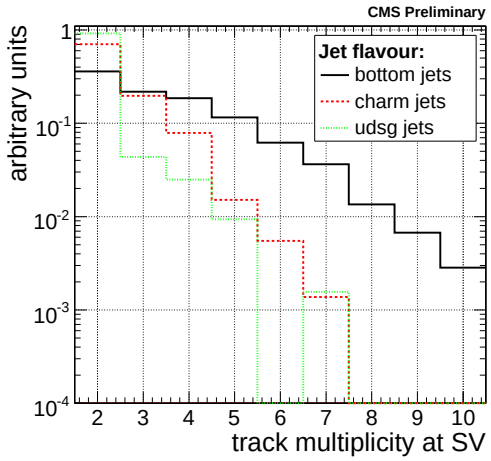


Figure C.8: Multiplicity of charged particle tracks associated to the secondary vertex. The black line shows the distribution in case of b-jets whereas red and green depicts c- and light flavour jets, respectively

C.2 b-Tagging Discriminator Distributions

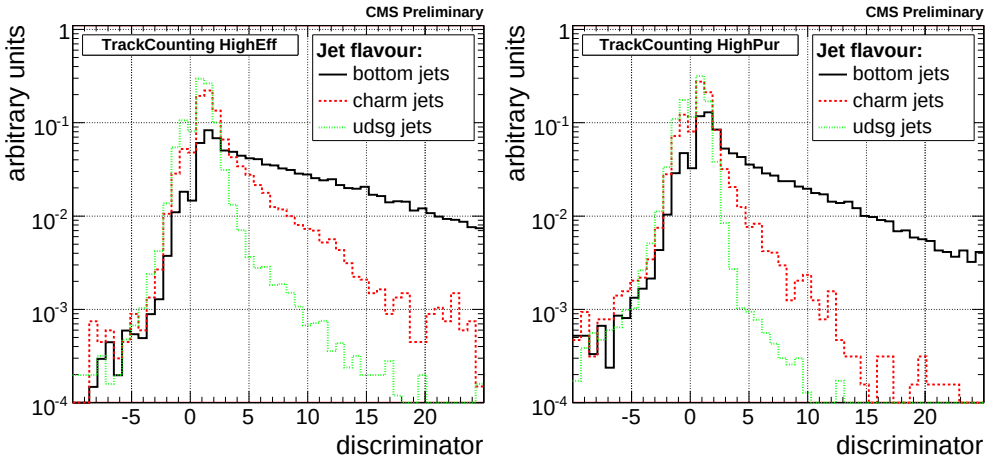


Figure C.9: Discriminator output for the TrackCounting algorithm using the second (high efficiency, left plot) and the third (high purity, right plot) track ordered by impact parameter significance as discriminator. The black line shows the distribution in case of b-jets whereas red and green depicts c- and light flavour jets including gluons, respectively

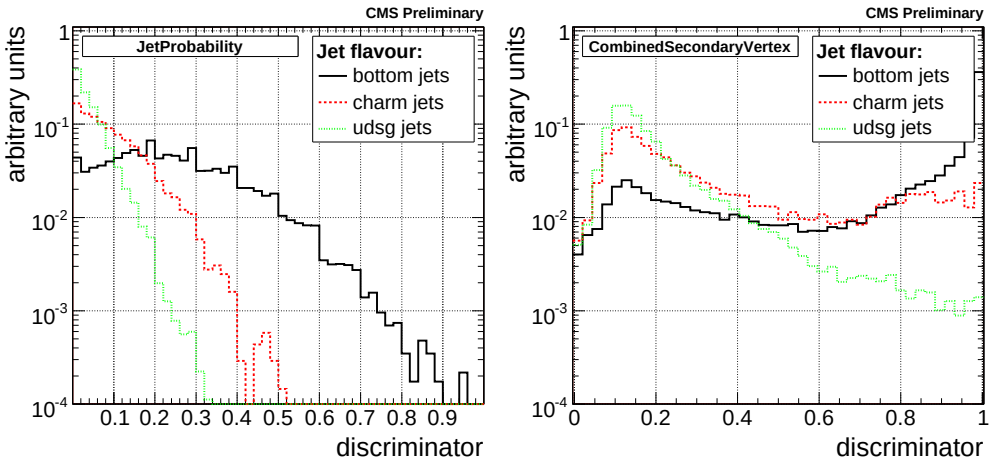


Figure C.10: Discriminator output for the JetProbability (left) and CombinedSecondaryVertex (right) algorithm. The black line shows the distribution in case of b-jets whereas red and green depicts c- and light flavour jets including gluons, respectively

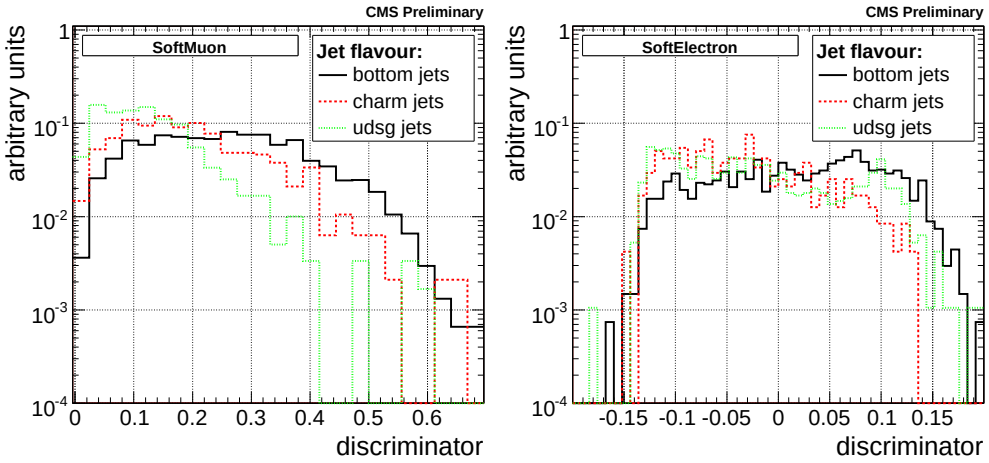


Figure C.11: Discriminator output for the SoftMuon (left) and SoftElectron (right) algorithm. The black line shows the distribution in case of b-jets whereas red and green depicts c- and light flavour jets including gluons, respectively

C.3 b-Tagging Performance Distributions

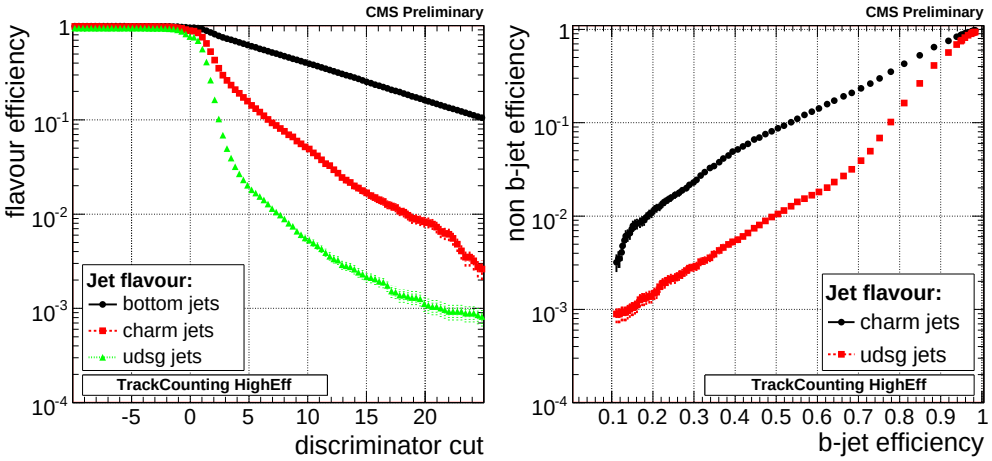


Figure C.12: The left plot shows the tagging efficiencies for the different jet flavours with respect to the discriminator cut and the right plot the b-tagging performance, both for the TrackCounting (high efficiency) algorithm.

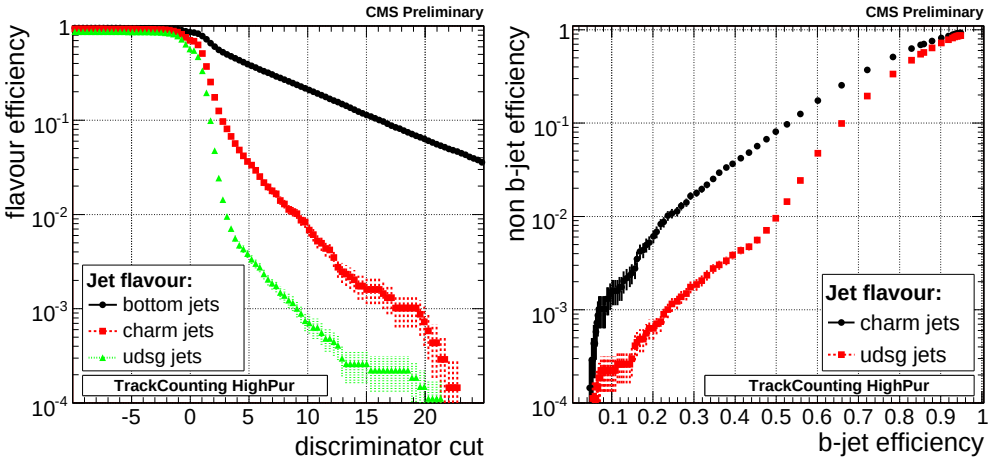


Figure C.13: The left plot shows the tagging efficiencies for the different jet flavours with respect to the discriminator cut and the right plot the b-tagging performance, both for the TrackCounting (high purity) algorithm.

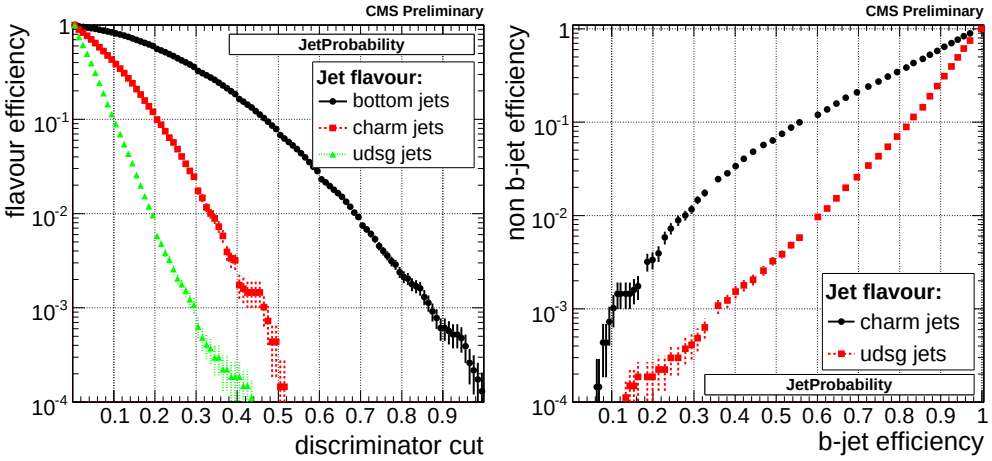


Figure C.14: The left plot shows the tagging efficiencies for the different jet flavours with respect to the discriminator cut and the right plot the b-tagging performance, both for the JetProbability algorithm.

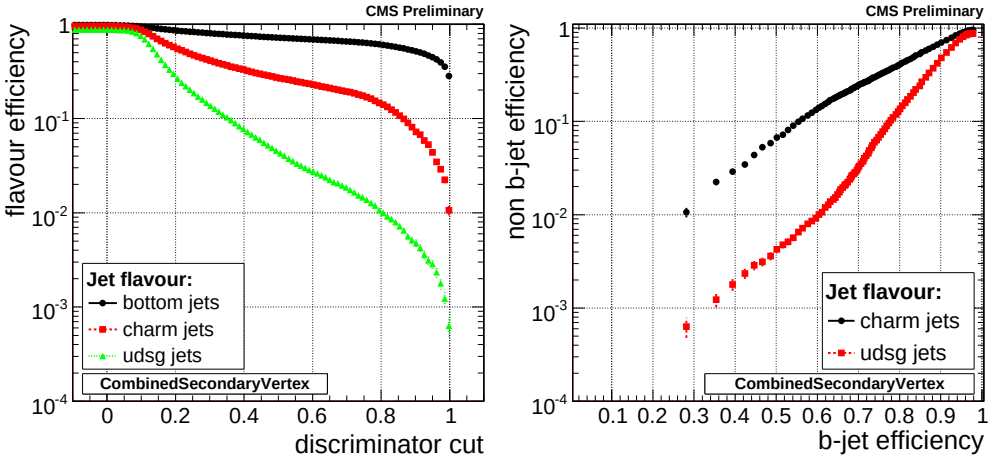


Figure C.15: The left plot shows the tagging efficiencies for the different jet flavours with respect to the discriminator cut and the right plot the b-tagging performance, both for the CombinedSecondaryVertex algorithm.

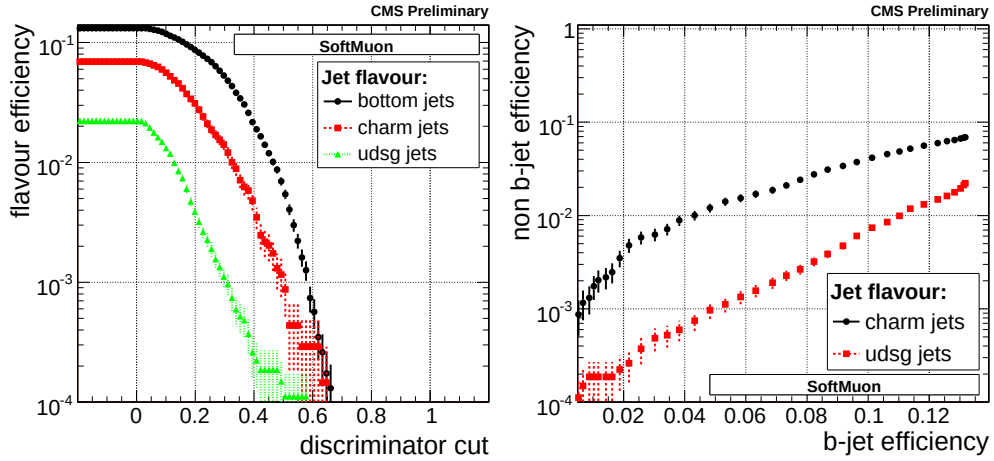


Figure C.16: The left plot shows the tagging efficiencies for the different jet flavours with respect to the discriminator cut and the right plot the b-tagging performance, both for the SoftMuon algorithm.

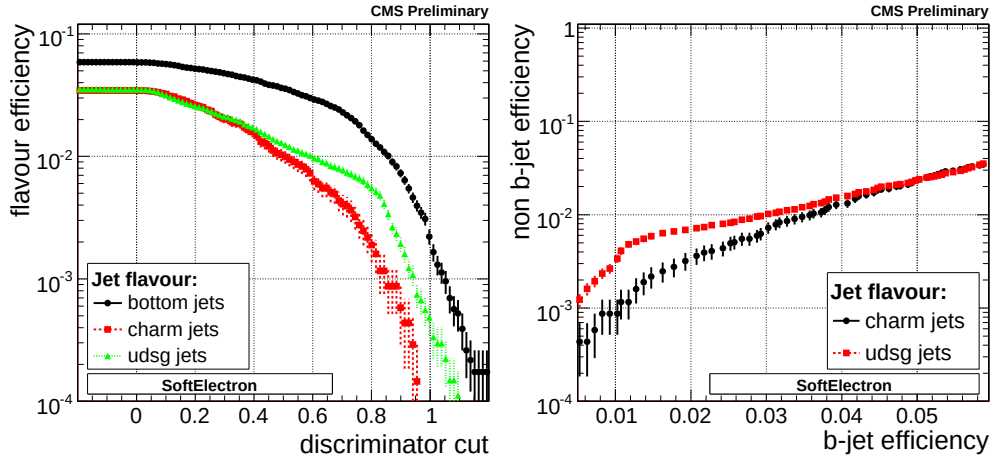


Figure C.17: The left plot shows the tagging efficiencies for the different jet flavours with respect to the discriminator cut and the right plot the b-tagging performance, both for the SoftElectron algorithm.

C.4 Track Refit Versus Track Re-Reconstruction

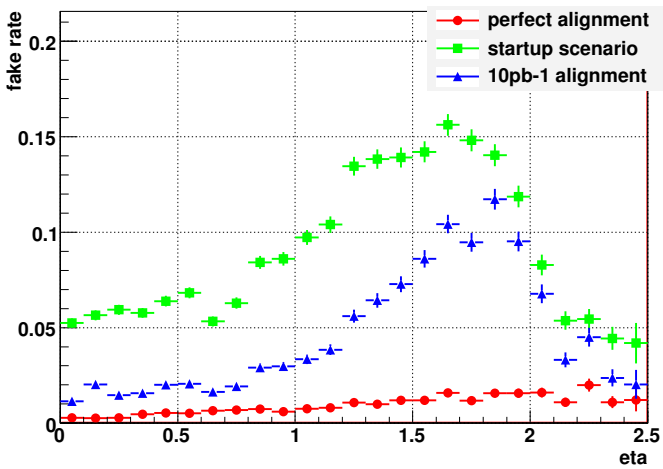


Figure C.18: Fake rate as a function of the pseudorapidity for tracks associated to jets which pass the b-tagging selection cuts presented in table 5.1

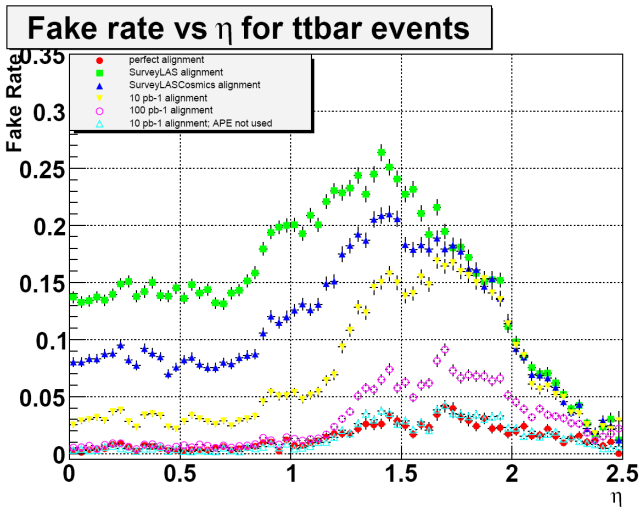


Figure C.19: Fake rate as a function of the pseudorapidity for all tracks in $t\bar{t}$-events [77]. The blue curve denoted as SurveyLASCosmics alignment corresponds to the startup scenario presented in the other plots

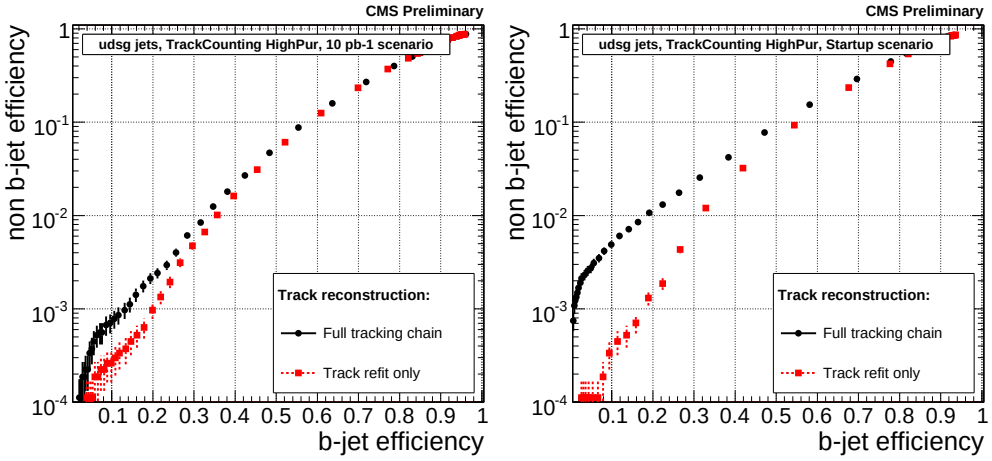


Figure C.20: Performance of the TrackCounting (high purity) algorithm for the refit-only and the full track reconstruction approach in case of the 10 pb^{-1} (left) and startup (right) misalignment scenario for light flavour and gluon jets

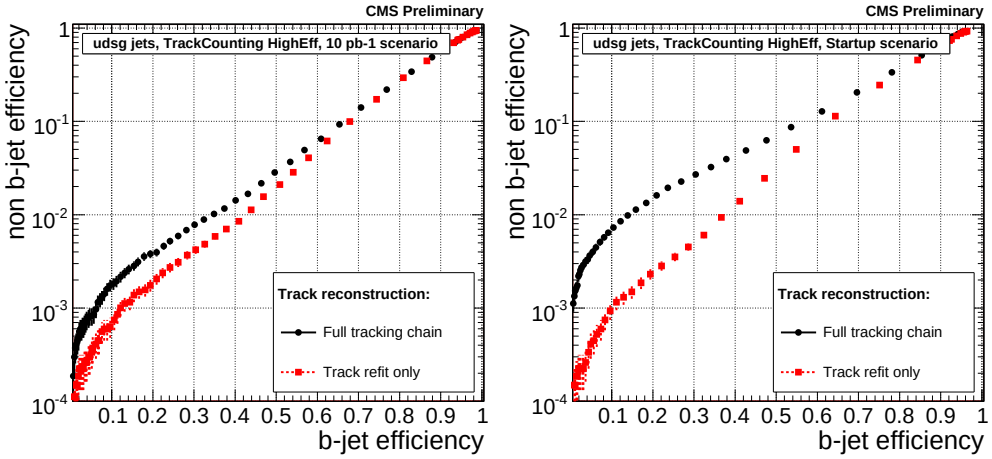


Figure C.21: Performance of the TrackCounting (high efficiency) algorithm for the refit-only and the full track reconstruction approach in case of the 10 pb^{-1} (left) and startup (right) misalignment scenario for light flavour and gluon jets

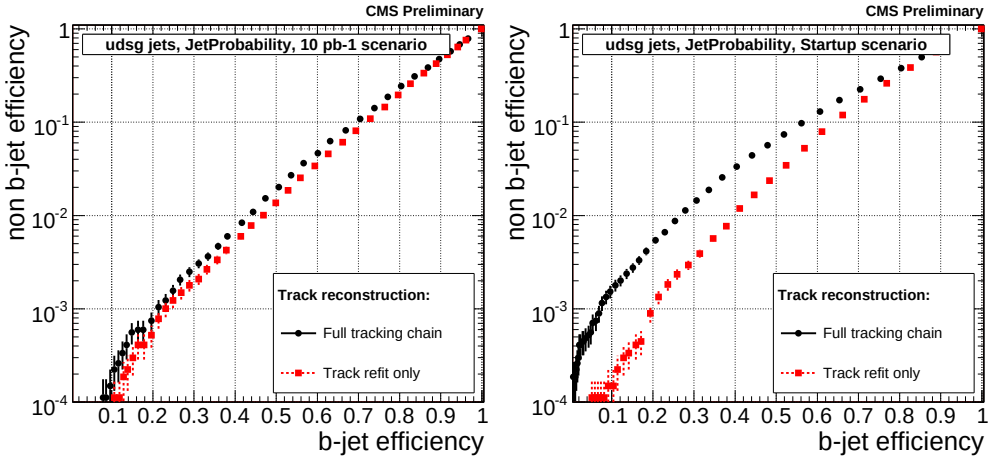


Figure C.22: Performance of the JetProbability algorithm for the refit-only and the full track reconstruction approach in case of the 10 pb^{-1} (left) and startup (right) misalignment scenario for light flavour and gluon jets

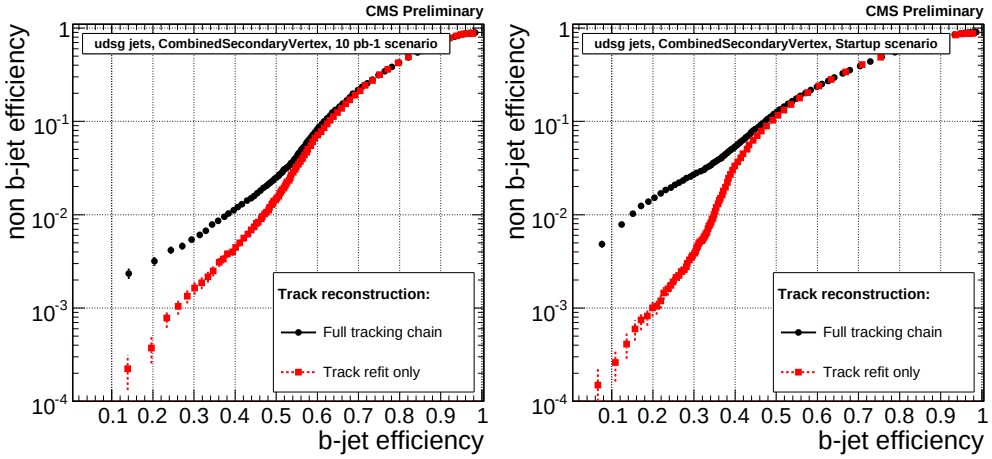


Figure C.23: Performance of the CombinedSecondaryVertex algorithm for the refit-only and the full track reconstruction approach in case of the 10 pb^{-1} (left) and startup (right) misalignment scenario for light flavour and gluon jets

C.5 Impact of Misalignment on the Input Variables

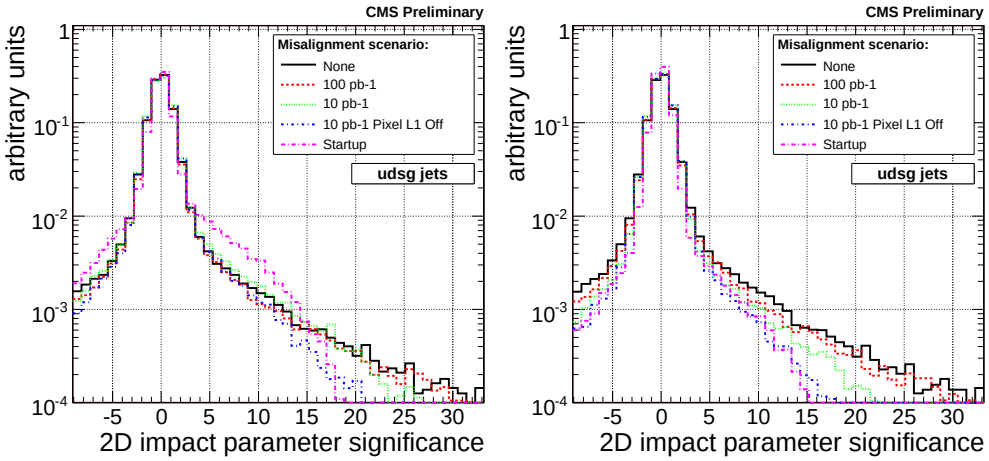


Figure C.24: Distribution of the transverse impact parameter significance of all tracks associated to a jet in case of a complete track reconstruction (left) and the refit-only approach (right). The distributions are presented for udsg-jets and the different misalignment scenarios

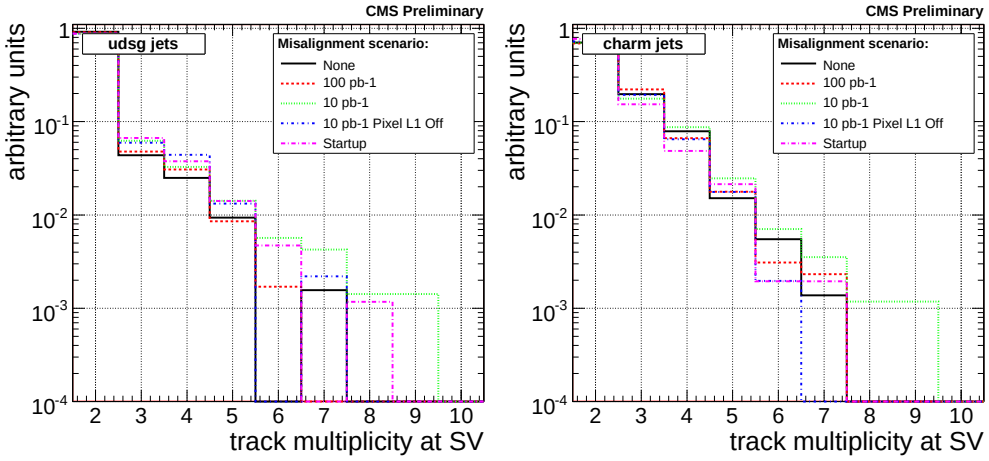


Figure C.25: Multiplicity of charged particle tracks associated to the secondary vertex in case of udsg-jets (left) and c-jets (right)

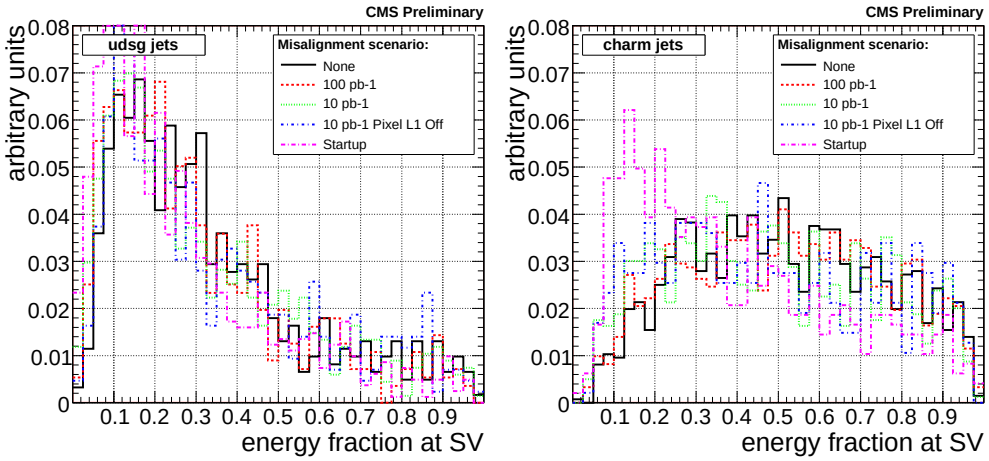


Figure C.26: Energy fraction of tracks associated to the secondary vertex with respect to the energy of all tracks associated to the jet in case of udsg-jets (left) and c-jets (right)

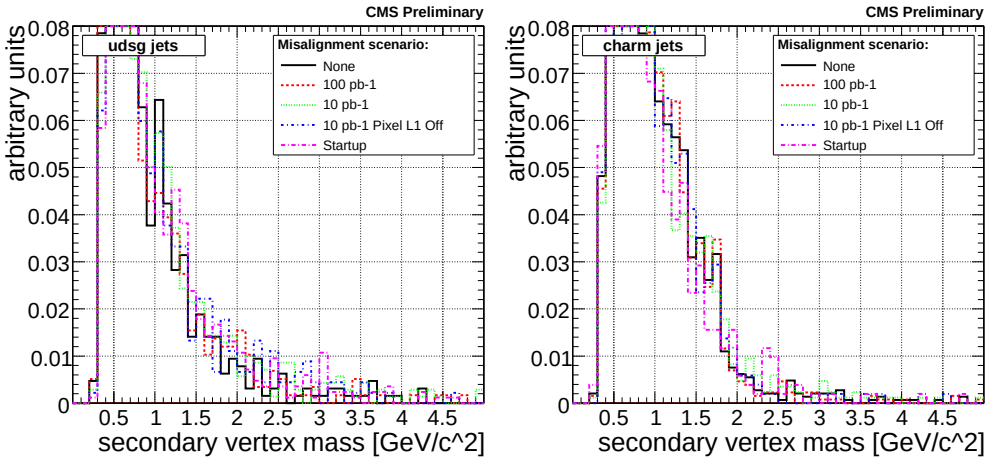


Figure C.27: Invariant mass of tracks associated to the secondary vertex in case of udsg-jets (left) and c-jets (right)

C.6 Impact of Misalignment on the Algorithm Performance

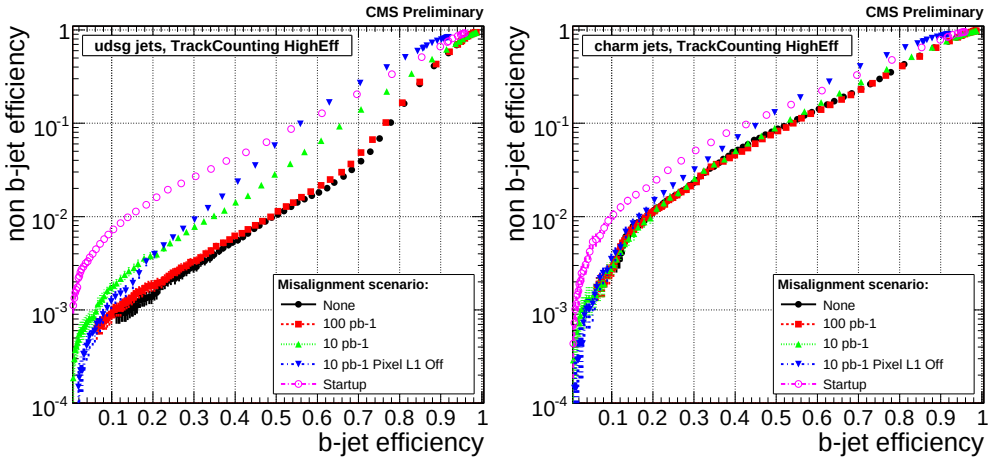


Figure C.28: b-jet efficiency versus non b-jet efficiency for the various misalignment scenarios in case of the TrackCounting (high efficiency) algorithm. The distributions are presented for light flavour and gluon (left) and charm (right) jets

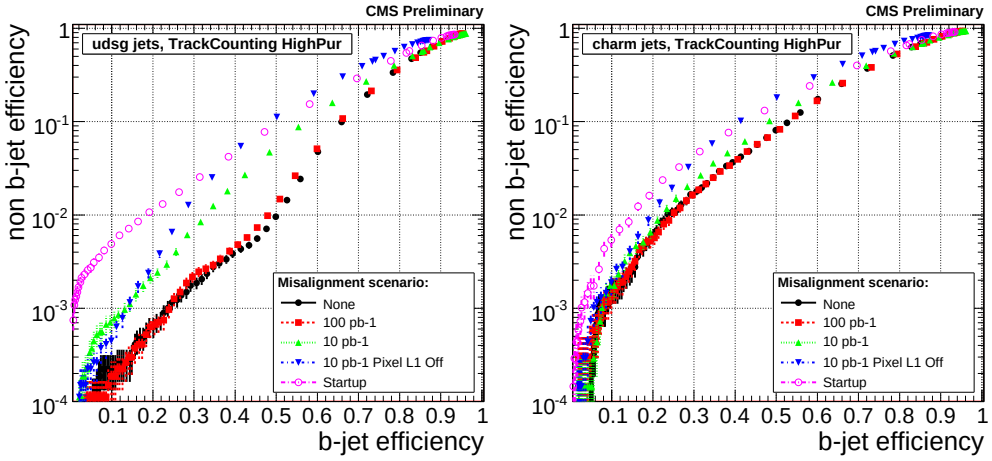


Figure C.29: b-jet efficiency versus non b-jet efficiency for the various misalignment scenarios in case of the TrackCounting (high purity) algorithm. The distributions are presented for light flavour and gluon (left) and charm (right) jets

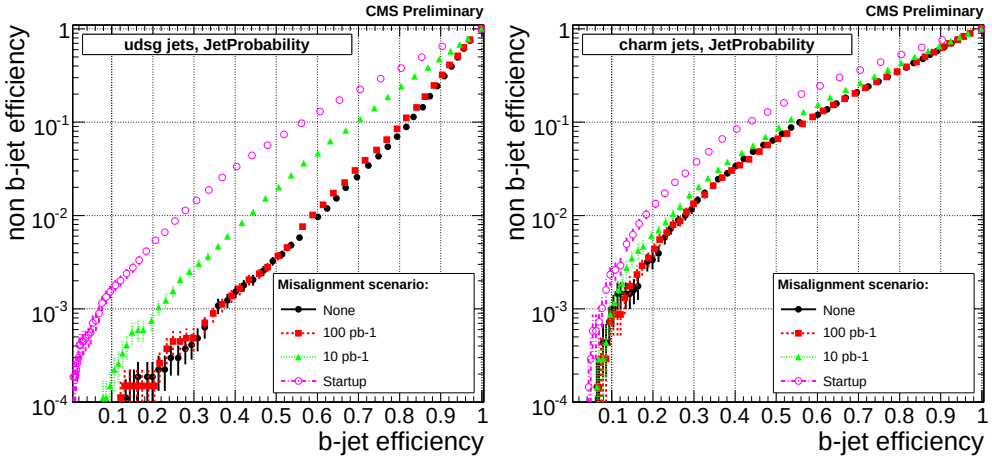


Figure C.30: b-jet efficiency versus non b-jet efficiency for the various misalignment scenarios in case of the JetProbability algorithm. The distributions are presented for light flavour and gluon (left) and charm (right) jets

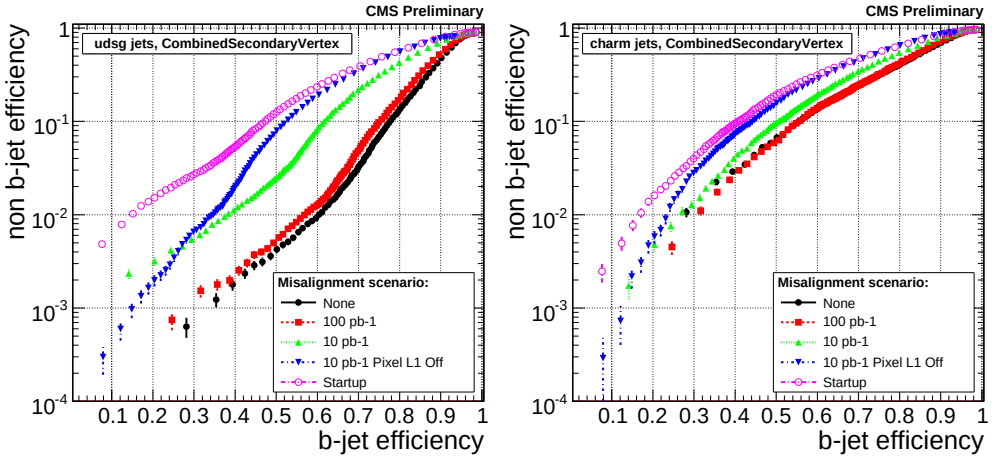


Figure C.31: b-jet efficiency versus non b-jet efficiency for the various misalignment scenarios in case of the CombinedSecondaryVertex algorithm. The distributions are presented for light flavour and gluon (left) and charm (right) jets

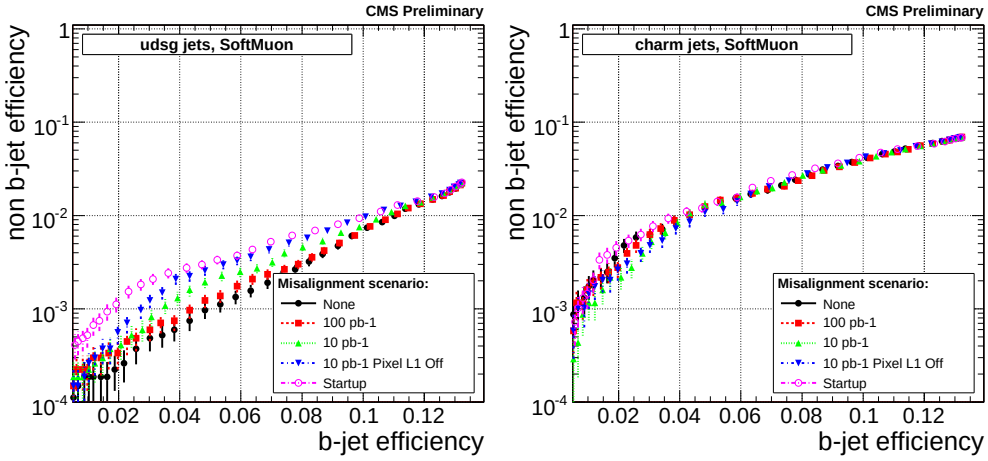


Figure C.32: b-jet efficiency versus non b-jet efficiency for the various misalignment scenarios in case of the SoftMuon algorithm. The distributions are presented for light flavour and gluon (left) and charm (right) jets

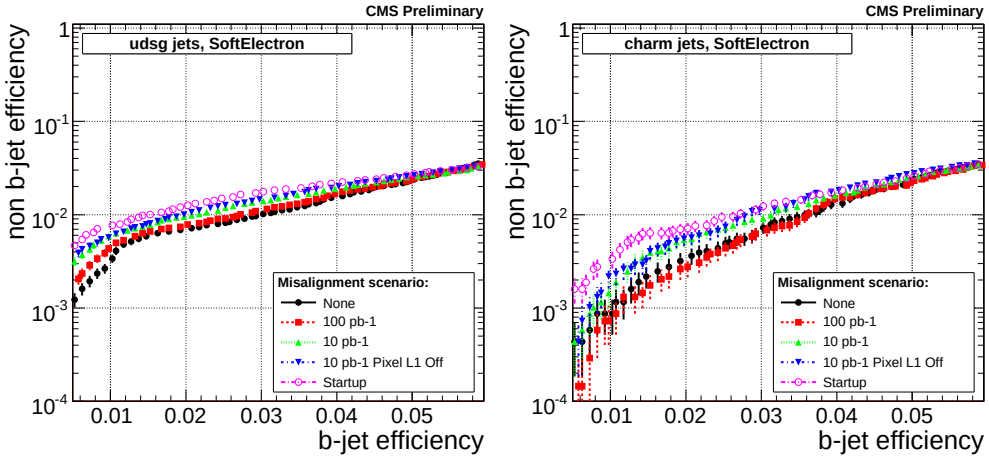


Figure C.33: b-jet efficiency versus non b-jet efficiency for the various misalignment scenarios in case of the SoftElectron algorithm. The distributions are presented for light flavour and gluon (left) and charm (right) jets

C.7 SimpleSecondaryVertex Candidates

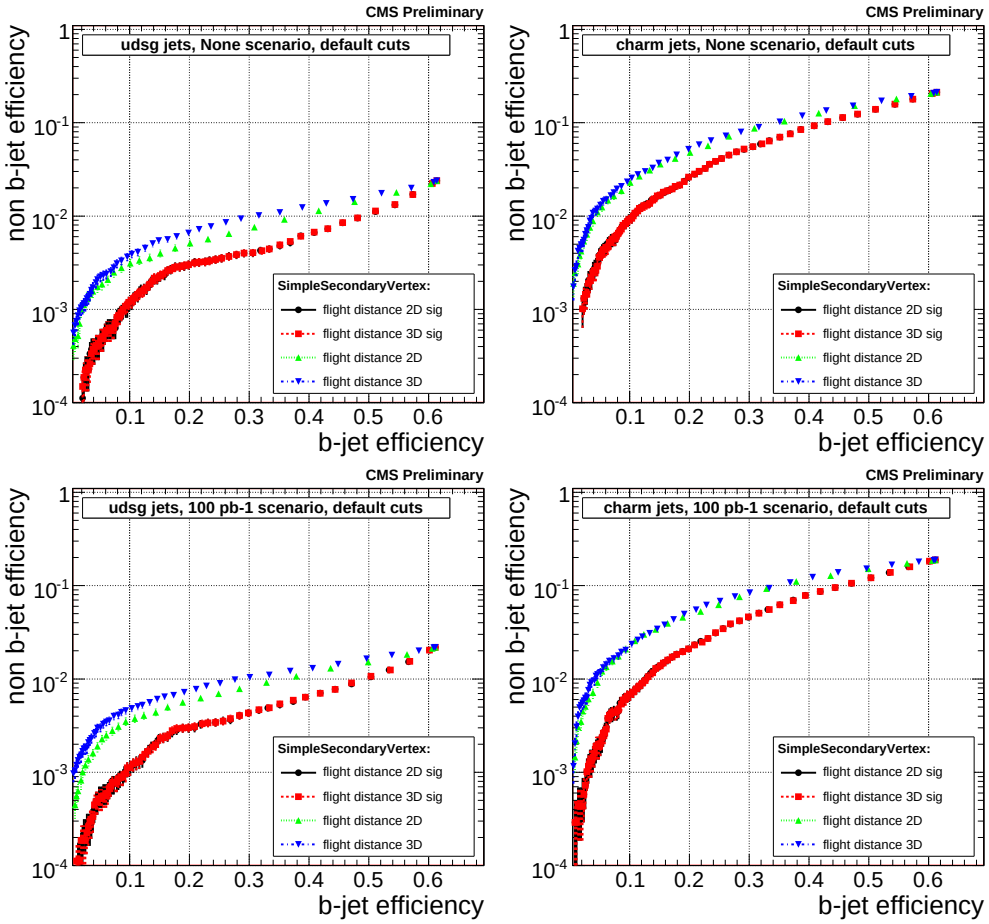


Figure C.34: b-jet efficiency versus non b-jet efficiency for the four SimpleSecondaryVertex candidates in case of an ideally aligned track detector (top) and the 100 pb^{-1} misalignment scenario (bottom). The distributions are presented for light flavour and gluon (left) and charm (right) jets

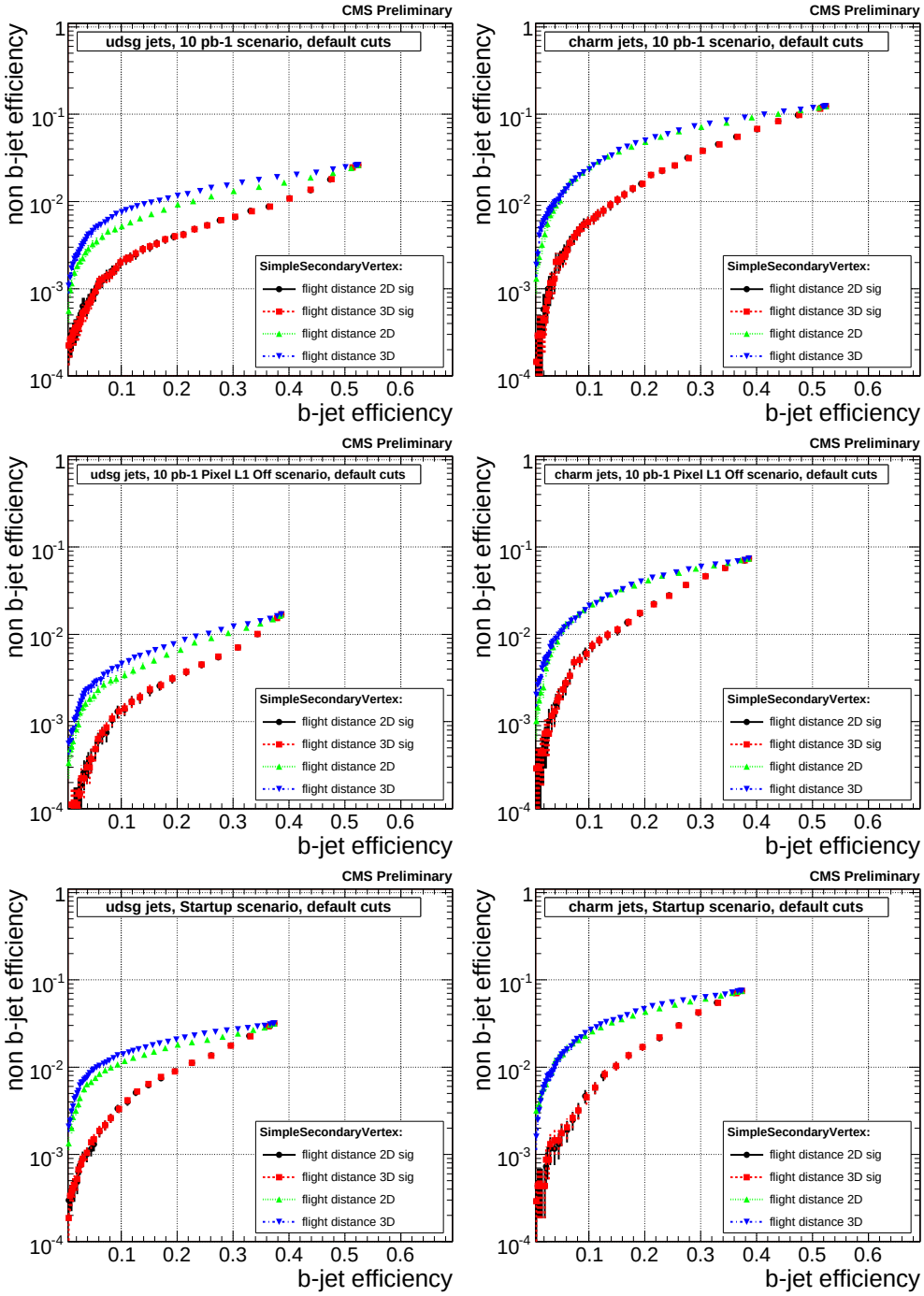


Figure C.35: b-jet efficiency versus non b-jet efficiency for the four SimpleSecondaryVertex candidates in case of the 10 pb^{-1} scenario (top), the missing first pixel layer scenario (middle) as well as startup tracker conditions (bottom). The distributions are presented for light flavour and gluon (left) and charm (right) jets

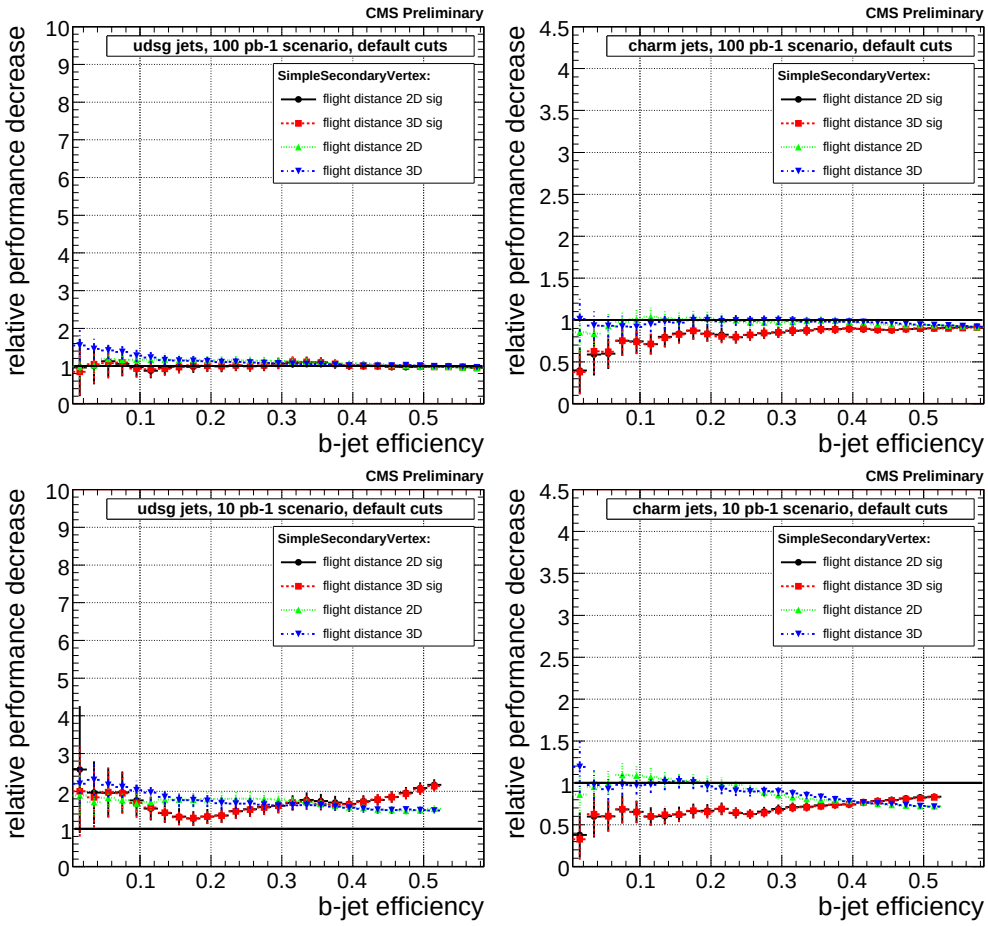


Figure C.36: Relative performance decrease for the four SimpleSecondaryVertex candidates in case of the 100 pb^{-1} (top) and 10 pb^{-1} (bottom) misalignment scenario. The distributions are presented for light flavour and gluon (left) and charm (right) jets

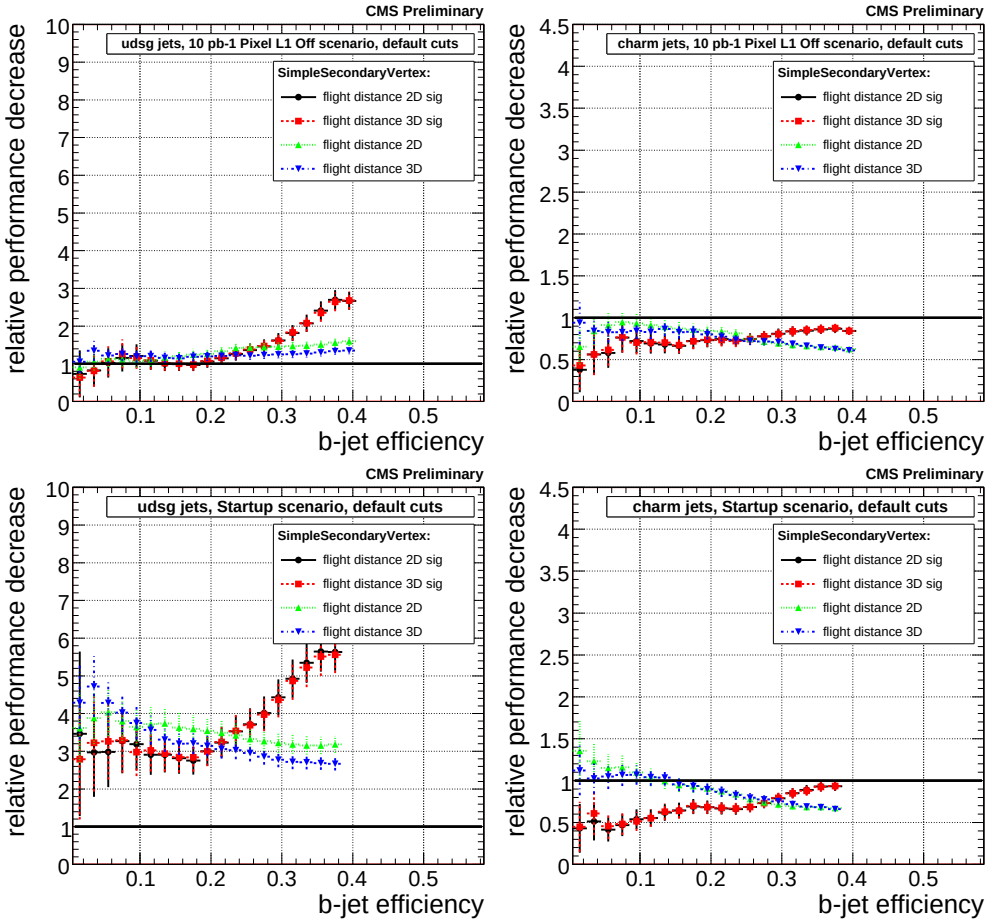


Figure C.37: Relative performance decrease for the four SimpleSecondaryVertex candidates in case of the missing first pixel layer scenario (top) and startup detector conditions (bottom). The distributions are presented for light flavour and gluon (left) and charm (right) jets

C.8 Flavour Efficiencies Versus Discriminator Cut

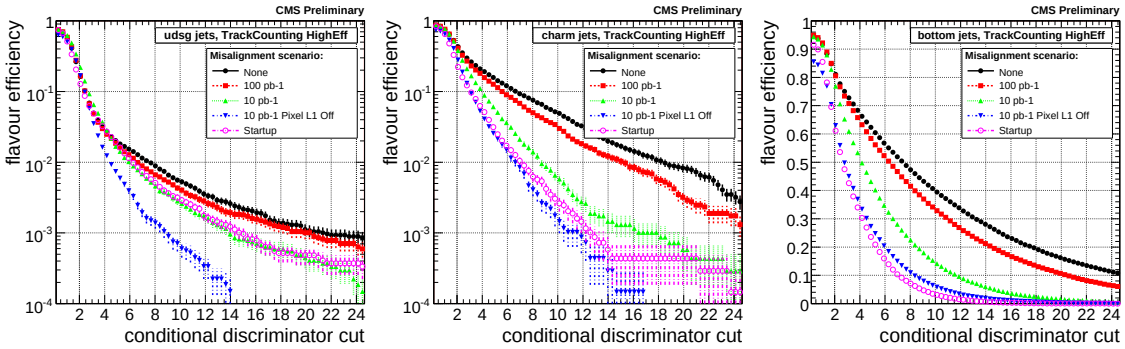


Figure C.38: Flavour tagging efficiencies with respect to the discriminator cut for the different misalignment scenarios. The distributions are presented for usdg (left), charm (middle) and bottom (right) jets in case of the TrackCounting (high efficiency) algorithm

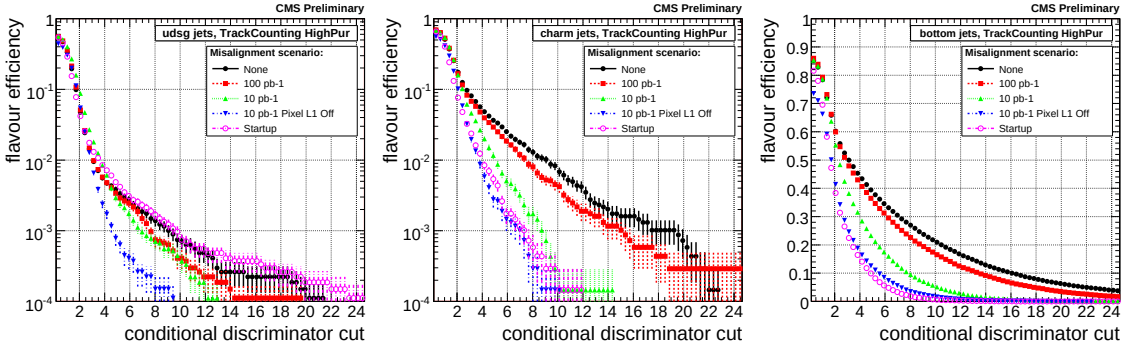


Figure C.39: Flavour tagging efficiencies with respect to the discriminator cut for the different misalignment scenarios. The distributions are presented for usdg (left), charm (middle) and bottom (right) jets in case of the TrackCounting (high purity) algorithm

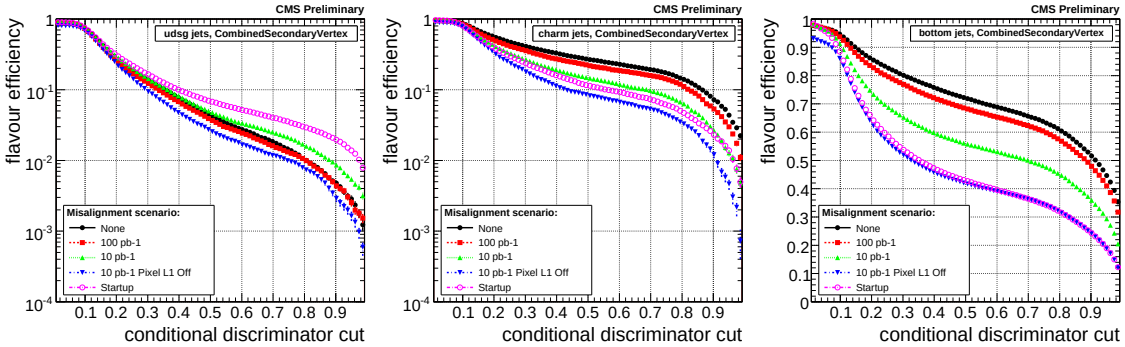


Figure C.40: Flavour tagging efficiencies with respect to the discriminator cut for the different misalignment scenarios. The distributions are presented for usdg (left), charm (middle) and bottom (right) jets in case of the CombinedSecondaryVertex algorithm

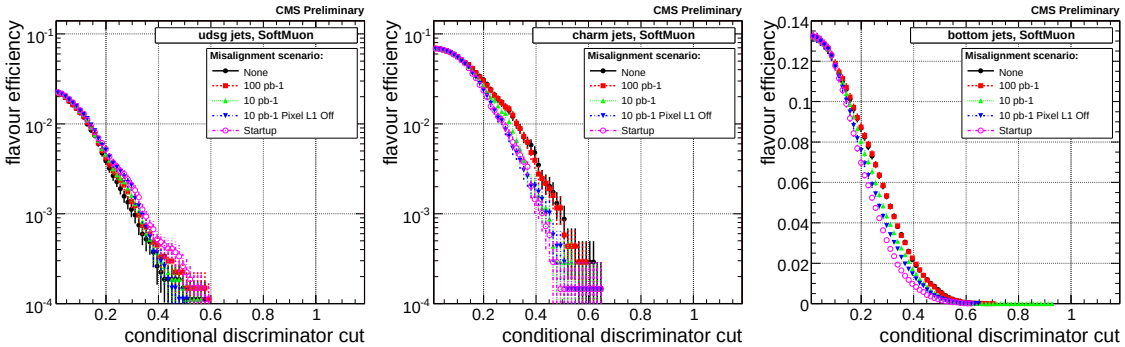


Figure C.41: Flavour tagging efficiencies with respect to the discriminator cut for the different misalignment scenarios. The distributions are presented for usdg (left), charm (middle) and bottom (right) jets in case of the SoftMuon algorithm

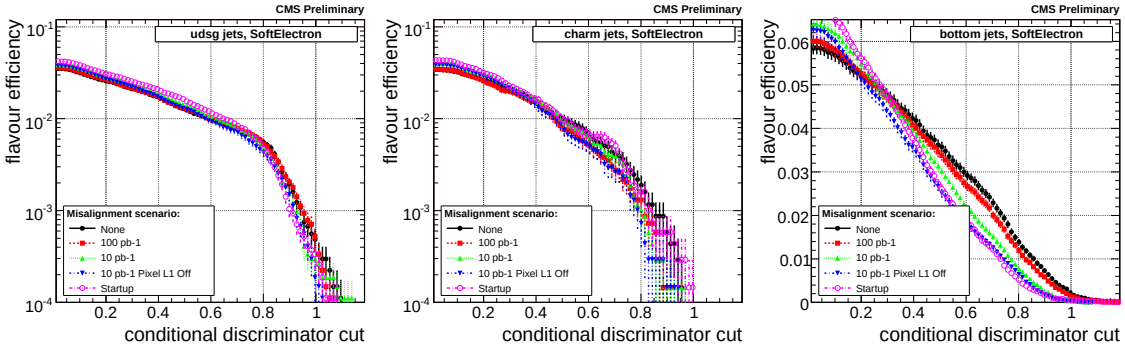


Figure C.42: Flavour tagging efficiencies with respect to the discriminator cut for the different misalignment scenarios. The distributions are presented for usdg (left), charm (middle) and bottom (right) jets in case of the SoftElectron algorithm

C.9 SV Optimisation Studies

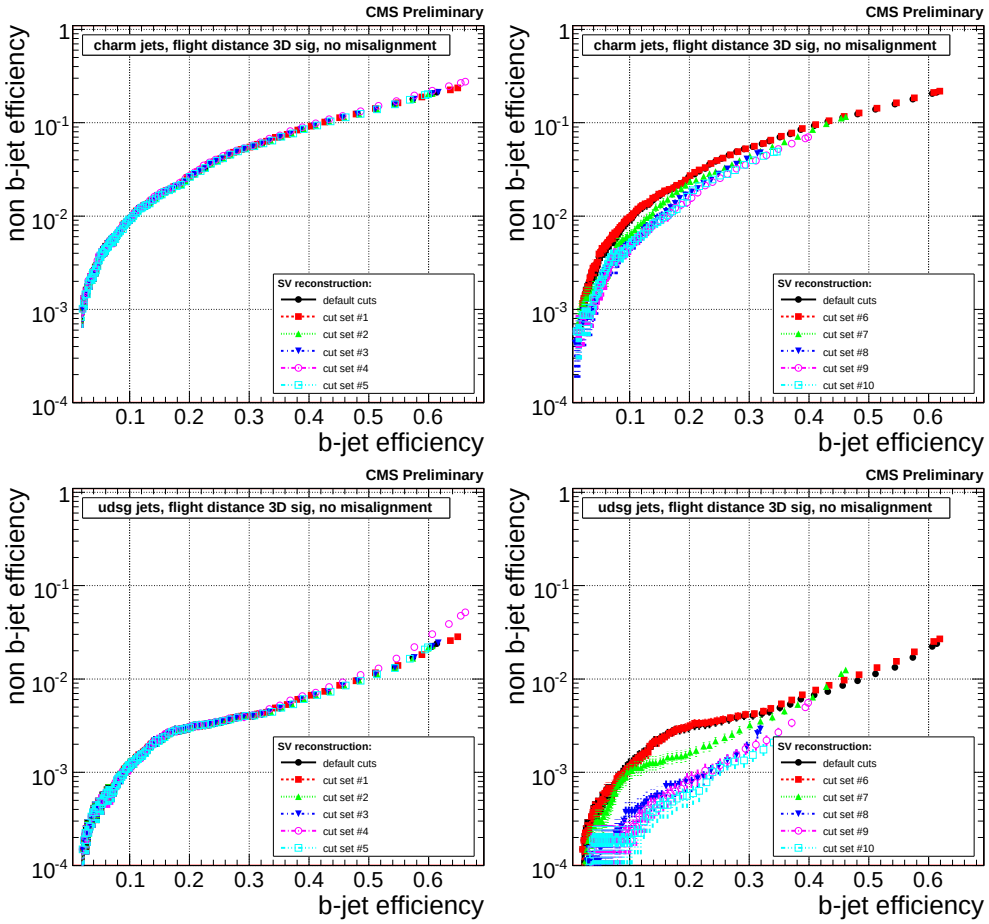


Figure C.43: Performance distributions for the SimpleSecondaryVertex algorithm with respect to the different cut sets presented for an ideally aligned tracker in case of charm (top) and udsg (bottom) jets

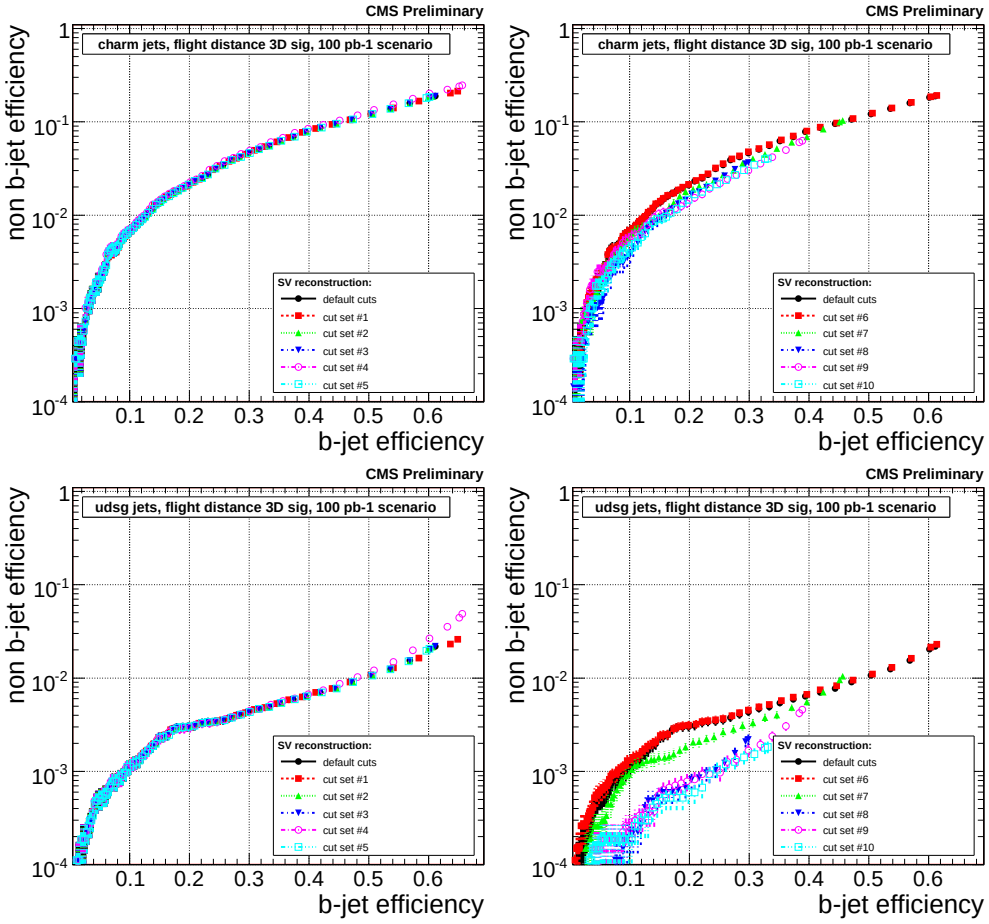


Figure C.44: Performance distributions for the SimpleSecondaryVertex algorithm with respect to the different cut sets presented for the 100 pb^{-1} misalignment scenario in case of charm (top) and udsg (bottom) jets

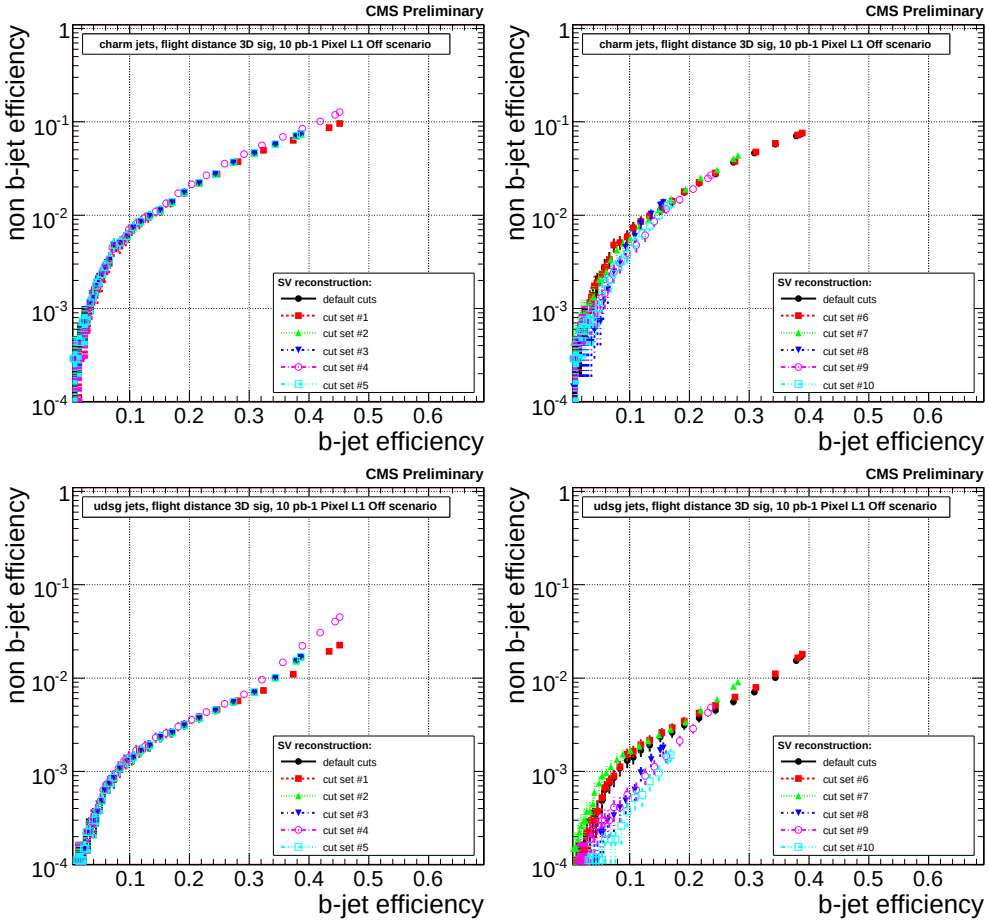


Figure C.45: Performance distributions for the SimpleSecondaryVertex algorithm with respect to the different cut sets presented for the first missing pixel layer misalignment scenario in case of charm (top) and udsg (bottom) jets

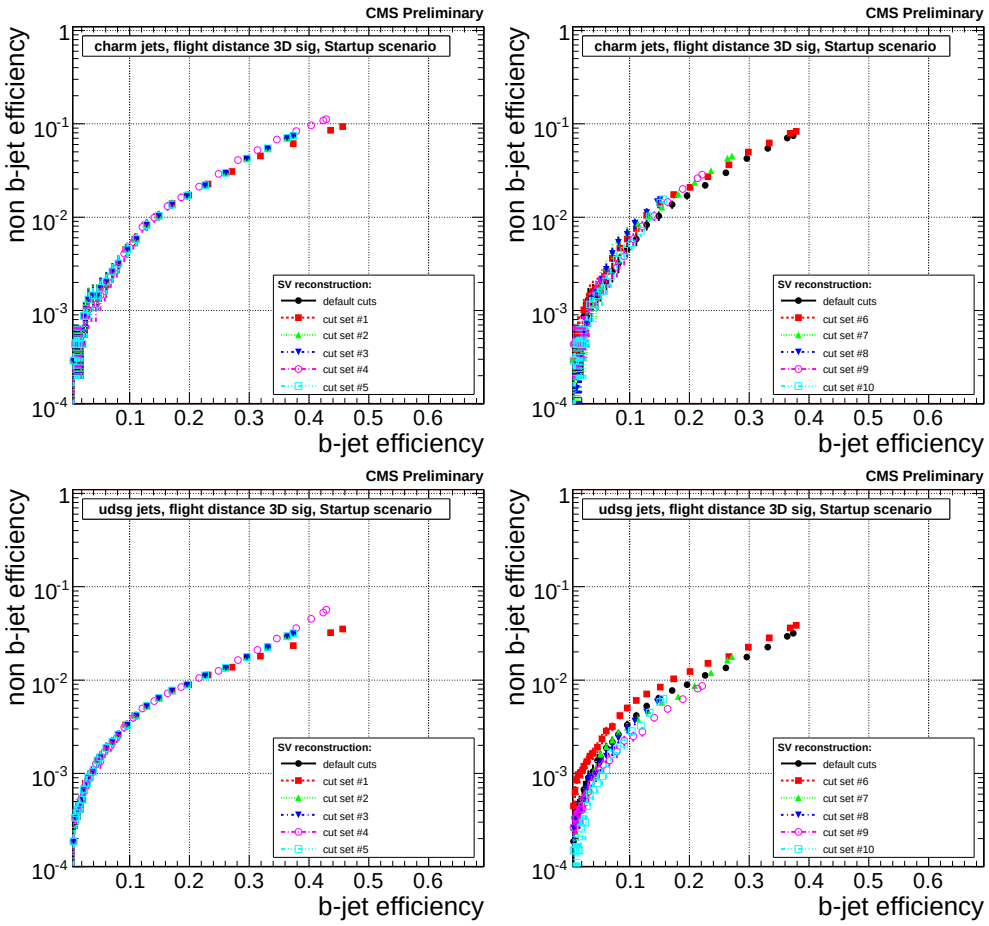


Figure C.46: Performance distributions for the SimpleSecondaryVertex algorithm with respect to the different cut sets presented for startup tracker conditions in case of charm (top) and udsg (bottom) jets

C.10 Variations of the Alignment Position Error

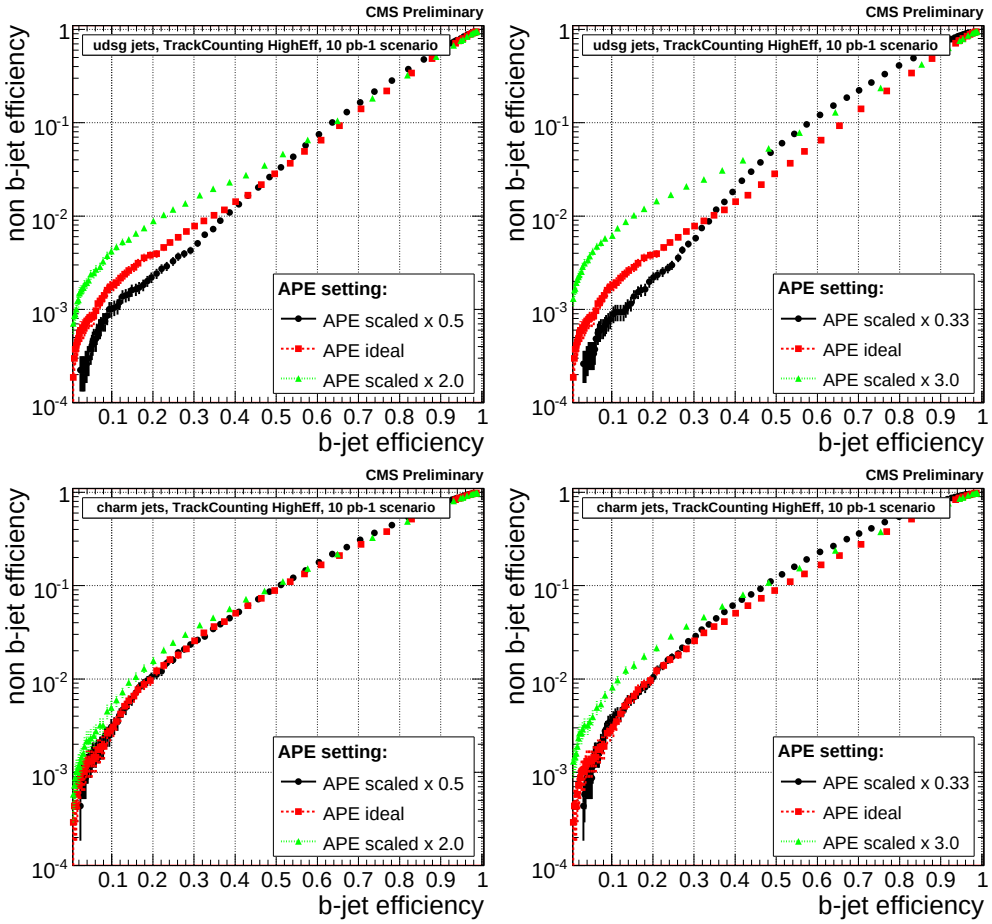


Figure C.47: b-tagging performance for the high efficiency TrackCounting algorithm with respect to scaled APEs by a factor of 1/3, 1/2, 2 and 3. The distributions are presented for light flavour (top) and charm (bottom) jets in case of the 10 pb^{-1} misalignment scenario

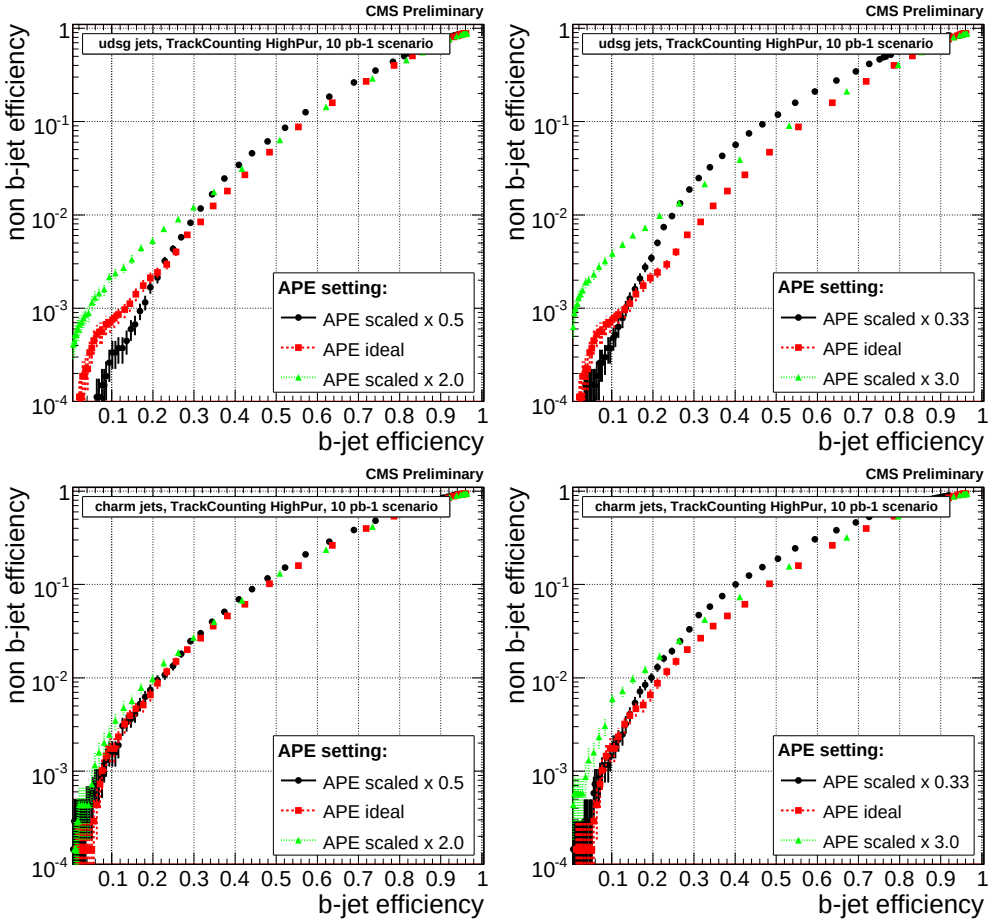


Figure C.48: b-tagging performance for the high purity TrackCounting algorithm with respect to scaled APEs by a factor of 1/3, 1/2, 2 and 3. The distributions are presented for light flavour (top) and charm (bottom) jets in case of the 10 pb^{-1} misalignment scenario

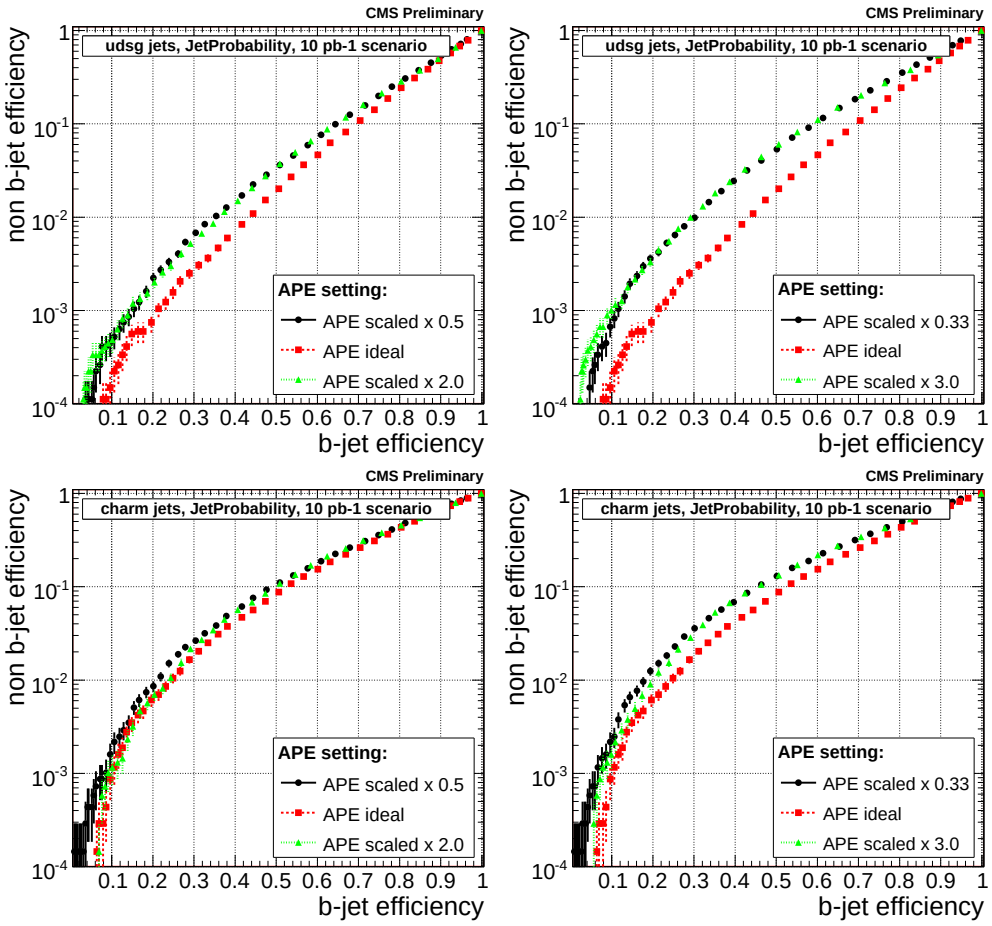


Figure C.49: b-tagging performance for the JetProbability algorithm with respect to scaled APEs by a factor of 1/3, 1/2, 2 and 3. The distributions are presented for light flavour (top) and charm (bottom) jets in case of the 10 pb^{-1} misalignment scenario

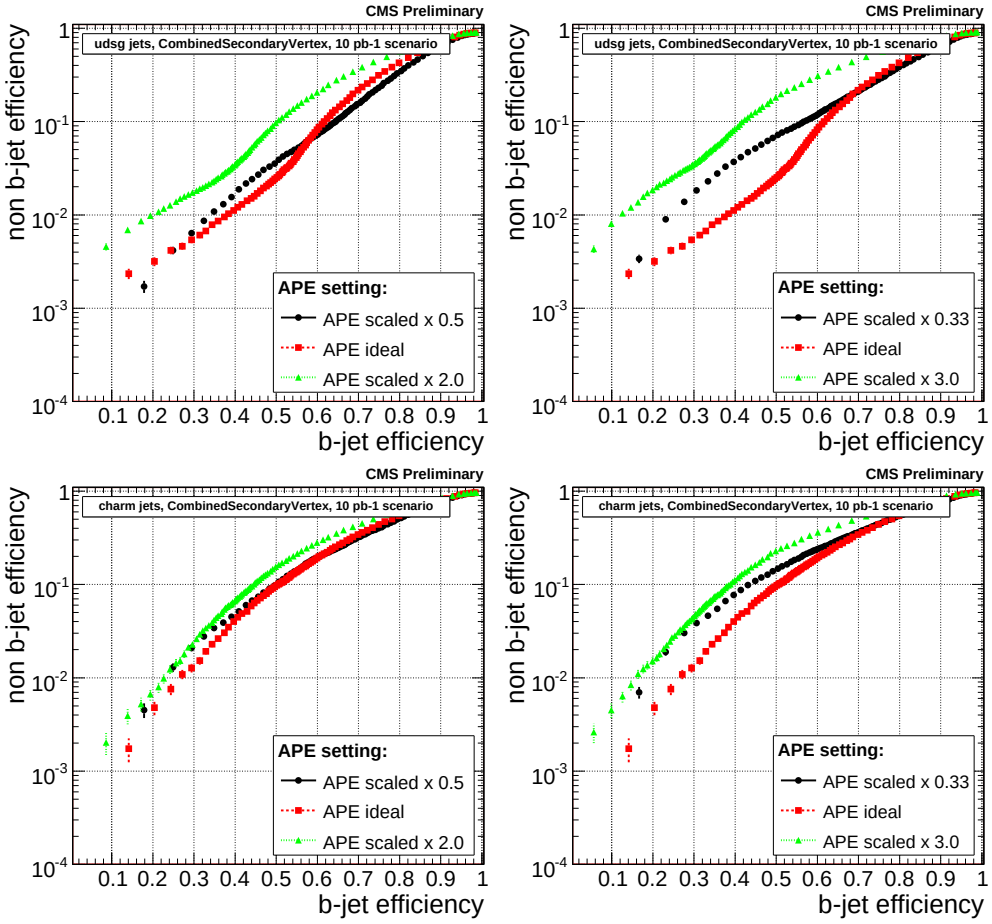


Figure C.50: b-tagging performance for the CombinedSecondaryVertex algorithm with respect to scaled APEs by a factor of 1/3, 1/2, 2 and 3. The distributions are presented for light flavour (top) and charm (bottom) jets in case of the 10 pb^{-1} misalignment scenario

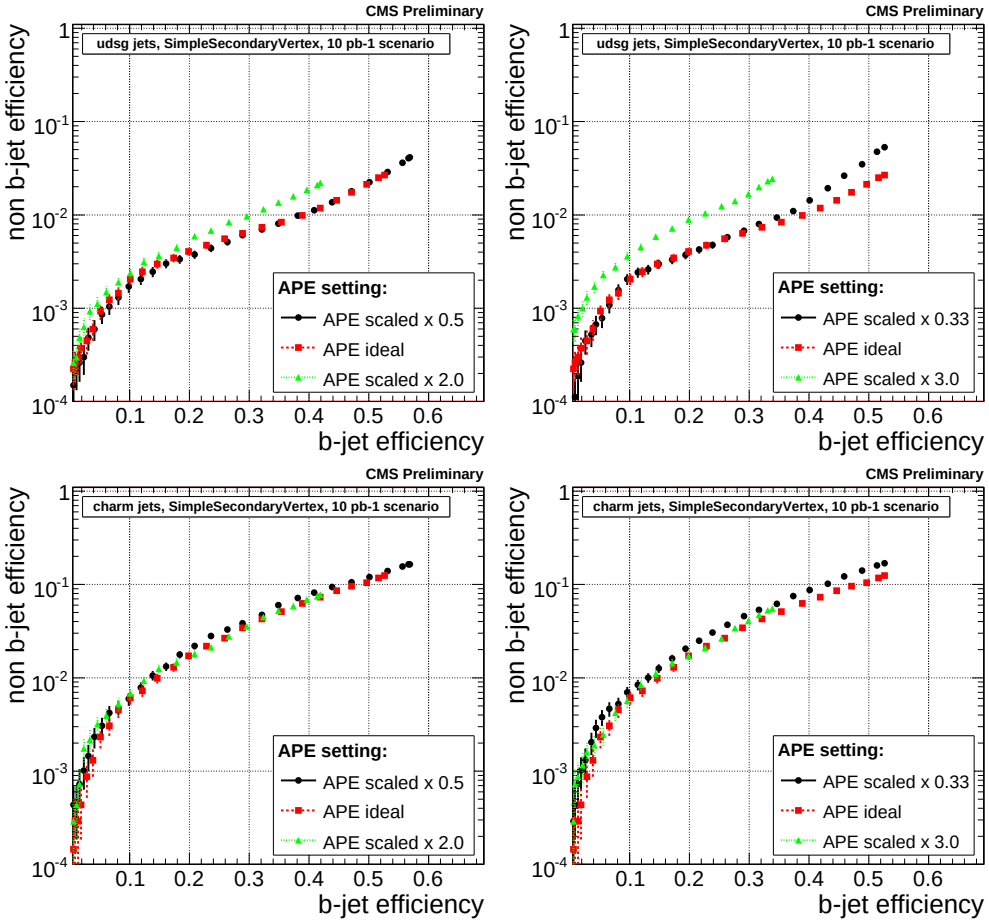


Figure C.51: b-tagging performance for the SimpleSecondaryVertex algorithm with respect to scaled APEs by a factor of 1/3, 1/2, 2 and 3. The distributions are presented for light flavour (top) and charm (bottom) jets in case of the 10 pb^{-1} misalignment scenario

C.11 Variations of the Track Association Cone Size

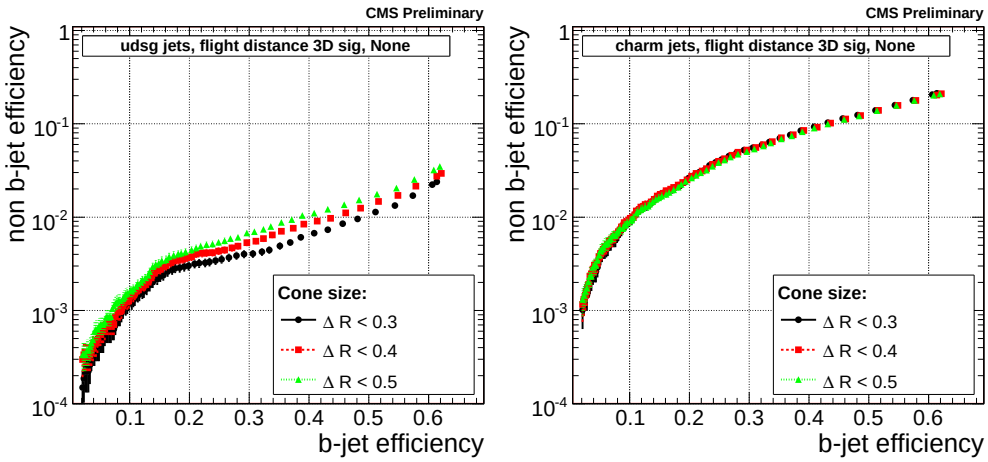


Figure C.52: Performance of the SimpleSecondaryVertex algorithm with respect to the track association cone size. The distributions are presented for light flavour (left) and charm (right) jets in case of an ideally aligned tracker

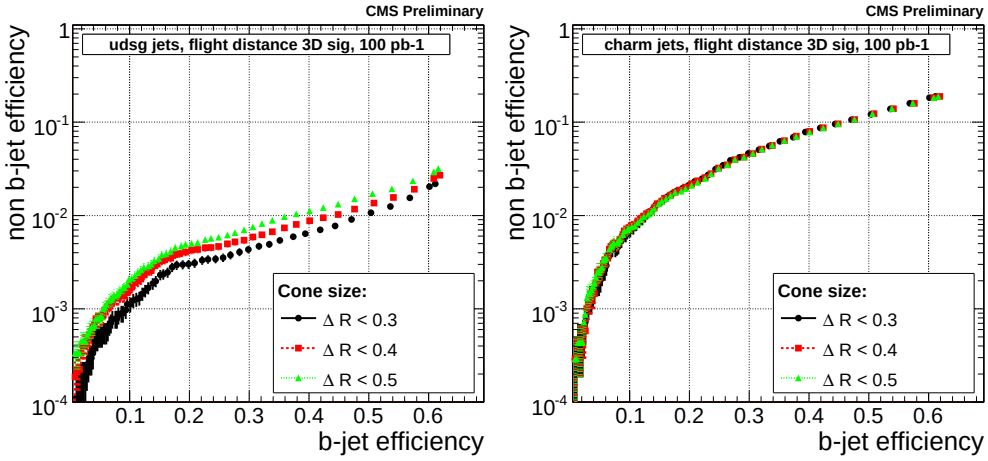


Figure C.53: Performance of the SimpleSecondaryVertex algorithm with respect to the track association cone size. The distributions are presented for light flavour (left) and charm (right) jets in case of the 100 pb^{-1} misalignment scenario

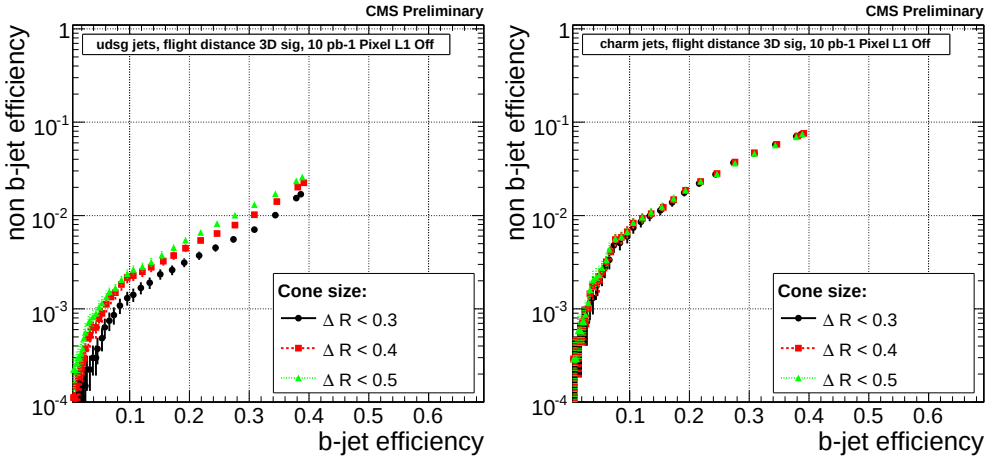


Figure C.54: Performance of the SimpleSecondaryVertex algorithm with respect to the track association cone size. The distributions are presented for light flavour (left) and charm (right) jets in case of the missing first pixel layer scenario

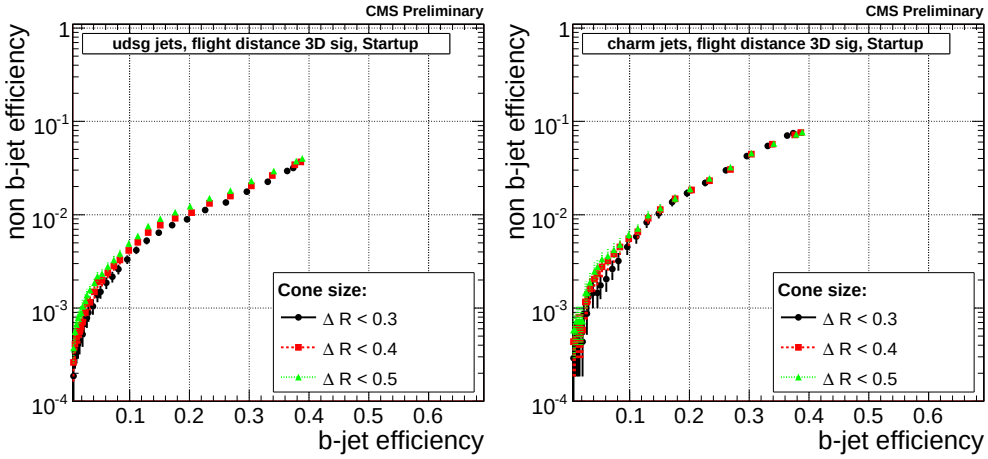


Figure C.55: Performance of the SimpleSecondaryVertex algorithm with respect to the track association cone size. The distributions are presented for light flavour (left) and charm (right) jets in case of startup tracker conditions

C.12 Performance of All Lifetime Based b-Tagging Algorithms

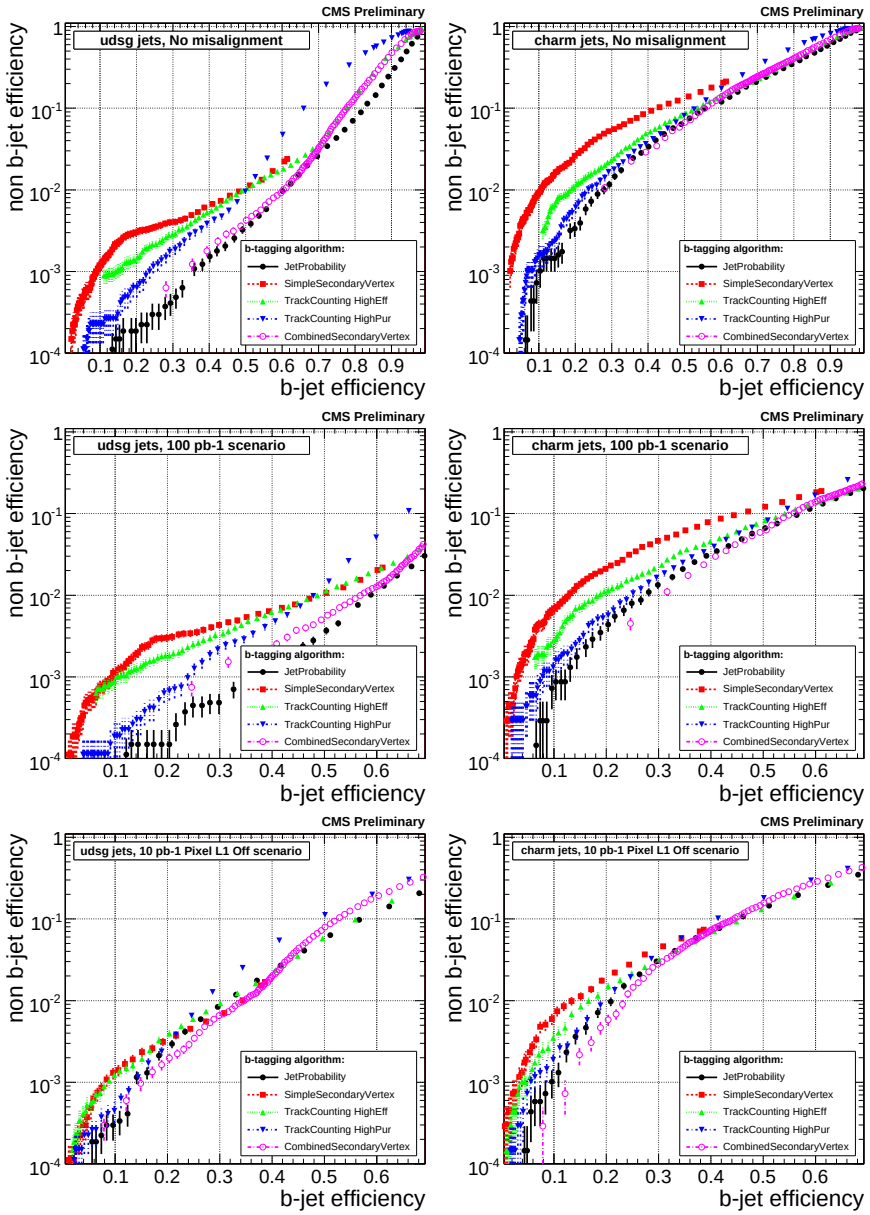


Figure C.56: Performances of the lifetime based b-tagging algorithms presented for light flavour (left) and charm (right) jets in case of an ideally aligned tracker (top), the 100 pb^{-1} (middle) and the missing first pixel layer (bottom) misalignment scenario

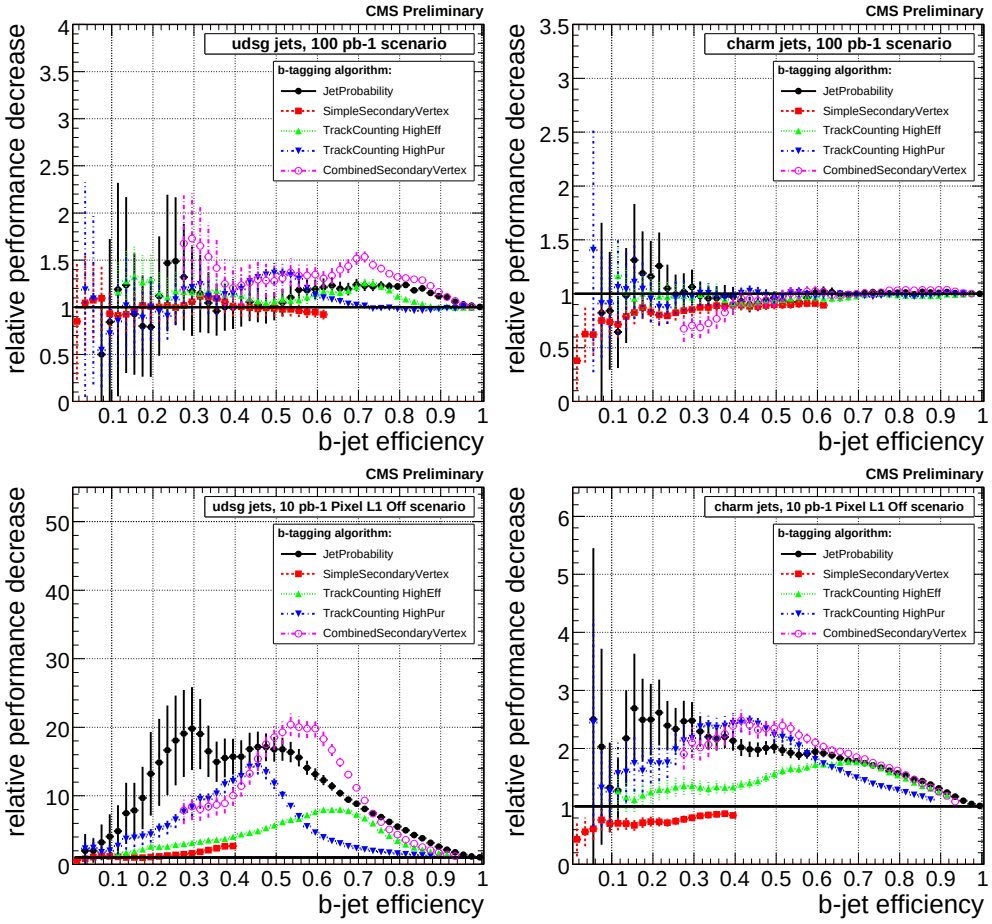


Figure C.57: Relative performance decrease of the lifetime based b-tagging algorithms with respect to an ideally aligned tracker, presented for light flavour (left) and charm (right) jets in case of the 100 pb^{-1} (top) and the missing first pixel layer (bottom) misalignment scenario

Appendix D

CMS Tracking Efficiencies

D.1 Track Reconstruction Efficiencies and Resolutions

The default tracking algorithm used in CMSSW is the "CombinatorialTrackFinder". Its most important tunable parameters for the "Pattern Recognition" step and their default values (in brackets) are:

- maximum number of track candidates propagated at each iteration step (5)
- maximum χ^2 of the hits which are considered compatible with the predicted track state (30)
- maximum number of invalid hits (1)
- minimum transverse momentum in GeV/c (0.9)
- minimum number of hits per track (5)

Figures D.1 and D.2 show the expected efficiency of the algorithm for single muons and pions with transverse momenta of 1, 10 and 100 GeV/c as well as for tracks in b-jets with transverse momenta between 120 and 170 GeV/c including low luminosity pile-up events. The reconstruction efficiency is defined as

$$\epsilon_{track} = \frac{number\ of\ reconstructed\ tracks}{number\ of\ simulated\ tracks} \tag{D.1}$$

There are two definitions of efficiencies presented here: the algorithmic efficiency is directly related to the performance of the track reconstruction algorithm itself whereas the global efficiency definition takes the acceptance of the tracking device, hit efficiencies and all other

factors influencing reconstruction into account as well. Figure D.3 shows the resolution of the track transverse momentum, the track in ϕ direction and the track impact parameters in transverse and longitudinal direction for single muons with transverse momenta of 1, 10 and 100 GeV/c.

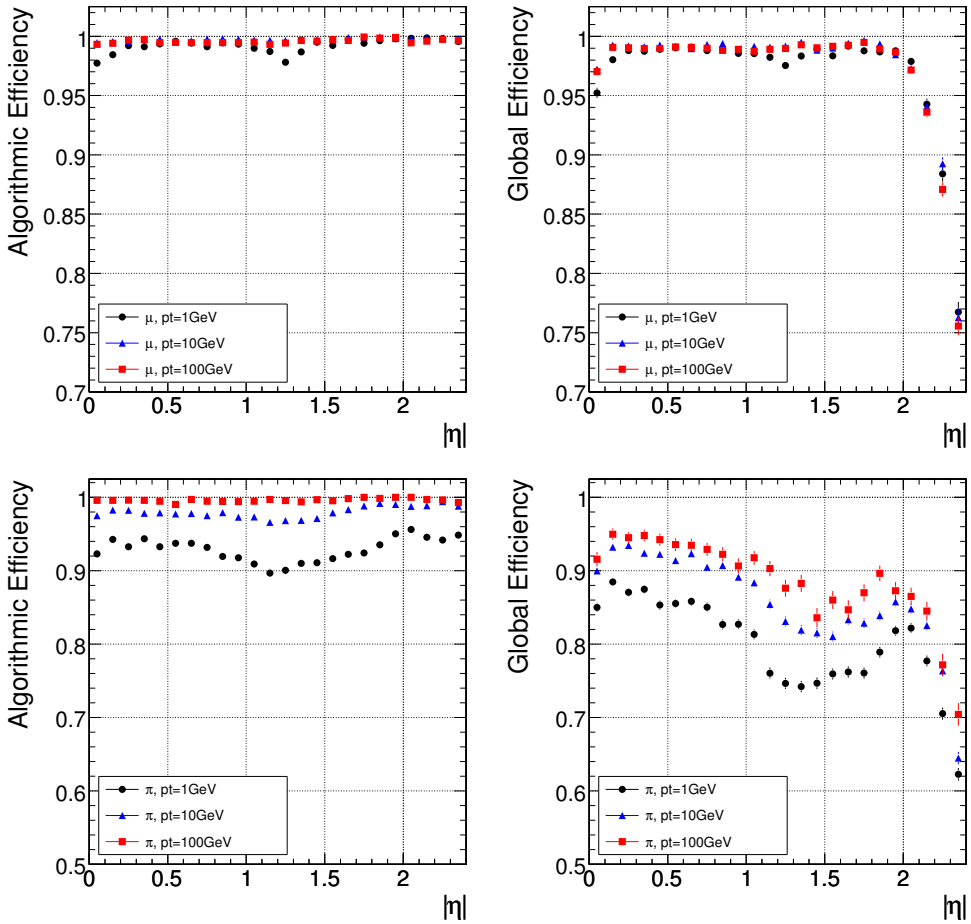


Figure D.1: Algorithmic (left) and global (right) tracking efficiency for single muons (top) and single pions (bottom) with transverse momenta of 1, 10 and 100 GeV/c [12]

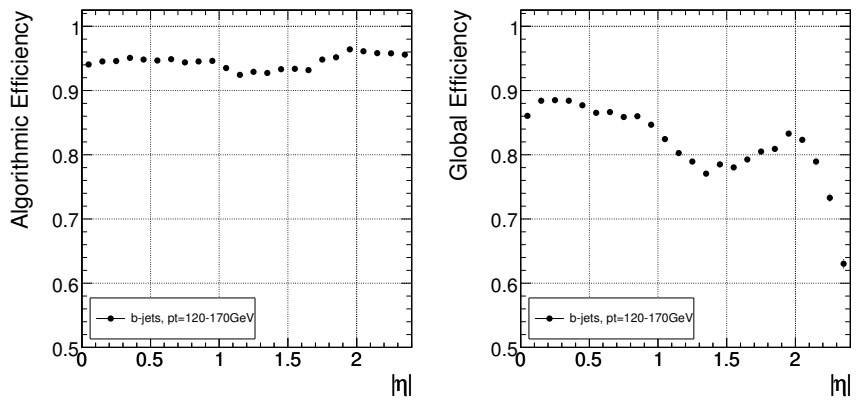


Figure D.2: Algorithmic (left) and global (right) tracking efficiency for tracks in b-jets with transverse momenta between 120 and 170 GeV/c [12]

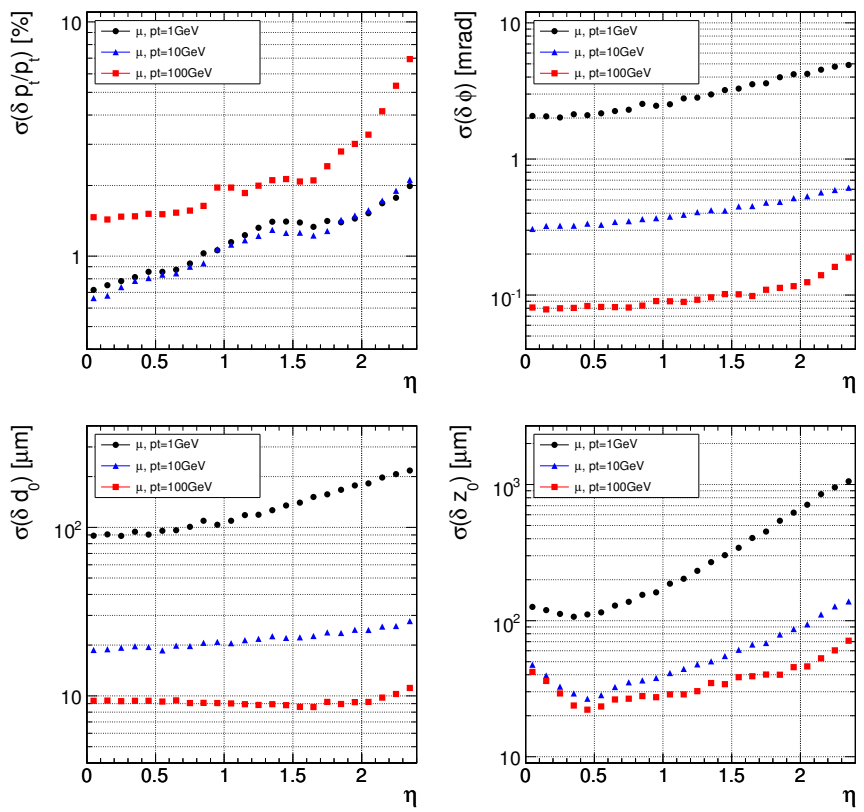


Figure D.3: Resolution of tracking parameters for single muons with transverse momenta of 1, 10 and 100 GeV/c: for the transverse momentum (top left), the track in ϕ direction (top right), the transverse (bottom left) and longitudinal (bottom right) impact parameter [12]

Appendix E

Old Versus New Software Framework

E.1 Results From the Physics TDR

The studies presented in this thesis have been performed by using the new CMS offline software framework CMSSW. It had been developed during the past years as a successor to the former software components ORCA and OSCAR, which were used for reconstruction and simulation, respectively. All software components have been rewritten and the performance of most of the new algorithms was verified with respect to their old behaviour documented in the experiment's physics technical design reports (TDR) [12] and [5]. The algorithms used for identifying b-quark jets have not been reinvestigated yet and as a side-result are therefore presented for the first time in coherence with the studies in this thesis and the accompanying publication [64]. The following section presents the b-tagging performances of the different algorithms as documented in the physics TDR.

Performance in QCD - Events

The distributions of the TrackCounting (high efficiency), JetProbability and CombinedSecondaryVertex algorithm have been determined by using a sample of QCD events with transverse jet momenta between 50 GeV/c and 80 GeV/c in a pseudorapidity region of $|\eta| < 1.4$ which corresponds to the barrel part of the detector. For the SoftLepton taggers, fully and semi-leptonic $t\bar{t}$-events have been used in addition to QCD events with transverse jet momenta between 50 GeV/c and 230 GeV/c. All events have been simulated with pile-up at a luminosity of $\mathcal{L}_{low} = 2 \times 10^{33}$ cm^{-2}s^{-1}.

Figure E.1 presents the b-tagging performances for the mentioned algorithms with respect to the described data samples. As stated in [66], the algorithm performance depends as well on the underlying event topology. Since the used data samples do not exactly match the ones used in this thesis, an absolute comparison is not implicitly meaningful, but the

right tendency should be clearly visible. Other aspects which could lead to discrepancies between the old and the new performance distributions are enhanced or even exchanged algorithms with respect to their implementation in the old framework. Slight differences in e.g. reconstruction efficiencies or track parameter resolutions might occur and thus, lead to differences which appear in the b-tagging performance distributions. Note that the "physics" definition has been used in order to determine the real jet flavour of a jet. Therefore, b-jets originating from gluon splitting are not treated as b-jets as it would be in case of the "algorithmic" definition. Therefore, the distributions for gluons and light flavour jets diverge towards the high purity region.

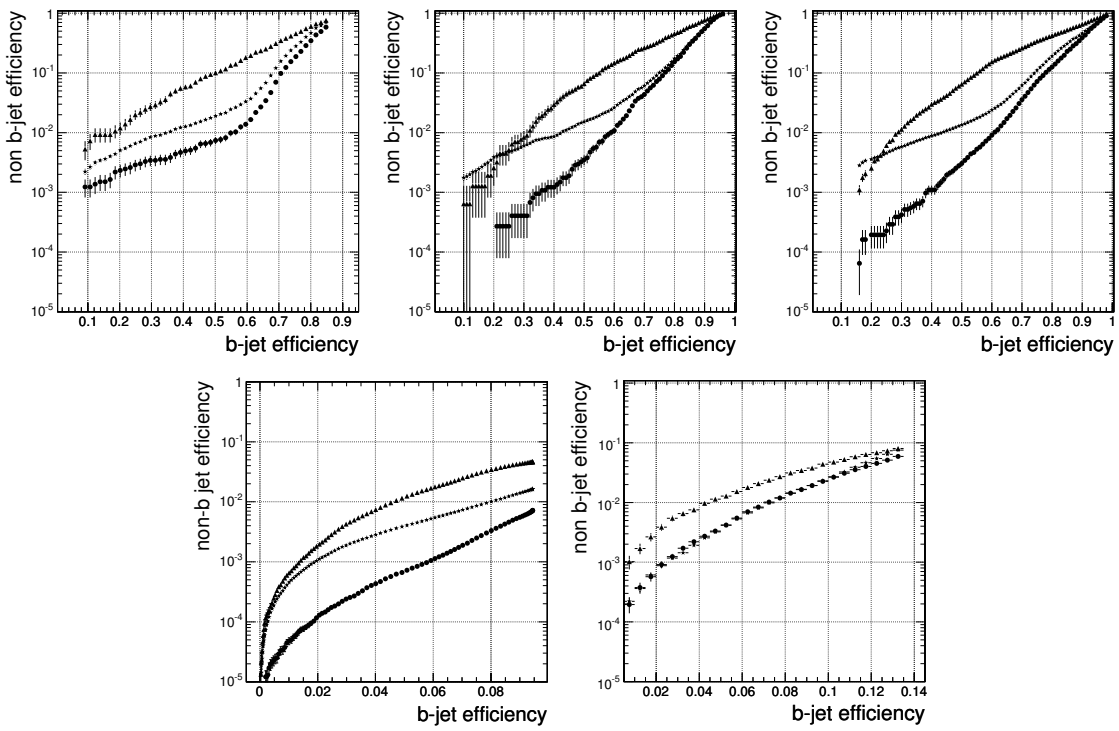


Figure E.1: b-tagging performance of the high efficiency TrackCounting (top left), JetProbability (top middle), CombinedSecondaryVertex (top right), SoftMuon (bottom left) and SoftElectron (bottom right) algorithm. The distributions are presented for c-jets (triangles), uds-jets (circles) and gluon jets (stars)

Performance in $t\bar{t}$ - Events

Figure E.2 shows the performance distributions of the JetProbability and CombinedSecondaryVertex algorithm for a sample of semi-leptonic $t\bar{t}$-events with transverse jet momenta larger than 30 GeV/c in the barrel region of the detector ($|\eta| < 1.4$).

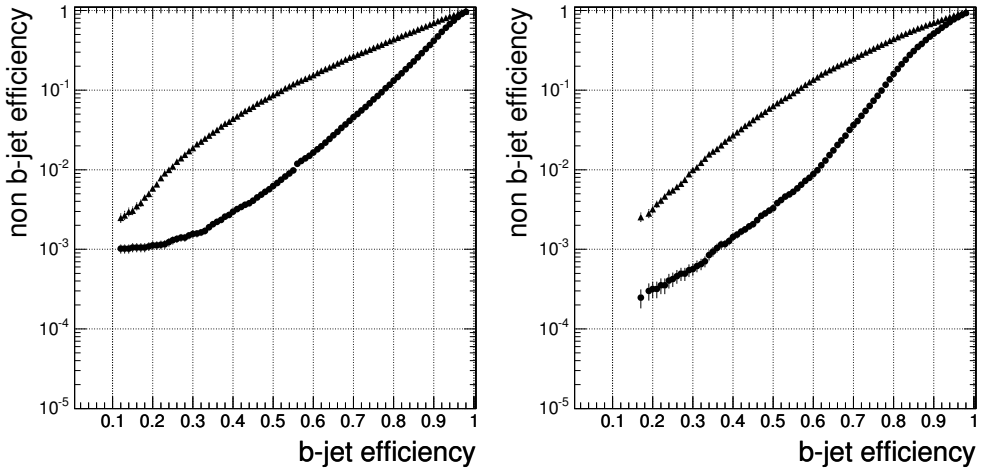


Figure E.2: b-tagging performance of the JetProbability (left) and the CombinedSecondaryVertex (right) algorithm. The distributions are presented for c-jets (triangles) and uds-jets (circles)

The rewritten TrackCounting, JetProbability and CombinedSecondaryVertex algorithms, as presented in chapter 5.2.4 and appendix C.3, perform comparable or slightly better with respect to the results documented in the physics TDR, although the latter two algorithms have not even been trained to determine new adjusted likelihood PDFs which are needed for the calculation of the b-tagging discriminator. The SoftLepton algorithms, however could not reach former mistagging rates yet. This is expected due to the use of a more sophisticated approach where a neural network is involved which has to be carefully trained in order to obtain its maximum performance.

List of Figures

2.1 NLO production cross sections for processes in proton-proton collisions . . 18

2.2 Feynman diagram of a leptonically and hadronically decaying top quark . . 20

2.3 Production cross sections and decay branching fractions for a SM Higgs . . 22

3.1 Aerial view of the LHC 26

3.2 CERN accelerator complex 27

3.3 Schematic view of CMS 34

3.4 Slice of the CMS detector 35

3.5 CMS tracker layout 36

3.6 Schematic layout of a silicon strip detector 37

3.7 CMS ECAL and HCAL layout 38

3.8 CMS muon system layout 40

3.9 CMS data acquisition and trigger system 41

3.10 CMSSW event set-up 43

3.11 Schematic view of the PYTHIA event generation workflow 44

4.1 The WLCG and CMS tier structure 50

4.2 Workflow of a typical WLCG Grid job 54

4.3 Cumulative amount of transferred data during CSA06 60

4.4 Transfer rates from the Tier-0 to the Tier-1 centres 61

4.5 Transfer quality from the Tier-0 to the Tier-1 centres 62

4.6 Transfer rate from the Tier-1 to the Tier-2 centres 63

4.7 Transfer quality from the Tier-1 to the Tier-2 centres 64
4.8 Transfer quality of the German Tier-2 centres 65
4.9 Summary of CSA06, 07 and 08 66
4.10 dCache transfer rates at GridKa during CSA08 66

5.1 Schematic view of a b-hadron decay 69
5.2 Lifetime signed track impact parameter 70
5.3 2D impact parameter value and significance per jet flavour 79
5.4 SV reconstruction efficiencies per jet flavour 80
5.5 SV flight distance significance and SV mass per jet flavour 81
5.6 TrackCounting and CombinedSV discriminator output per jet flavour . . . 82
5.7 b-tagging performance of the CombinedSecondaryVertex algorithm 82
5.8 Comparison of all b-tagging algorithms 83
5.9 Track fake rate versus eta, tight track quality cuts 88
5.10 Performance of the TC (high purity) algorithm, refit vs. reco 88
5.11 Track impact parameter significance per scenario 90
5.12 Track impact parameter value per scenario 91
5.13 Track impact parameter error per scenario 92
5.14 Pull distribution of the 2D impact parameter 93
5.15 Flight distance significance per scenario 95
5.16 Flight distance value per scenario 96
5.17 Flight distance error per scenario 97
5.18 SV based variables with respect to misalignment, b-jets 98
5.19 Relative performance decrease, c-jets 99
5.20 Relative performance decrease, udsg-jets 100
5.21 Performance of the SimpleSV algorithm, 10 pb-1 scenario 102
5.22 Performance decrease of the SimpleSV algorithm, 10 pb-1 scenario 102
5.23 SV reconstruction efficiency versus 3D flight distance 103

5.24 Pull distribution of the 3D flight distance 105
5.25 Flavour tagging efficiency versus discriminator cut, SimpleSV 105
5.26 Flavour tagging efficiency versus discriminator cut, JetProbability 106
5.27 SimpleSV performance with respect to the SV cut sets, 10 pb-1 scenario . . 108
5.28 SimpleSV performance with respect to the SV finder algorithm 109
5.29 APE Variations for the TrackCounting and SimpleSV 110
5.30 SimpleSV performance with respect to the cone size, 10 pb-1 scenario . . . 113
5.31 Performance of lifetime based b-tagging algorithms, part one 114
5.32 Performance decrease of lifetime based b-tagging algorithms, part one . . . 115

C.1 2D and 3D impact parameter significance per jet flavour 135
C.2 Impact parameter value and error per jet flavour 136
C.3 Track multiplicity of a jet per jet flavour 136
C.4 SV flight distance value per jet flavour 137
C.5 SV flight distance error per jet flavour 137
C.6 SV flight distance significance per jet flavour 138
C.7 Invariant mass and energy fraction at the SV per jet flavour 138
C.8 Track multiplicity at the SV per jet flavour 139
C.9 TrackCounting discriminator output per jet flavour 140
C.10 JetProbability and CombinedSV discriminator output per jet flavour . . . 140
C.11 SoftLepton discriminator output per jet flavour 141
C.12 b-tagging performance of the TrackCountingHighEff algorithm 142
C.13 b-tagging performance of the TrackCountingHighPur algorithm 142
C.14 b-tagging performance of the JetProbability algorithm 143
C.15 b-tagging performance of the CombinedSecondaryVertex algorithm 143
C.16 b-tagging performance of the SoftMuon algorithm 144
C.17 b-tagging performance of the SoftElectron algorithm 144
C.18 Track fake rate versus eta, tight track quality cuts 145

C.19 Track fake rate versus eta, loose track quality cuts 145
C.20 Performance of the TC (high purity) algorithm, refit vs. reco 146
C.21 Performance of the TC (high efficiency) algorithm, refit vs. reco 146
C.22 Performance of the JetProbability algorithm, refit vs. reco 147
C.23 Performance of the CombinedSecondaryVertex algorithm, refit vs. reco . . 147
C.24 Impact parameter significance, refit vs. reco 148
C.25 SV track multiplicity with respect to misalignment 148
C.26 SV energy fraction with respect to misalignment 149
C.27 SV mass with respect to misalignment 149
C.28 Performance of the TC (high efficiency) algorithm, misalignment 150
C.29 Performance of the TC (high purity) algorithm, misalignment 150
C.30 Performance of the JetProbability algorithm, misalignment 151
C.31 Performance of the CombinedSecondaryVertex algorithm, misalignment . . 151
C.32 Performance of the SoftMuon algorithm, misalignment 152
C.33 Performance of the SoftElectron algorithm, misalignment 152
C.34 Performance of the SimpleSV algorithm, part one 153
C.35 Performance of the SimpleSV algorithm, part two 154
C.36 Performance decrease of the SimpleSV algorithm, part one 155
C.37 Performance decrease of the SimpleSV algorithm, part two 156
C.38 Flavour tagging efficiency versus discriminator cut, TCHighEff 157
C.39 Flavour tagging efficiency versus discriminator cut, TCHighPur 157
C.40 Flavour tagging efficiency versus discriminator cut, CombinedSV 158
C.41 Flavour tagging efficiency versus discriminator cut, SoftMuon 158
C.42 Flavour tagging efficiency versus discriminator cut, SoftElectron 158
C.43 SimpleSV performance with respect to the SV cut sets, no misalignment . 159
C.44 SimpleSV performance with respect to the SV cut sets, 100 pb-1 160
C.45 SimpleSV performance with respect to the SV cut sets, 1st PL miss. 161
C.46 SimpleSV performance with respect to the SV cut sets, startup 162

C.47 APE Variations for the TrackCounting (high efficiency) algorithm 163
C.48 APE Variations for the TrackCounting (high purity) algorithm 164
C.49 APE Variations for the JetProbability algorithm 165
C.50 APE Variations for the CombinedSecondaryVertex algorithm 166
C.51 APE Variations for the SimpleSecondaryVertex algorithm 167
C.52 SimpleSV performance depending on the cone size, no misalignment 168
C.53 SimpleSV performance depending on the cone size, 100 pb-1 168
C.54 SimpleSV performance depending on the cone size, missing pixel layer . . . 169
C.55 SimpleSV performance depending on the cone size, startup 169
C.56 Performance of lifetime based b-tagging algorithms, part two 170
C.57 Performance decrease of lifetime based b-tagging algorithms, part two . . . 171

D.1 Tracking efficiency for single muons and pions 174
D.2 Tracking efficiency for tracks in b-jets 175
D.3 Resolution of tracking parameters for single muons 175

E.1 b-Tagging Performance in the old framework, QCD 178
E.2 b-Tagging Performance in the old framework, ttbar 179

List of Tables

2.1 Fundamental forces 14

2.2 Fundamental bosons 15

2.3 Fundamental fermions 15

3.1 Energy of the proton beam at each acceleration step 29

3.2 Important LHC parameters in case of proton-proton collisions 30

4.1 Intended and achieved transfer rates from CERN to the Tier-1 centres . . . 59

5.1 Basic track selection cuts 76

5.2 Configuration parameters of the "Adaptive Vertex Finder" 77

5.3 Secondary vertex selection cuts 77

5.4 SV reconstruction efficiencies per jet flavour 80

5.4 SV reconstruction efficiency with respect to misalignment 94

List of Abbreviations

ALICE ..... A Large Ion Collider Experiment

AOD ..... Analysis Object Data

APD ..... Silicon Avalanche Photodiode

APE ..... Alignment Position Error

ATLAS ..... A Toroidal LHC AparatuS

AVF ..... Adaptive Vertex Finder/Fitter

CASTOR ..... CERN Advanced STORage Manager

CE ..... Computing Element

CERN ..... Conseil Européen pour la Recherche Nucléaire

CKM ..... Cabibbo-Kobayashi-Maskawa

CMS ..... Compact Muon Solenoid

CMSSW ..... CMS SoftWare

CRAB ..... CMS Remote Analysis Builder

CSA06 ..... Computing, Software and Analysis Challenge 2006

CSC ..... Cathode Strip Chamber

CTF ..... Combinatorial Track Finder

DAQ ..... Data AcQuisition

DBS ..... Dataset Bookkeeping System

DLS ..... Data Location Service

DPM ..... Disk Pool Manager

DT ..... Drift Tube chamber

EB ..... ECAL Barrel detector

ECAL ..... Electromagnetic Calorimeter

EDM ..... Event Data Model

EE ..... ECAL Endcap detector

EGEE ..... Enabling Grids for E-sciencE

FTS ..... File Transfer Service

GOC ..... Grid Operations Centre

HB ..... HCAL Barrel detector

HCAL ..... Hadron Calorimeter

HE ..... HCAL Endcap detector

HF ..... HCAL Forward detector

HLT ..... High Level Trigger

HO ..... HCAL Outer detector

IOV ..... Interval Of Validity

IP ..... Impact Parameter

IP ..... Interaction point

IS ..... Information Service

JDL ..... Job Description Language

LEP ..... Large Electron-Positron Collider

Level-1 ..... Level-1 trigger

LFC ..... Logical File Catalog

LFN ..... Logical File Name

LHC ..... Large Hadron Collider

LHCb ..... Large Hadron Collider beauty

LHCf ..... Large Hadron Collider forward

LINAC2 ..... CERN Linear Accelerator 2

MB ..... Muon Barrel

MC ..... Monte Carlo

MDS ..... Globus Monitoring and Discovery Service

ME ..... Muon Endcap

MoU ..... Memorandum of Understanding

MSSM ..... Minimal Supersymmetric Standard Model

MTCC ..... Magnet Test and Cosmic Challenge

PDF ..... Probability Density Function

PhEDEx ..... Physics Experiment Data Export

PHF ..... Pixel Hit Filter

ProdAgent ..... CMS Monte Carlo Production Agent

PS ..... Proton Synchrotron

PSB ..... Proton Synchrotron Booster

PV ..... Primary Vertex

PXB ..... Pixel Barrel

PXE ..... Pixel Endcap

QCD ..... Quantum ChromoDynamics

QED ..... Quantum ElectroDynamics

QFD ..... Quantum FlavourDynamics

R-GMA ..... Relational Grid Monitoring Architecture

RB ..... Resource Broker

RFQ ..... Radio Frequency Quadrupole

RPC ..... Resistive Plate Chamber

RS ..... Road Search

SE ..... Silicon Electromagnetic calorimeter

SE ..... Storage Element

SM ..... Standard Model of particle physics

SPS ..... Super Proton Synchrotron

SRM ..... Storage Resource Manager

SSL ..... Secure Sockets Layer

SUSY ..... SUperSymmetry

SV ..... Secondary Vertex

TDR ..... Technical Design Report

TEC ..... Tracker Endcap

TFC ..... Trivial File Catalog

TIB ..... Tracker Inner Barrel

TID ..... Tracker Inner Disk

TKF ..... Trimmed Kalman Fitter

TKVF ..... Trimmed Kalman Vertex Finder

TMDB ..... Transfer Management Database

TOB ..... Tracker Outer Barrel

TOTEM ..... TOTal Elastic and diffractive cross section Measurement

UI ..... User Interface

VO ..... Virtual Organisation

VPT ..... Vacuum Phototriode

WLCG ..... Worldwide LHC Computing Grid

WLS ..... WaveLength-Shifting

WMS ..... Workload Management System

Bibliography

[1] David Griffiths. *Introduction to Elementary Particles*, 2004. ISBN 0-471-60386-4.

[2] Donald H. Perkins. *Hochenergiephysik*, 1991. ISBN 3-89319-236-0.

[3] The Particle Data Group. *Particle Physics Booklet - July 2006*, 2006. LBL-61033.

[4] S. Catani et al. *QCD*, 2000. CERN-TH/2000-131 [arXiv:hep-ph/0005025v1].

[5] The CMS Collaboration. *CMS physics : Technical Design Report II*, 2006. CERN-LHCC-2006-021.

[6] J. Fernandez. *Search for MSSM heavy neutral Higgs bosons in the four-b final state*, 2006. CMS Note 2006-023.

[7] Povh, Rith, Scholz, and Zetsche. *Teilchen und Kerne*, 2006. Springer, ISBN 3-540-36685-7.

[8] CERN Editorial Board. *LHC Design Report - Volume I: The LHC Injector Chain*, 2004. CERN-2004-003-V-1.

[9] CERN Editorial Board. *LHC Design Report - Volume II: The LHC Infrastructure and General Services*, 2004. CERN-2004-003-V-2.

[10] CERN Editorial Board. *LHC Design Report - Volume III: The LHC Main Ring*, 2004. CERN-2004-003-V-3.

[11] CERN Communication Group. *CERN FAQ LHC the guide*, 2008. CERN-Brochure-2008-001-Eng.

[12] The CMS Collaboration. *CMS physics : Technical Design Report*, 2006. CERN-LHCC-2006-001.

[13] The CMS Collaboration. *CMS computing : Technical Design Report*, 2005. CERN-LHCC-2005-023.

[14] The CMS Collaboration. *The CMS experiment at the CERN LHC*, 2008. Journal of Instrumentation - JINST, in press.

[15] The CMS Collaboration. *The Tracker Project Technical Design Report*, 1998. CERN-LHCC-98-006.

[16] L. Borrello, A. Messineo, E. Focardi, and A. Macchiolo. *Sensor design for the CMS Silicon Strip Tracker*, 2003. CMS Note 2003-020.

[17] The CMS Collaboration. *The Electromagnetic Calorimeter Technical Design Report*, 1997. CERN-LHCC-97-033.

[18] The CMS Collaboration. *The Hadron Calorimeter Technical Design Report*, 1997. CERN-LHCC-97-031.

[19] The CMS Collaboration. *The Muon Project Technical Design Report*, 1997. CERN-LHCC-97-032.

[20] The CMS Collaboration. *The TriDAS Project Technical Design Report, Volume 1: The Trigger Systems*, 2000. CERN-LHCC-2000-38.

[21] The CMS Collaboration. *The TriDAS Project Technical Design Report, Volume 2: Data Acquisition and High-Level Trigger*, 2002. CERN-LHCC-2002-26.

[22] The CMS collaboration. *The CMS Magnet Test and Cosmic Challenge (MTCC Phase I and II) Operational Experience and Lessons Learnt*, 2007. CMS Note 2007-005.

[23] The CMS Tracker Group. *Tracker Operation and Performance at the Magnet Test and Cosmic Challenge*, 2007. CMS Note 2007-029.

[24] *CMS Software*. https://twiki.cern.ch/twiki/bin/view/CMS/SWGuide.

[25] *ROOT*. http://root.cern.ch.

[26] *PROOF*. http://root.cern.ch/twiki/bin/view/ROOT/PROOF.

[27] Torbjörn Sjöstrand. *Pythia 6.4 Physics and Manual*, 2006. hep-ph/0603175.

[28] G. Corcella, I.G. Knowles, G. Marchesini, S. Moretti, K. Odagiri, P. Richardson, M.H. Seymour, and B.R. Webber. *HERWIG 6.5*, 2005. hep-ph/0011363, hep-ph/0210213.

[29] *Sherpa Event Generator*. http://www.sherpa-mc.de.

[30] *Geant4*. http://geant4.web.cern.ch/geant4.

[31] *The CMS Fast Simulation*. https://twiki.cern.ch/twiki/bin/view/CMS/WorkBookFastSimulation.

[32] W. Adam, B. Mangano, Th. Speer, and T. Todorov. *Track Reconstruction in the CMS tracker*, 2006. CMS Note 2006-041.

[33] *Development of the RoadSearch Track Reconstruction*. https://twiki.cern.ch/twiki/bin/view/CMS/RoadSearch.

[34] R. Frühwirth. *Application of Kalman Filtering to Track and Vertex Fitting*, 1987. doi:10.1016/0168-9002(87)90887-4.

[35] A. Heister, O. Kodolova, V. Konoplianikov, S. Petrushanko, J. Rohlf, C. Tully, and A. Ulyanov. *Measurement of Jets with the CMS Detector at the LHC*, 2006. CMS NOTE 2006-036.

[36] Gerald C. Blazey et al. *Run II jet physics*, 2000. hep-ex/0005012.

[37] S. V. Chekanov. *Jet algorithms: A mini review*, 2002. hep-ph/0211298.

[38] G. Salam and G. Soyez. *A practical seedless infrared-safe cone jet algorithm*, 2007. JHEP05 (2007) 086.

[39] J. M. Butterworth, J. P. Couchman, B. E. Cox, and B. M. Waug. *KtJet: A C++ implementation of the kT clustering algorithm*, 2003. hep-ph/0210022.

[40] G. Bruno, B. V. de Vyver, P. Cox, S. Lacaprara, N. Neumeister, S. Villa, and R. Wilkinson. *Local Reconstruction in the Muon Detectors*, 2002. CMS Note 2002-043.

[41] E. Meschi, T. Monteiro, C. Seez, and P. Vika. *Electron Reconstruction in the CMS Electromagnetic Calorimeter*, 2001. CMS Note 2001-034.

[42] S. Baffioni, C. Charlot, F. Ferri, D. Futyan, P. M. I. Puljak, C. Rovelli, R. Salerno, and Y. Sirois. *Electron reconstruction in CMS*, 2006. CMS Note 2006-040.

[43] R. Frühwirth and T. Speer. *A Gaussian-sum filter for vertex reconstruction*, 2004. doi:10.1016/j.nima.2004.07.090.

[44] *The Worldwide LHC Computing Grid*. http://lcg.web.cern.ch/LCG/.

[45] *Enabling Grids for E-sciencE*. http://eu-egee.org/.

[46] *CERN Advanced STORage Manager*. http://castor.web.cern.ch/castor/.

[47] LHC and WLCG Collaboration. *Memorandum of Understanding for Collaboration in the Deployment and Exploitation of the Worldwide LHC Computing Grid*, 2005. CERN-C-RRB-2005-01/Rev, http://lcg.web.cern.ch/LCG/C-RRB/MoU/WLCGMoU.pdf.

[48] *Lightweight Middleware for Grid Computing*. http://glite.web.cern.ch/glite/.

[49] *Scientific Linux*. http://www.scientificlinux.org/.

[50] *Grid Operations Centre*. http://www.ukiroc.eu/content/view/115/235/.

[51] *The dCache Project*. http://www.dcache.org/.

[52] *Globus Monitoring and Discovery Service*. http://www.globus.org/toolkit/mds/.

[53] *Relational Grid Monitoring Architecture*. http://www.r-gma.org/.

[54] A. Scheurer, M. Ernst, A. Flossdorf, C. Hof, T. Kress, K. Rabbertz, and G. Quast. *Challenges of the LHC Computing Grid by the CMS experiment*, 2007. ID: 316623.0, Max Planck Digital Library / German e-Science Conference.

[55] The CMS Collaboration. *CMS Computing, Software and Analysis Challenge in 2006 (CSA06) Summary*, 2007. CMS Note 2007-006. CERN-LHCC 2007-010. LHCC-G-128.

[56] *The CMS Production Agent*. https://twiki.cern.ch/twiki/bin/view/CMS/ProdAgent.

[57] *CMS Remote Analysis Builder*. https://twiki.cern.ch/twiki/bin/view/CMS/SWGuideCrab.

[58] *CMS Data Bookkeeping System*. https://twiki.cern.ch/twiki/bin/view/CMS/DBS-TDR.

[59] *CMS Data Location Service*. https://twiki.cern.ch/twiki/bin/view/CMS/DLS.

[60] *Physics Experiment Data Export*. http://cmsdoc.cern.ch/cms/aprom/phedex/.

[61] *Frontier Webpage*. http://frontier.cern.ch/.

[62] *Squid Webpage*. http://www.squid-cache.org/.

[63] C. Saout, A. Scheurer, F.-P. Schilling, and A. Schmidt. *Impact of Tracker Misalignment on the CMS b-Tagging Performance*, 2008. CMS AN-2007-047.

[64] C. Saout, A. Scheurer, F.-P. Schilling, and A. Schmidt. *Impact of Tracker Misalignment on the CMS b-Tagging Performance*, 2008. CMS PAS BTV-07-003.

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

[66] C. Weiser. *A Combined Secondary Vertex Based B-Tagging Algorithm in CMS*, 2006. CMS Note 2006-014.

[67] A. Bocci, P. Demin, R. Ranieri, and S. de Visscher. *Tagging b jets with electrons and muons at CMS*, 2006. CMS Note 2006-043.

[68] R. Frühwirth, P. Kubinec, W. Mitaroff, and M. Regler. *Vertex reconstruction and track bundling at the LEP collider using robust algorithms*, 1996. doi:10.1016/0010-4655(96)00040-9.

[69] T. Speer et al. *Vertex Fitting in the CMS Tracker*, 2006. CMS Note 2006-032.

[70] P. Vanlaer, V. Barbone, N. De Filippis, T. Speer, O. Buchmueller, and F.-P. Schilling. *Impact of CMS Silicon Tracker Misalignment on Track and Vertex Reconstruction*, 2006. CMS Note 2006-029.

[71] T. Muller, C. Piasecki, G. Quast, and C.Weiser. *Inclusive Secondary Vertex Reconstruction in Jets*, 2006. CMS Note 2006-027.

[72] P. Schleper, G. Steinbrück, and M. Stoye. *Software Alignment of the CMS Tracker using MILLEPEDE II*, 2006. CMS Note 2006-011.

[73] V. Karimaki, T. Lampen, and F.-P. Schilling. *The HIP Algorithm for Track Based Alignment and its Application to the CMS Pixel Detector*, 2006. CMS Note 2006-018.

[74] E. Widl, R. Frühwirth, and W. Adam. *A Kalman Filter for Track-based Alignment*, 2006. CMS Note 2006-022.

[75] G. Flucke, P. Schleper, and M. Stoye G. Steinbrück. *A Study of Full Scale CMS Tracker Alignment using High Momentum Muons and Cosmics*, 2008. CMS Note 2008-008.

[76] N. de Filippis, T. Lampen, F.-P. Schilling, A. Schmidt, and M.Weber. *Update of Misalignment Scenarios for the CMS Tracker*, 2007. CMS IN 2007-036.

[77] T. Lampen, M. Weber, N. de Filippis, and A. Schmidt. *Misalignment Scenarios for the Startup Conditions of the CMS Tracker*, 2007. CMS Note 2007-061.

[78] T. Lampen, N. de Filippis, F.-P. Schilling, A. Schmidt, and M. Weber. *Comprehensive Set of Misalignment Scenarios for the CMS Tracker*, 2008. CMS Note in preparation.

[79] The CMS Collaboration. *Performance Measurement of b-tagging Algorithms Using Data containing Muons within Jets*, 2008. CMS PAS BTV-07-001.

[80] The CMS Collaboration. *Evaluation of udsg Mistags for b-tagging using Negative Tags*, 2008. CMS PAS BTV-07-002.

Danksagung

Ich möchte mich in erster Linie bei Prof. Dr. Günter Quast bedanken, der mich in den vergangenen Jahren in verschiedensten Belangen hervorragend und kompetent betreut hat. Vor allem die zahlreichen Aufenthalte am CERN, die mir einen direkten Zugang zum Experiment erlaubt haben, wären ohne seine Unterstützung nicht möglich gewesen.

Ich danke weiterhin Prof. Dr. Michael Feindt für die Übernahme des Korreferats und die gute Zusammenarbeit, die schon mit dem Beginn meiner Diplomarbeit unter seiner Betreuung begann.

Des Weiteren möchte ich meinen Kollegen Christophe Saout, Dr. Frank-Peter Schilling und Dr. Alexander Schmidt für die gute Zusammenarbeit und ihre Unterstützung, nicht nur im Bereich des b-Tagging, danken.

Den folgenden Personen, die mich bei der Korrektur dieser Arbeit unterstützt haben, danke ich vielmals für ihre Mühe, ihre investierte Zeit und ihre wertvollen Anmerkungen: Michael Heinrich, Simone Herberger, Benjamin Klein, Viktor Mauch, Oliver Oberst, Andreas Oehler, Christophe Saout, Fred-Markus Stober und Manuel Zeise.

Ich danke außerdem Volker Büge, der mit mir in den letzten Wochen und Monaten das gleiche Schicksal geteilt und damit die Zeit deutlich erträglicher gemacht hat.

Ich danke Volker Büge, Dr. Ulrich Felzmann, Dr. Joachim Heuser, Jasmin Kiefer, Dr. Philipp Mack, Dr. Michael Milnik, Oliver Oberst, Andreas Oehler, Dr. Svenja Richter, Christophe Saout, Dr. Andreas Scharf und Dr. Thorsten Scheidle für die verschiedensten Aktivitäten, auch außerhalb des Instituts.

Ich danke dem gesamten Institut für Experimentelle Kernphysik und vor allem der CMS Gruppe für die tolle Atmosphäre in den letzten Jahren. Außerdem verdienen alle, die ihre

Zeit zum Wohle des Instituts eingesetzt haben, ein aufrichtiges Dankeschön: die Administratoren, die Hardware-Beauftragten und die Getränke-Besteller.

Ich danke meinen Freunden und vor allem Julian Gonska und Marwin Rieger für Ihre langjährige Freundschaft.

Mein ganz besonderer Dank gilt meiner Familie, vor allem meinen Eltern und meiner Freundin Simone für die uneingeschränkte Unterstützung in jeder Hinsicht!

Schließlich möchte ich noch all denen danken, die ich in dieser Liste vergessen haben sollte!