Implementation of a Model-Independent Search for New Physics with the CMS Detector Exploiting the World-Wide LHC Computing Grid

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Diplom-Physiker

Carsten Hof

aus Betzdorf (Sieg)

Berichter: Univ.-Prof. Dr. Thomas Hebbeker
Univ.-Prof. Dr. Christopher Wiebusch

Tag der mündlichen Prüfung: 04.12.2009

Abstract

With this year's start of CERN's Large Hadron Collider (LHC) it will be possible for the first time to directly probe the physics at the TeV-scale at a collider experiment. At this scale the Standard Model of particle physics will reach its limits and new physical phenomena are expected to appear.

This study performed with one of the LHC's experiments, namely the Compact Muon Solenoid (CMS), is trying to quantify the understanding of the Standard Model and is hunting for deviations from the expectation by investigating a large fraction of the CMS data. While the classical approach for searches of physics beyond the Standard Model assumes a specific theoretical model and tries to isolate events with a certain signature characteristic for the new theory, this thesis follows a model-independent approach.

The method relies only on the knowledge of the Standard Model and is suitable to spot deviations from this model induced by particular theoretical models but also theories not yet thought of. Future data are to be compared to the expectation in several hundreds of final state topologies and a few variables of general sensitivity to deviations like invariant masses. Within this feasibility study, events are classified according to their particle content (muons, electrons, photons, jets, missing energy) into so called event classes. A broad data scan is performed by investigating distributions searching for significant deviations from the Standard Model. Systematic uncertainties are rigourously taken into account within the analysis. Several theoretical models such as supersymmetry, new heavy gauge bosons and microscopic black holes as well as possible detector effects in the early data have been fed into the search algorithm as benchmark scenarios and proof the ability to supplement the traditional model-driven searches.

Due to the enormous computing resource required for such an analysis performing a multitude of classical analyses in parallel the approach would not be feasible without the increasing performance and decreasing costs of modern computing systems. The LHC and its experiments with expected data rates of several 10 PetaBytes per year face this challenge with a distributed, locally organized computing and storage network: the LHC Computing Grid. The CMS tools embedded in such an environment and its application are demonstrated within this work.

Zusammenfassung

Der diesjährige Start des Large Hadron Colliders (LHC) am CERN ermöglicht es erstmals auf direktem Wege die Physik an der TeV-Skala zu untersuchen. An dieser Skala stößt das Standard-Modell der Elementarteilchenphysik an seine Grenzen und man erwartet die Entdeckung neuer Phänomene.

Die vorliegende Studie wurde an einem der LHC Experimente, dem Compact Muon Solenoid (CMS), durchgeführt und hat zur Aufgabe, das Verständnis des Standard-Modells zu überprüfen und Abweichungen von den Erwartungen aufzuspühren. Dazu wird ein Großteil der zukünftig aufgezeichneten Daten untersucht. Während die klassische Suche nach neuer Physik in der Hochenergiephysik ein spezielles theoretisches Modell annimmt und versucht Ereignisse mit einer für die Theorie spezifischen passenden Signatur zu finden, verfolgt diese Analyse einen neueren model-unabhängigen Ansatz.

Diese Methode beruht nur auf der Annahme des Standard-Modells und ist daher in der Lage Abweichungen zu finden, die durch bestimmte theoretische Modelle, aber auch durch Theorien, die bisher noch nicht formuliert wurden, beschrieben werden. Dazu werden die Daten in mehreren hunderten von Endzuständen jeweils auf Abweichungen überprüft. Hierzu werden Variablen wie invariante Massen, von denen man erwartet, dass sie besonders sensitiv auf Physik jenseits des Standard Models sind, untersucht. Die Ereignisse werden anhand der gemessenen Teilchenarten (Myonen, Elektronen, Photonen, Jets und fehlende transversale Energie) und Häufigkeiten in sogenannte Ereignisklassen einsortiert. Innerhalb dieser Klassen werden in bestimmten Verteilungen und mit einem dedizierten Such-Algorithmus Abweichungen von der Standard-Modell-Erwartung gesucht.

Besondere Aufmerksamkeit wird den systematischen Unsicherheiten gewidmet, da ihre Berücksichtigung ein kritischer Bestandteil einer jeden Analyse ist. Dies stellt jedoch bei einer so generischen Suche eine anspruchsvolle Herausforderung dar. Der Algorithmus und sein Potential wurde mit einer Reihe von theoretischen Modellen unter Beweis gestellt: unter anderem Supersymmetrie, neue schwere Eichbosonen und Leptoquarks, aber auch mögliche Detektor-Effekte. Der breite Anwendungsbereich zeigt die vielfältigen Möglichkeiten der Analyse auf und demonstriert ihr Potential, die traditionellen modellbasierten Suchen zu ergänzen.

Auf Grund der enormen Rechenkapazitäten, die eine solche Analyse bedarf, ist es erst im letzten Jahrzehnt durch die steigenden Rechenleistungen und fallenden Preise moderner Computersysteme möglich eine solche Vielzahl von parallelen Einzelanalysen durchzuführen. Der LHC und seine Experimente, die ein jährliches Datenvolumen von mehreren 10 Peta-Bytes erwarten, haben hierzu ein verteiltes, dezentral organisiertes Rechen- und Speichernetzwerk entwickelt und installiert: das weltweite LHC Computing Grid.

Contents

Introduction

The introduction marks a special entry point to every thesis, not to say to every book. In that sense the author needs to think of a clever and interesting way of introducing the work done. Many theses try this by quoting wise persons or cite the ever existing human aspiration to gain deep insight into nature questing for the "Theory of Everything". The glory details of history are stated in many text books and thus will not be repeated here. Instead we will start with a comprehensive overview of our current understanding and modelling of nature within elementary particle physics.

To our today's knowledge four fundamental forces, the electromagnetic, the weak, the strong and the gravitational force, interact between twelve elementary fermionic particles by mediating bosonic particles. All forces except gravity have been implemented in the framework of gauge theories, combined in the "Standard Model of Particle Physics" (SM). The gauge groups model the fundamental degrees of freedom and reflect the importance of the underlying symmetries. The identification of basic symmetries has played a crucial role in the description of the fundamental reactions up to a very precise level. Nonetheless, the Standard Model - as every appealing scientific model - already points to its limitations: it does not incorporate gravity and thus must fail at least at energy scales where gravity is not negligible any more (Planck scale). But already at the terra-scale, which will be probed by CERN's Large Hadron Collider (LHC) and its experiments, questions beyond the Standard Model might be solved: How do particles acquire mass? How are the symmetries of the Standard Model broken? Has nature realized even more symmetries then the ones incorporated within the Standard Model? Searches at colliders might even provide solutions to astronomical challenges, such as explanations for the origin of the asymmetry between particles and their counter-partners, the anti-particles, or discover new particles explaining the unknown "dark matter" or even the dark energy within the universe. The LHC might produce these particles so that its properties can be measured in order to gain a deeper understanding of our universe and its origin.

Theorists have thought of many extensions of the Standard Model, but for none of them appealing hints are visible. The traditional approach for searches at collider experiments are "theory driven": the data are investigated for promising signatures predicted by a certain theoretical model. Special selection criteria are developed to distinguish those events from already known processes.
In this feasibility study a different, model-independent approach is developed. Without any theoretical bias beyond the Standard Model, events are classified by their particle contents (number of electrons, muons, particle jets etc.). Within each class characteristic variables

are used to compare the data to the expectations provided by the Standard Model and implemented in Monte Carlo event generators.

This generic approach has the advantage of performing a broad data scan, which might be able to spot discrepancies where no dedicated signature-driven analysis is at hand. Assuming only the SM, this strategy is not bound to a special theory. Of course the generality has its price: investigating so many regions of the phase space it is not possible to look at every detail as accurate as a dedicated search. Nonetheless, this model-independent approach should not be seen as a concurrent strategy or even a replacement for conventional analyses. It is a valuable complementary approach with a variety of possible interplays between this ansatz and a conventional analysis: dedicated analyses provide measurements of useful parameters as input to the model-independent search which might in return check their validity in a broader context. On the other hand every interesting deviation spotted by the model-independent search might trigger a dedicated analysis, investigating the discrepancy in greater detail.

The analysis of data via a model-independent approach would not be possible without the developments in computer science and industry. Hardware prices rapidly decrease, fast networks connect the world and the World Wide Web provides the basis of information sharing. Still the demands of the LHC and its detectors as the world largest experiment exceed the present capacities, leading to the development of the Worldwide LHC Computing Grid (WLCG). It provides computing resources, utilizable without the need to know where the data are located or where the calculation is actually done. In a hierarchial model of computing centres the WLCG provides the needed resources to reconstruct the several ten million Gigabytes of raw data taken from the LHC experiments annually. Further it is used to safely store and uniformly distribute the data around the world to finally provide them for the physicists to analyse.

This thesis covers the development of a model-independent analysis at one of the LHC experiments, the general-purpose Compact Muon Solenoid (CMS) detector. After a review of the Standard Model and possible extensions, the experimental machinery consisting of the LHC, the CMS detector and the WLCG Computing Grid will be introduced. The following chapters cover the basic principles and details of the search strategy and its implementation. Exemplary models beyond the Standard Model such as Supersymmetry are fed as benchmark tests into the search algorithm and illustrate the feasibility and power of such a method. Of course the true value of such an analysis will only be proven once the LHC starts taking data, which is close, but still out of scope for this thesis.

Chapter 1

The Standard Model

During the last century the fundamental constituents of matter and the interactions among them have been merged into one model of great beauty and simplicity known as the Standard Model of Particle Physics. All known particles seem to be built up from quarks and leptons, which are to today's knowledge point like, structureless, spin-1/2 particles (fermions). The interactions among them can be classified into four categories: gravitation, weak, electromagnetic and strong interaction, where the former one can be neglected at distances viewed at in particle physics. They vastly differ in their range: whereas the electromagnetic and gravitational force act over infinite distances, weak and strong interactions are limited to a very small region.

Force	Range [m]	Relative strength	Force carrier
Strong force	10^{-15}	1	8 gluons (g)
Electromagnetic force	∞	10^{-2}	photon (γ)
Weak force	10^{-18}	10^{-2}	W, Z^0
Gravitational force	∞	10^{-40}	graviton (?)

Table 1.1: *The fundamental forces sorted by their relative strengths and the force carrying bosons.*

Beside the gravitational force, all interactions can be described by local gauge theories, where the forces are carried by fundamental spin-1 gauge bosons. The gravitation is expected to be mediated through a spin-2 boson called graviton, but no direct evidence for this particle has been found so far. Quarks, which do not exist freely in nature (see section 1.2) participate in all interactions, whereas leptons do not take part in strong interactions. The systematics of weak interactions with charged leptons such as the β-decay motivate the pairing of leptons in three families (see section 1.3). This is not only a nice ordering, but reflects basic symmetries of nature. For every generation an additive quantum number can be defined, which is conserved in all investigated reactions of fundamental particles (lepton number conservation).

A priori all particles within the Standard Model are massless. This is no problem for the massless photons and gluons, but all fermions and the weak force mediating W and Z^0 bosons are known to have a mass. The problem is possibly solved by a spontaneous symmetry breaking and the so called "Higgs-mechanism". It demands a new fundamental

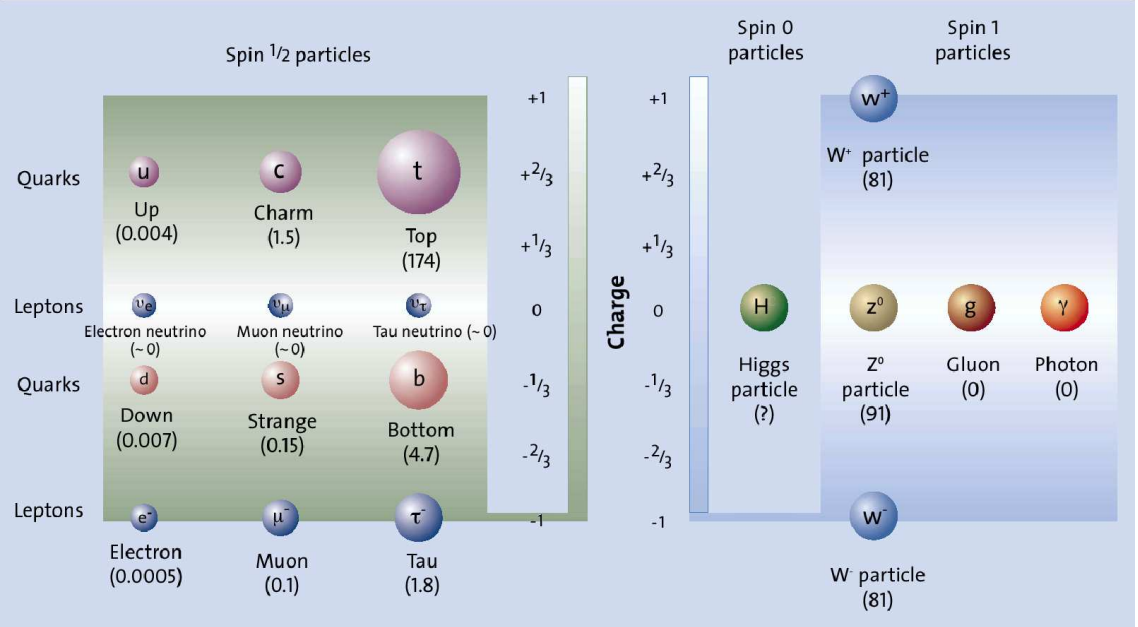

Figure 1.1: *Overview of the Standard Model particles, their charge and mass (in parenthesis) [1].*

spin-0 particle, the Higgs boson, whose discovery or exclusion is one of the main tasks of the LHC and the experiments located there.

This chapter gives a brief overview of the successes of the SM, but will also enlight the facts where it hits its limitations. The Standard Model has been validated in precision measurements, but for sure new physics will enter at the electroweak scale to be probed by the LHC.

1.1 The Standard Model of Particle Physics

The Standard Model of elementary particle physics is based on the Glashow-Salam-Weinberg model of the weak interaction and Quantum Chromodynamics (QCD). A favoured supplement is the Higgs-mechanism and the Higgs particle, which provides a way to give mass to particles. For a detailed introduction into the Standard Model see for example [2–6].

The fundamental particles within the Standard Model are described by space-time coordinate dependent fields $\psi(x)$. Symmetries observed in nature are mathematically reflected by the fact, that the solution of the equation of motion does not change under a certain unitary transformation[1]. In other words: a theory is invariant under a symmetry group G represented by a unitary operator U if the fields $\psi(x)$ and $\psi'(x)$ given by

$$\psi(x) \rightarrow \psi'(x) = U\psi(x) \tag{1.1}$$

[1] A transformation U is unitary, if the adjoint operator $U^\dagger$ equals the inverse operator U^{-1}. This is equivalent to the preservation of the inner product (within a Hilbert space) for all vectors x and y: $\langle Ux, Uy\rangle = \langle x, y\rangle$.

follow the same equation of motion.

In the framework of a Lagrangian field theory with a given Lagrangian $\mathcal{L}(\psi_i, \partial_\mu\psi_i)$ as a function of the fields ψ_i and their first derivatives $\partial_\mu\psi_i$ the equation of motion is the Euler-Lagrange equation

$$\frac{\delta\mathcal{L}}{\delta\psi_i} = \partial_\mu\left(\frac{\delta\mathcal{L}}{\delta(\partial_\mu\psi_i)}\right). \tag{1.2}$$

It can be derived by minimizing the action S, which is a functional of ψ_i and $\partial_\mu\psi_i$

$$S = \int d^4x\ \mathcal{L}\left(\psi_i, \partial_\mu\psi_i\right). \tag{1.3}$$

A symmetry acting in the way

$$\psi \to \psi + \delta\psi, \qquad \partial_\mu\psi_i \to \partial_\mu\psi_i + \delta\partial_\mu\psi_i, \qquad \mathcal{L} \to \mathcal{L} + \delta\mathcal{L} =: \mathcal{L} + \alpha\partial_\mu\mathcal{J}^\mu(x) \tag{1.4}$$

is exact if

$$\delta\mathcal{L} = 0. \tag{1.5}$$

Note that one can allow for an arbitrary divergence term ($\alpha\partial_\mu\mathcal{J}^\mu(x)$) as this leaves the equation of motion derived from the Euler-Lagrange equation (1.2) unchanged. Associated with each exact symmetry is a so called Noether current $j^\mu(x)$ and a corresponding charge Q

$$j^\mu(x) = \frac{\delta\mathcal{L}}{\delta(\partial_\mu\psi)}\delta\psi - \mathcal{J}^\mu \qquad \text{and} \qquad Q = -i\int d^3x\ j^0(x), \tag{1.6}$$

which are conserved

$$\partial_\mu j^\mu = 0 \qquad \text{and} \qquad \frac{dQ}{dt} = 0 \qquad \text{if } \delta\mathcal{L} = 0. \tag{1.7}$$

The gauge symmetries of the Standard Model are all local ones. From an aesthetic point of view this appears much more plausible, since global symmetries act on different space-time points in an exact manner - no matter how far they are separated or how they are causally connected. The local symmetries are implemented by making parameters of the gauge group G and thus their representations $U = U(x)$ space-time dependent. The elements $U(x)$ of G can be expressed in terms of the generators Λ_a via

$$\psi(x) \to U(x)\psi(x) = e^{i\epsilon_a(x)\Lambda_a}\psi(x). \tag{1.8}$$

They satisfy the Lie-Algebra with the structure constants f_{abc}

$$[\Lambda_a, \Lambda_b] = if_{abc}\Lambda_c, \tag{1.9}$$

whose knowledge is sufficient to construct the whole group.

Any Lagrangian containing derivatives, like the Lagrangian for a free particle

$$\mathcal{L}_{\text{free}} = \overline{\psi}(i\gamma_\mu\partial^\mu - m)\psi, \tag{1.10}$$

is not invariant under local gauge transformations[2]. A solution known as minimal substitution is the replacement of the derivative ∂_μ by a covariant derivative D_μ, which satisfies

$$D_\mu\psi(x) \to e^{i\epsilon_a(x)\Lambda_a}D_\mu\psi(x). \tag{1.11}$$

[2] $\partial_\mu\psi(x) \to e^{i\epsilon_a(x)\Lambda_a}\partial_\mu\psi(x) + i\partial_\mu\epsilon_a(x)\Lambda_a e^{i\epsilon_a(x)\Lambda_a}\psi(x)$ spoils gauge invariance.

For this purpose it is necessary to introduce a vector field A_μ

$$D_\mu \psi(x) := (\partial_\mu + i\epsilon_a(x) A_\mu(x))\psi(x), \tag{1.12}$$

which transforms under the unitary operator $U(x)$ (see eqation (1.8)) as

$$A_\mu(x) \to A_\mu - \partial_\mu \epsilon_a(x). \tag{1.13}$$

In addition a kinematic term for the field A_μ has to be added to the Lagrangian. The process of restoring the gauge invariance of the Lagrangian and choice of the vector field A_μ is called gauging.

Aiming towards the understanding of quarks and leptons and their interactions in a framework of a local gauge theory, one has to discover the underlying fundamental symmetries of the different forces, i.e. to identify the basic degrees of freedom on which the symmetries operate. As will be discussed in the following sections the Standard Model is based on the gauge group

$$\mathrm{SU(3)_C} \times \mathrm{SU(2)_L} \times \mathrm{U(1)_Y}.$$

The former term describes the colour degree of freedom of the theory of quarks, quantum chromodynamics. The rest reflects the symmetry of the electroweak unification of the weak and electromagnetic force, with its charges of weak isospin T_3 and electric charge Q, respectively. They are connected to the quantum number Y (hypercharge) related to the $\mathrm{U(1)_Y}$ symmetry via the Gell-Mann-Nishijima formula

$$Q = T_3 + \frac{Y}{2}. \tag{1.14}$$

1.2 Quantum Chromodynamics

The introduction of quarks[3] (spin-1/2, fractional charge) as constituents of hadrons, divided into (anti-)baryons as three-(anti-)quark-states and also mesons as quark-antiquark-states, can describe the huge variety of particles (Gell-Mann and Zweig). The ordering of the spectrum in the baryon-meson world is achieved by the assignment of a degree of freedom to the quarks known as flavour. This global flavour symmetry, which is also retrieved in the lepton sector ("quark-lepton-symmetry"), is described by the gauge group SU(6). Due to the different charges and masses of the quarks and leptons the flavour symmetry is only an approximate one.

Historically, the concept of quark substructures showed two significant problems: first, free quarks have never been observed and second, baryons with three equal quarks, such as Δ^{++} violate the Pauli principle. In 1964, only one year after the proposal of the quark-model, this drawback was bypassed by the introduction of a new "hidden" quantum number, called colour, which can hold the three values red (R), green (G) and blue (B) plus

[3] Many famous physicists denied the existence of quarks as particles for a long time and treated them only as a formal concept. In 2004 D. J. Gross, H. D. Politzer and F. Wilczek won the Nobel prize for the discovery of the asymptotic freedom in the theory of strong interactions.

their three counterparts ($\overline{\mathrm{R}}$, $\overline{\mathrm{G}}$, $\overline{\mathrm{B}}$). (Anti-)quarks carry (anti-)colour, whereas the known hadrons appear as "colourless" particles. Thus the hadrons transform as colour-singlets under this new degree of freedom based on the gauge group $\mathrm{SU(3)_C}$. In general colour can be interpreted as the charge of the strong interaction. In analogy to optics a mixture of (anti-)red, (anti-)green and (anti-)blue quarks in case of baryons or colour plus anti-colour in case of mesons results in a "white" particle. The non-observability of free quarks is interpreted in such a way, that only colourless (white) particles can be seen. Experimental evidence has been gained for example by the measurement of the cross-section ratio

$$R := \frac{\sigma(e^- e^+ \rightarrow hadrons)}{\sigma(e^- e^+ \rightarrow \mu^- \mu^+)} = N_{\mathrm{C}} \sum_{2m_q < E_{\mathrm{CMS}}} Q_q^2, \tag{1.15}$$

which depends on the "colour factor" $\mathrm{N_C}$, i.e. the number of colours and Q_q the charge of the quarks being available at a certain centre of mass energy $\mathrm{E_{CMS}}$. The data taken with several experiments require $\mathrm{N_C} \equiv 3$.

While the strength of the electromagnetic interactions, described by the fine structure constant α, increases with higher momentum transfer Q^2, the coupling constant of the strong interaction α_s decreases. This behaviour, called "asymptotic freedom", states that at small distances quarks behave like free particles. It describes also why it is not possible to see free coloured particles (confinement).

After the success of local gauge theories in the field of electromagnetic and weak interactions (see section 1.3) theorists tried to construct a theory of strong interactions between quarks, which is based on local gauge transformations with colour as the interaction charge.

In 1973 Gross and Wilczek discovered that non-abelian gauge groups can describe theories with asymptotic freedom and managed to formulate the theory of quantum chromodynamics based on the local gauge group $\mathrm{SU(3)_C}$. Strong interactions stay invariant under the colour transformation

$$U_{\mathrm{C}}(x) = exp\left(i \frac{g_s}{2} \sum_{j=1}^{8} \lambda_j \beta_j(x) \right). \tag{1.16}$$

These are described by eight independent rotations β_j in the colour space, by the QCD coupling constant g_s and by the Gell-Mann-matrices λ_j. To guarantee the invariance of the equations of motion eight additional vector fields G_j^μ and a covariant derivative D^μ have to be introduced

$$D^\mu = \partial^\mu + i \frac{g_s}{2} \sum_{j=1}^{8} \lambda_j G_j^\mu. \tag{1.17}$$

The massless particles related to the vector fields are the eight different coloured gluons, which mediate the strong interaction. The first evidence for gluons was observed at the PETRA collider in 1979 in three jet events [7], where one jet originates from a radiated gluon. In contrast to photons, which are electrically neutral, gluons carry the interaction charge (colour). Thus additional terms appear in the transformed gluon fields performing a rotation in the colour space (last term of (1.19).

$$\psi(x) \rightarrow U_{\mathrm{C}}\, \psi(x) \tag{1.18}$$

$$G_j^\mu(x) \rightarrow G_j^\mu(x) - \partial^\mu \beta_j(x) - g_{\mathrm{S}}\, f_{jkl}\, \beta_k(x)\, G_l^\mu(x) \tag{1.19}$$

1.3 The GSW-Model of Electroweak Interactions

Symmetries, broken or not, like the broken flavour symmetry or the exact colour symmetry, do not only appear in QCD, but the way they are hidden in weak interactions makes them less obviously discernible. While the flavour symmetry is visible in the spectrum of particles and their approximate mass degeneracy, the observed universality of the Fermi coupling of weak-decay processes suggests the existence of a hidden symmetry in weak interactions. This is an outstanding fact since the weakly interacting particles have widely varying masses. The symmetry manifests itself not through the existence of degenerated multiplets, but through broken local symmetries.

In the 1960s Glashow, Salam and Weinberg (see [8–10]) were the first, who realised a unified theory of weak and electromagnetic interactions in the framework of a renormalizable field theory. It is based on the gauge group $\mathrm{SU(2)_L}\times\mathrm{U(1)_Y}$.

In order to describe the interaction, they assigned the quarks and leptons to representations of the gauge groups arranged in multiplets as shown in Table 1.2. As seen first in β-decays by Wu et al. [11] parity is violated maximally in weak interactions and weak charged currents couple only to left-handed particles, where the handedness is determined by the projection operators

$$\mathrm{P_L} = \frac{1}{2}(1-\gamma_5) \qquad \mathrm{P_R} = \frac{1}{2}(1+\gamma_5). \tag{1.20}$$

This experimental result is included in the Standard Model by the assignment of the left-handed fermions to $\mathrm{SU(2)_L}$ doublets, while the right-handed fermions transform as singlets under $\mathrm{SU(2)_L}$.

Except for the Higgs sector (see below) the Lagrangian is completely dictated by (the desired feature of) gauge invariance and renormalisability[4]. It can be separated into these parts:

$$\mathcal{L}_{\mathrm{GSW}} = \mathcal{L}_{\mathrm{fermion}} + \mathcal{L}_{\mathrm{gauge}} + \mathcal{L}_{\mathrm{Higgs}} + \mathcal{L}_{\mathrm{Yukawa}} \tag{1.21}$$

The first term describes the kinematic of the free fermion fields and their interaction with the gauge fields. It has the form

$$\mathcal{L}_{\mathrm{fermion}} = i\overline{\psi}\gamma^\mu D_\mu \psi \tag{1.22}$$

with ψ as the combined spinor of all fermionic fields and D_μ as the covariant derivative of the $\mathrm{SU(2)_L}\times\mathrm{U(1)_Y}$ gauge group[5]

$$\psi = \begin{pmatrix} \nu_{e_\mathrm{L}} \\ e_\mathrm{L} \\ \vdots \\ t_\mathrm{R} \end{pmatrix}, \qquad D_\mu = \partial_\mu + i\,g\,\frac{T_a}{2}W_\mu^a + i\,g'\,B_\mu Y. \tag{1.23}$$

[4] Renormalisability reflects the fact, that the predicted interaction probabilities stay finite by including higher order corrections and self-couplings of bosons. As proven by 't Hooft local gauge invariance is a condition for the renormalisability of gauge theories with massless and massive gauge bosons. It can be shown, that the Lagrangian can only contain terms with dimensions less than or equal to 4.

[5] Equation (1.23) is only a symbolic notation! W_μ^a acts only on the left-handed fermions (isospin doublets), while B_μ acts on both, right- and left-handed particles.

Fermions (Spin 1/2)							
	Generation			Quantum Number			
	1.	2.	3.	Q	T	T_3	Y
Quarks	$\begin{pmatrix} u \\ d' \end{pmatrix}_{\mathrm{L}}$	$\begin{pmatrix} c \\ s' \end{pmatrix}_{\mathrm{L}}$	$\begin{pmatrix} t \\ b' \end{pmatrix}_{\mathrm{L}}$	$2/3$ $-1/3$	$1/2$ $1/2$	$1/2$ $-1/2$	$1/3$ $1/3$
	u_{R} $d_{\mathrm{R}'}$	c_{R} $s_{\mathrm{R}'}$	t_{R} $b_{\mathrm{R}'}$	$2/3$ $-1/3$	0 0	0 0	$4/3$ $-2/3$
Leptons	$\begin{pmatrix} \nu_e \\ e \end{pmatrix}_{\mathrm{L}}$	$\begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix}_{\mathrm{L}}$	$\begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}_{\mathrm{L}}$	0 -1	$1/2$ $1/2$	$1/2$ $-1/2$	-1 -1
	e_{R}	μ_{R}	τ_{R}	-1	0	0	-2

Bosons (Spin 1)					
Interaction	Gauge Boson	Q	T	T_3	Y
Electromagnetic	γ	0	0	0	0
Weak	Z^0	0	1	0	0
	W	1	1	± 1	0
Strong	$g_{1\ldots 8}$	0	0	0	0

Table 1.2: *The particles of the Standard Model with their electroweak quantum numbers. Fermions are assigned to left-handed doublets and right-handed singlets. The primes on the left-handed down-type-quarks indicate, that these are not the physical mass eigenstates, but the electroweak eigenstates. They are related via the 3×3 Cabbibo-Kobayashi-Maskawa-matrix. Note that neutrino oscillations also require a mixing matrix for the neutrino sector. Q denotes the electromagnetic charge, Y the weak hypercharge and T_3 the eigenvalue of the third component of the weak isospin T, where $Q - T_3 = Y/2$.*

Since any arbitrary special unitary group SU(N) is built up by N^2 - 1 generators and the unitary groups U(N) by $\mathrm{N} \cdot (\mathrm{N} + 1)/2$ generators, the gauge group $\mathrm{SU(2)_L} \times \mathrm{U(1)_Y}$ contains 3 + 1 gauge fields. These are denoted by W^a_μ ($a = 1, 2, 3$) and B_μ. The variables g and g' represent the coupling constants of the unified electroweak theory[6] and the matrices T_a (Pauli matrices) and Y the generators of the corresponding groups $\mathrm{SU(2)_L}$ and $\mathrm{U(1)_Y}$.

[6] The unification is not perfect, since it contains not only *one* coupling constant.

The boson fields W, Z^0 and the massless photon (A_μ) corresponding to the observed mass eigenstates are the linear combinations

$$W_\mu^\pm = \frac{1}{\sqrt{2}}\left(W_\mu^1 \mp i W_\mu^2\right) \tag{1.24}$$

$$Z_\mu = -B_\mu \sin\theta_\mathrm{W} + W_\mu^3 \cos\theta_\mathrm{W} \tag{1.25}$$

$$A_\mu = B_\mu \cos\theta_\mathrm{W} + W_\mu^3 \sin\theta_\mathrm{W} \tag{1.26}$$

where the electroweak mixing angle (Weinberg angle) is given by the coupling constants

$$\cos\theta_\mathrm{W} = \frac{g}{\sqrt{g^2+g'^2}} \qquad \sin\theta_\mathrm{W} = \frac{g'}{\sqrt{g^2+g'^2}}. \tag{1.27}$$

By inserting the fields $W_\mu^\pm$, Z_μ and A_μ into (1.22) one receives a representation of the interaction of gauge bosons and fermions by the exchange of currents. By construction the charged W bosons couple only to left-handed particles and right-handed antiparticles.

With the introduction of the covariant derivative (1.23) and the addition of a kinematic term for the gauge bosons, the Lagrangian (1.21) contains terms, which are bilinear in the gauge fields and thus describe the interactions among them. The occurrence of such terms is not trivial, since they do not occur in case of the photon.

$$\mathcal{L}_\mathrm{gauge} = -\frac{1}{4} W_{\mu\rho} W^{\mu\rho} + B_{\mu\rho} B^{\mu\rho} \tag{1.28}$$

using the field tensors

$$W_{\mu\rho} = \left(\partial_\mu W_\rho^a - \partial_\rho W_\mu^a - g\epsilon_{abc} W_\mu^b W_\rho^b\right) T_a \tag{1.29}$$

$$B_{\mu\rho} = \partial_\mu B_\rho - \partial_\rho B_\mu \tag{1.30}$$

T_a are the $\mathrm{SU(2)_L}$ generators and ϵ_{abc} are the structure constants.

Up to now the electroweak theory is a well formulated gauge theory describing the discovered particle spectrum, especially the gauge bosons W and Z^0. But there is one problem: none of the particles have masses, neither the fermions nor the massive gauge bosons. The simple addition of mass terms like $m^2 Z_\mu Z^\mu$ to the Lagrangian spoils local gauge invariance. One solution is the spontaneous symmetry breaking via the Higgs-mechanism.

1.4 The Higgs-Mechanism

A formulation of the electroweak symmetry breaking is given as the Higgs-mechanism. It solves the problem of the unitarity of the WW scattering amplitude, gives mass to the particles and is in accordance with electroweak precision tests. The Higgs mechanism spontaneously breaks the $\mathrm{SU(2)_L} \times \mathrm{U(1)_Y}$ symmetry without destroying the gauge invariance of the Lagrangian. In the simplest nontrivial implementation the Higgs boson field ϕ transforms as an isospin doublet under the gauge group $\mathrm{SU(2)_L}$

$$\Phi(x) = \begin{pmatrix} \phi^+(x) \\ \phi^0(x) \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1(x) + i\phi_2(x) \\ \phi_3(x) + i\phi_4(x) \end{pmatrix}. \tag{1.31}$$

The Higgs field couples to the gauge bosons ($\mathcal{L}_{\text{Higgs}}$) as well as to the fermions ($\mathcal{L}_{\text{Yukawa}}$)

$$\mathcal{L}_{\text{Higgs}} = (D_\mu \Phi(x))^\dagger (D^\mu \Phi(x)) - V(\Phi) \tag{1.32}$$

$$\mathcal{L}_{\text{Yukawa}} = \overline{\psi}\, \Phi_i(x)\, C_i\, \psi + h.c. \qquad (i = 1, 2), \tag{1.33}$$

where the matrix $\mathbf{C_i}$ contains the masses of the fermions i.e. the strength of the coupling to the Higgs field. The potential V(Φ),

$$V(\Phi) = \mu^2 \Phi^\dagger \Phi + \lambda (\Phi^\dagger \Phi)^2 \tag{1.34}$$

is chosen to be symmetric $V(\Phi) = V(-\Phi)$, so that only even powers of Φ occur and higher orders are neglected. To have a reasonable theory the potential has to tend to infinity for the limit $\Phi \to \pm\infty$, thus $\lambda > 0$ and must have a lower bound. The potential has only a non-trivial minimum for $\mu^2 < 0$, which is given by

$$\Phi^\dagger \Phi = -\frac{\mu^2}{2\lambda} =: \frac{v^2}{2}. \tag{1.35}$$

Only in this case it is possible to break the $\text{SU(2)}_\text{L} \times \text{U(1)}_\text{Y}$ symmetry. A possible solution, which sets the vacuum expectation value of the charged Higgs field ϕ^+ to zero and guarantees the photon mass to be zero is

$$\Phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ \rho(x) \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}. \tag{1.36}$$

Then the neutral part of the Higgs field can be expressed in terms of the vacuum expectation value v and a scalar field $h(x)$. By the special choice of the vacuum expectation the SU(2)_L as well as the U(1)_Y is broken, but the U(1)_em symmetry remains untouched.

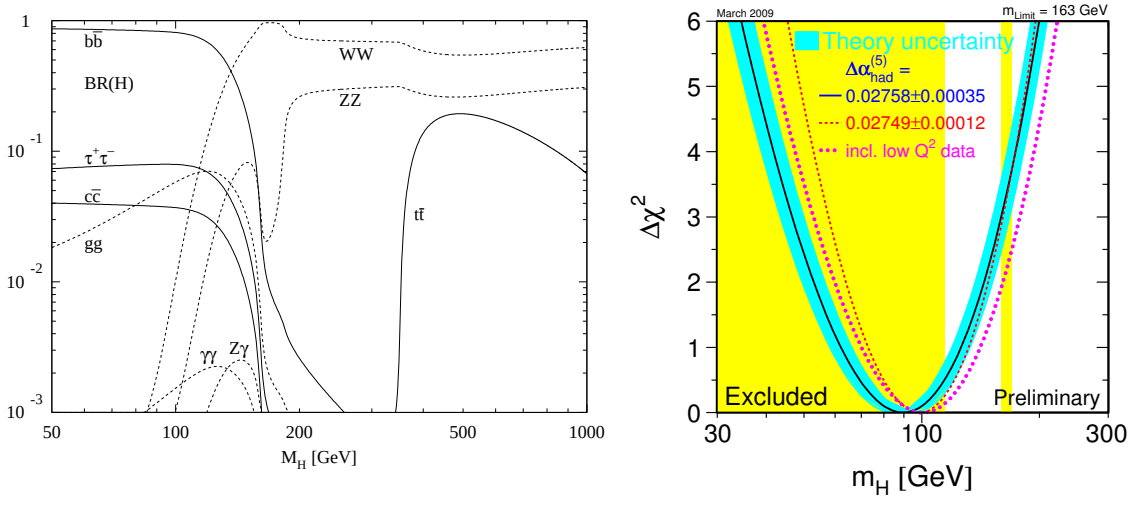

Figure 1.2: ***Left:*** *Branching fractions of the Higgs boson [12]. For mass ranges beyond the LEP limits up to a mass of approximately 150* GeV *the $\gamma\gamma$-decay is the most promising discovery channel. Beyond the decay into gauge bosons is said to be gold-plated.* ***Right:*** *The so called blue band plot [13] shows the regions (yellow) which have been experimentally excluded by direct searches at LEP and recently at the Tevatron. The parabola shows the $\Delta\chi^2$-distribution from a theory fit to electro-weak precision data, which are in favour of a light Higgs.*

The Goldstone boson arising naturally in a broken symmetry is absorbed as longitudinal degree of freedom of the massive gauge bosons W and Z^0.

Replacing $\Phi(x)$ in equation (1.32) by (1.36) mass terms like $\frac{1}{4}g^2v^2W_\mu^+W^{-\mu}$ arise and give mass to the weak gauge bosons

$$m_{Z^0} = \frac{1}{2}\sqrt{g^2+g'^2}\,v, \qquad m_W = \frac{1}{2}gv, \qquad m_\gamma = 0, \qquad m_{\mathrm{H}} = v\sqrt{2\lambda}. \tag{1.37}$$

The interaction of the fermions with the Higgs field leads to mass terms and to couplings to the Higgs field $h(x)$, which are proportional to the fermion mass ($\mathcal{L}_{\mathrm{Yukawa}}$).

Today's knowledge of the Higgs is impressive: all its couplings, its cross section and branching ratio are known (as a function of the mass) with a high precision. The missing piece of information is if it exists at all and if with which mass. Figure 1.2 shows the branching ratios of the Higgs as a function of its mass. The famous "blue band plot" reflects the regions where it is likely to be found. Higgs masses beyond 300 GeV are excluded by electroweak precision data [13].

Chapter 2

Beyond the Standard Model

Despite the experimental success of the Standard Model there are strong experimental and theoretical indications that the model does not describe nature in every detail. Instead the Standard Model is thought of as an effective theory valid up to energies probed at past and recent collider experiments. It is thus only an approximative model of a more generic and complete theory.

In the following experimental and theoretical evidences which point to physics beyond the Standard Model are discussed. Models which try to extent the Standard Model in order to solve some of its limitations are sketched. The focus of the introduced theories is on models which can be probed at the LHC and which are used within this analysis. Although this work deals with a model-independent search without any theoretical bias beyond the Standard Model, representative signatures of new physics can be used as benchmark scenarios to demonstrate its feasibility.

2.1 Pointers to Physics beyond the Standard Model

One of the main tasks of the LHC is the discovery or the exclusion of a Standard Model like Higgs boson. Even if nature has realized the Higgs bosons there are various reasons why one could expect to spot further signatures of new physics. Lacking the inclusion of quantum gravity our current model will at last finally fail around the Planck scale. Things get even more exciting without a Higgs boson. Then, other mechanisms (e.g. described by little higgs models) have to come in to explain the electroweak symmetry breaking and the unitarization of the WW cross section. However, is there a chance to detect first hints of these new theories at LHC energies?

This chapter tries to shed some light on this question. It discusses theoretical and experimental hints for new physics at the TeV-scale and describes some of the candidate models, which might enter at this scale and which are used as benchmark scenarios within this feasibility study.

2.1.1 Experimental Hints

In the last decade the most spectacular discoveries concerning particle physics have been made at non-collider experiments. New inspiring results have been obtained by neutrino experiments like K2K and in the field of astro-particle physics like WMAP.

Neutrino Oscillations and Neutrino Masses

Today's experiments only yield upper limits for neutrino masses, but the recently observed neutrino oscillations require neutrinos to have a non vanishing mass [14]. In the Standard Model a neutrino mass can not be implemented ad hoc. By choice of the multiplets there is no simple theoretical solution. Since Dirac neutrinos[1] are not foreseen in the Standard Model, they can be added to the Standard Model only as gauge singlets, which would naturally result in neutrino masses of the order of their charged counterparts [15]. Since there are no gauge singlet or triplet Higgs scalars, Majorana masses[2] cannot be generated either [15]. However, it is possible to add terms to the Lagrangian, which result in neutrino masses. But they predict a new mass scale beyond the SM [6] and thus point towards new physics.

Dark Energy and Dark Matter

The current cosmological model of the universe is based on the two unknown components of dark energy and dark matter. The evidence for dark i.e. non-luminous and non-absorbing matter has already emerged 50 years ago from the observations of the rotation curves of galaxies.

While the balance of the gravitational and the centrifugal force would yield a rotation speed $v \sim r^{-1/2}$ as a function of the distance r from the galactical centre, the rotation curves of many galaxies are flat as e.g. given in figure 2.1 (left). This leads to the conclusion of a halo of dark weakly interacting particles surrounding the visible matter. Detailed studies conclude that around 25% of our universe is made out of dark matter.

Although there are evidences for dark matter at both large- and galactic-scales no experiment to date has successfully detected dark matter particles. Many extensions of the Standard Model predict the existence of particles with the characteristics of dark matter. For example, cold dark matter, such as the weakly interacting massive particles predicted by supersymmetry, which is favoured by cosmological numerical simulations compared to hot dark matter based on neutrinos.

Recent cosmological observations are in favour of a flat, but accelerated universe [19]. Given the estimate of the complete energy within our universe, this is in contrast to the attractive power of gravitation. Proposed solutions include either a modification of general relativity and the gravitational laws at cosmological scales or a new repulsive gravitational energy form called dark energy. The currently most accepted Standard Model of Cosmology is based on dark energy with a constant energy density (cosmological constant). Dark energy is supposed to make up around 75% of the energy in our universe.

[1] Dirac particles are distinguishable from their anti-particles.
[2] Majorana particles are their own anti-particles.

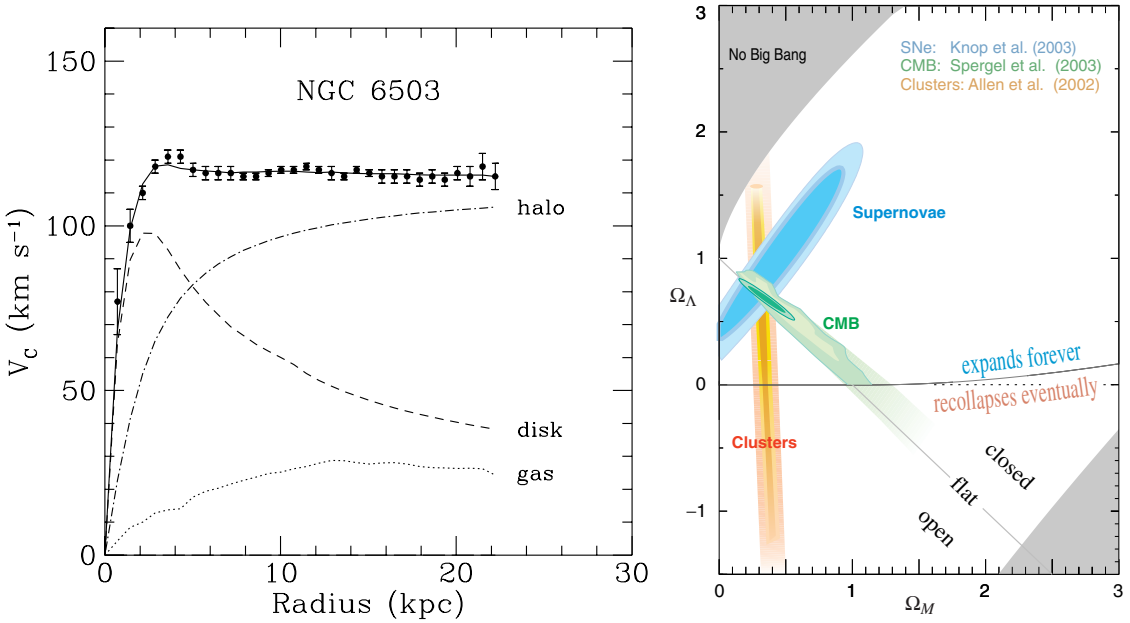

Figure 2.1: ***Left:*** *Rotation curve of an galactic object. The rotation speed as a function of the radius from the galactical centre implies the existence of a halo of dark matter surrounding the object [16].* ***Right:*** *Measurements of the current matter density Ω_{M} (sum of baryonic and dark matter density) versus the cosmological constant Ω_Λ usually identified as dark energy. The densities are normalized to the critical density of the universe and are in favour of a flat universe ($\Omega_{\mathrm{tot}} = 1$, see e.g. [17]) which consists to roughly 25% out of matter and to 75% out of dark energy [18].*

2.1.2 Theoretical Hints

Also from the theoretical point of view there are several arguments at hand why the Standard Model is not the ultimate answer. Just to name a few: it does not incorporate gravity, it contains many free parameters and is unstable against corrections from the Planck or any intermediate scale. The desired solution would be a grand unified theory or a theory of everything which unifies all forces and gives an inherent explanation for all its parameters and symmetries. Here some of the main theoretical limitations of the Standard Model should be briefly discussed.

The Hierarchy Problem

Probably the most serious theoretical issue of the Standard Model is its instability against the huge hierarchy of vastly different scales relevant in high energy physics. The well-known electroweak scale at a few hundred GeV is opposed to the Planck scale $\sim 10^{19}$ GeV where the influence of gravity is comparable to other forces. Consider the case where the electroweak symmetry breaking is mediated by a Standard Model Higgs boson. Quantum loop corrections to the mass of the Higgs boson are sensitive to a cut off parameter Λ

naturally thought of as the scale up to which the theory is valid. One loop diagrams yield a corrections to the Higgs mass which are quadratically divergent:

$$\Delta m_H^2 = -\frac{3\lambda_q^2}{8\pi^2}\Lambda^2 \quad \text{(for fermions)} \qquad \Delta m_H^2 = \frac{g^2}{16\pi^2}\Lambda^2 \quad \text{(for bosons).} \tag{2.1}$$

If one assumes that the Standard Model is valid up to the Planck scale ($\Lambda \sim 10^{19}$ GeV) the Higgs mass should be naturally in the order of the Planck scale. However, the LEP data and other theoretical constraints expect the Higgs to have a mass of 100 – 300 GeV. Extremely precise cancelations at all pertubation levels are required to fix the Higgs mass to something of the order of 100 GeV. This does not expose a mathematical contradiction in the theory, but is merely an aesthetical problem of "fine-tuning" or "naturalness". Theorists are in favour of a natural cancelation or physics at intermediate scales in order to avoid these highly tuned cancelations. If one accepts corrections to the Higgs mass which are not larger than ten times the Higgs mass itself, the scale at which one expects new physics to enter is at $\Lambda \sim 2$ TeV, which sounds very promising for the LHC.

The variety of ideas to solve this blemish of the Standard Model is overwhelming. They involve new symmetries (e.g. supersymmetry), new particles canceling these divergencies (e.g. Little Higgs), new physics at intermediate scales or even the effective lowering of the Planck scale within models with large extra dimensions.

Unification of Forces

A shortcoming of the Standard Model is the inclusion of gravity, whose strength should become comparable with that of other interactions at the Planck scale (10^{19} GeV). The problem that gravitation is still far outside the Standard Model, is given by the fact that its addition spoils the feature of renormalisability. Theories beyond the Standard Model, like string theory, try to address this unification of local gauge invariance and the principle of equivalence.

The Standard Model is based on three different gauge groups associated with arbitrary coupling constants. From the theoretical and aesthetic point of view one unified gauge group, which contains the SM as a subgroup, seems to be much more satisfying. In such a "grand unified theory" (GUT) the strong, electromagnetic and weak interactions can be understood as being just three different manifestations of a single fundamental interaction. Due to spontaneous symmetry breakings at different energy scales the observed low energy behaviour and thus the well-established Standard Model could be restored as an effective theory of the physics at the electroweak scale. In order to unify the different interactions, the coupling constants need to converge into a single value at a certain energy scale where the unification takes part. However, this fails within the Standard Model as shown in figure 2.2. Certain extensions of the Standard Model such as low mass supersymmetry entering at the tera-scale would modify the running of the coupling constants in such a way that they cross in exactly one point. This unification of couplings would mark the basis of a grand unified theory. The predicted GUT scale of typically 10^{16} GeV is approximately the same as the one that would give rise to neutrino masses consistent with the experimental observations.

The hope is that a unified theory can also include the description of quantum gravity relevant at these scales. In addition a satisfying GUT should not only give a description, but instead a natural explanation of all its parameters and symmetries.

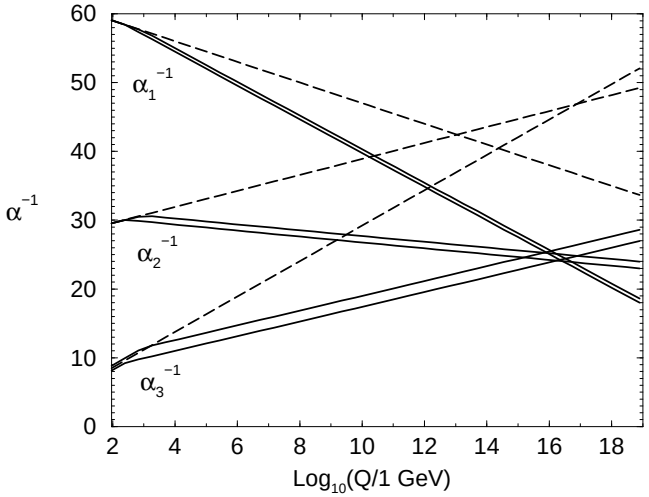

Figure 2.2: *Evolution of the inverse gauge couplings within the Standard Model (dashed) compared to the MSSM (solid). While the unification fails within the SM the additional MSSM particles ensure that the gauge couplings can unify at a scale of* $\sim 10^{16}$ GeV. *The sparticle masses are varied between 250* GeV *and 1* TeV, *and* $\alpha_3(m_Z)$ *between 0.113 and 0.123. The calculation includes two-loop effects. [20].*

Yukawa Couplings and other Arbitrary Parameters

The Standard Model contains various free parameters which are determined by measurements, but are lacking a fundamental explanation. There are at least 19 absolute arbitrary parameters in the SM, and more are needed to incorporate neutrino masses. In a fundamental physical model all these parameters should not appear as totally free.

The assignment of the left-handed fermions to doublets and the right-handed to singlets is only justified by the fact, that it fits to data. There is no explanation why charged weak currents are strictly left-handed as well why there are three fermion generations. Their mixing and the masses given through Yukawa couplings stay arbitrary in the SM. The hierarchical pattern of the quark masses m_t, $m_b \gg m_c$, $m_s \gg m_u$, m_d, but also for charged leptons $m_\tau \gg m_\mu \gg m_e$ (for neutrinos the mass hierarchy still has to be confirmed) might be hints for additional hidden symmetries.

Also, there is no explanation both for the origin of the three-family structure and the breaking of the generational symmetry (flavour symmetry) and for the fact that the particle masses would be significantly smaller than the energy scale up to which the theory remains valid. Another puzzle of nature is the quantization of the electric charge. One would expect, that a fundamental theory predicts the value of the elementary electric charge.

Thus to judge the meaning and importance of tests, which stress the Standard Model, it is necessary to work in a more general framework and be aware in which direction it can be modified. In the following models which try to supersede the limitations of the Standard Model are outlined.

2.2 Selected Theoretical Models Beyond the Standard Model

A variety of theoretical models addressing the discussed limitations of the Standard Model have been proposed in the last decades begging for their investigation at the LHC. In the following some appealing models which have been probed in this model-independent search as benchmark points are discussed. These include supersymmetry, models with large extra dimensions or theories predicting new heavy gauge bosons.

2.2.1 New Heavy Gauge Bosons

Within the Standard Model the origin of parity violation in weak interactions stays unexplained. A priori the multiplets are explicitly designed to break parity in the weak sector. As displayed in table 1.2 the left-handed particles are assigned to doublets, whereas the right-handed particles do not participate in charged weak interactions, since they are $\mathrm{SU}(2)_\mathrm{L}$ singlets. Thus, the introduction of parity violation within the Standard Model has nothing to do with the spontaneous symmetry breaking of the gauge groups or any other mechanism, but has just been included by hand.

Left-Right-Symmetric Models (LRSM) [21–23] address this problem and provide an attractive extension of the Standard Model (for a review see [6, 15]). The general feature of these models is the intrinsic exact parity symmetry of the Lagrangian and an additional SU(2) gauge group, resulting in an observable W' and Z'. To match the low-energy behaviour of maximum parity violation in weak interactions, the symmetry is spontaneously broken by a scalar Higgs field.

In addition LRSM incorporate full quark-lepton symmetry and turn the quantum number of the U(1) from the hypercharge Y to the value of baryon-minus-lepton number $B - L$. Finally, in choosing an appropriate Higgs sector the theory gives a natural explanation for the smallness of the neutrino masses, by relating it to the observed suppression of $V + A$ currents. Variants of the model can be derived from grand unified theories, superstring inspired models or other theories based on extended gauge groups, which contain the LSRM as a subgroup.

The simplest realization of a Left-Right-Symmetric Model is based on the gauge group

$$SU(2)_\mathrm{L} \times SU(2)_\mathrm{R} \times U(1)_{\widetilde{\mathrm{Y}}}. \tag{2.2}$$

The SM fermion doublets are mirrored by arranging the right-handed singlets of the Standard Model together to form another SU(2) doublet. In the lepton sector this can only be done by predicting a neutrino singlet ν_R for each generation, which is a massive Majorana particle

$$u_\mathrm{R}, d_\mathrm{R} \rightarrow \begin{pmatrix} u_\mathrm{R} \\ d_\mathrm{R} \end{pmatrix}; \qquad \nu_\mathrm{R}, l_\mathrm{R} \rightarrow \begin{pmatrix} \nu_\mathrm{R} \\ l_\mathrm{R} \end{pmatrix}. \tag{2.3}$$

The spontaneous symmetry breaking occurs in two steps with appropriate Higgs sectors i.e. the parity symmetry of the Lagrangian is broken by an Higgs bosons, whose vacuum expectation value is not parity conserving. This first stage gives mass to the W_R and Z_R, which are bosons in the right-handed sector. The properties of the W_R are different

compared to the Standard Model W and thus, match with the given definition of a W'. In addition right-handed neutrinos occur, which have to be very heavy.

The masses of the other boson fields, W_{L} and Z_{L}, result from the subsequent symmetry breaking. This step is in principle equivalent to the Higgs-mechanism in the Standard Model and the arising bosonic fields can therefore be identified with the Standard Model W and Z^0.

Beside the additional vector bosons and numerous Higgs scalars, an important feature of LRSM models is the generation of neutrino masses. Due to the existence of right-handed neutrinos, the neutrinos obtain Majorana masses through the symmetry breaking. Via a see-saw mechanism [24, 25] the Standard Model neutrinos obtain small masses, whereas the right-handed neutrinos N obtain masses in the order of the breaking mass scale u_{R}

$$m_{\mathrm{N}} \sim u_{\mathrm{R}} \qquad \text{and} \qquad m_{\nu_l} \sim m_l^2/m_N. \tag{2.4}$$

As stated before Left-Right-Symmetry can occur in models with larger gauge symmetry groups as intermediate state of a symmetry breaking pattern. Thus, the variety of such models is in principle arbitrary large [26]: they range from SO(10) over supersymmety to extra dimensions. Little Higgs models being in the actual focus of some theorists, are mentioned here as a theory predicting a W' at energies of the LHC.

Little Higgs

Little Higgs models provide a relative new formulation of the physics of electroweak symmetry breaking. The key features of those models are summarized here:

- The Higgs fields are Goldstone bosons, which are associated with some global symmetry breaking at a higher scale.
- The Higgs fields acquire mass and become pseudo Goldstone bosons via symmetry breaking at the electroweak scale.
- The Higgs fields remain light since they are protected by the global symmetry and free from a 1-loop quadratic sensitivity to the cutoff scale.

The interested reader is referred to dedicated papers (see for example [27]).

Here the motivation of new gauge bosons within these models should be mentioned briefly: a set of heavy gauge bosons are included in Little Higgs models having the same quantum numbers as the gauge bosons of the Standard Model. By the choice of the gauge coupling constants to the Higgs boson, quadratic divergencies induced by the SM gauge boson loops are canceled by quadratic divergencies of the new heavy gauge bosons. These new particles are expected to have masses of a few TeV pushing the hierarchy problem to a higher scale ($\mathcal{O}(10\ \mathrm{TeV})$). The entire reasonable parameter space of Little Higgs models can already be discovered or excluded with one year of LHC data [28].

The W' Reference Model

Given the large numbers of models which predict new heavy charged gauge bosons, it is a natural approach to use a simplified ansatz for their search. After a discovery of

signatures related to a new boson, detailed studies can be performed to distinguish between these models and to determine whether the boson belongs to a Little Higgs model, a Left-Right-Symmetric or a totally different one. The advantage of such an approach is the independence from other constraints. For example a search for a W' within a LRSM in the decay $W' \rightarrow \mu\nu_{\mathrm{R}}$ channel is confronted with the problem of right-handed massive neutrinos. In this case additional assumptions have to be made to get a discovery limit. Direct searches for new heavy gauge bosons are traditionally based on a simplified Reference Model first discussed by G. Altarelli [29].

The Reference Model is obtained by simply introducing ad hoc new heavy gauge bosons, two charged W' vector bosons as well as one neutral Z', as carbon copies of the Standard Model ones. The couplings are chosen to be the same as for the ordinary W and Z^0. The only parameters are the masses of the new vector bosons. While the coupling of the so constructed bosons with leptons is comparable to those obtained in extended gauge theories, the couplings to the massive Standard Model gauge bosons are enlarged [29]. For W' masses larger than 500 GeV this leads to a W' width larger than its mass. Since such a state is not interpreted as a particle any more, the couplings of W' and Z' to the Standard Model W and Z^0 are suppressed manually in the Reference Model. One should notice that such a suppression arises naturally in extended gauge theories when the new gauge bosons belong to a different gauge group than the heavy SM bosons. Additional (heavy) neutrinos are not taken into account within this model.

For W' masses below the top mass (~ 175 GeV) the kinematically allowed decay channels are identical for the SM W and the W'. W' masses larger than about 400 GeV allow the decay $W' \rightarrow tb$. Since the phase space is enlarged it results in an increase of the width by a factor of about 4/3 [3]. In the intermediate region the factor is between 1 and 4/3 since the decay into a tb-pair is in principle possible, but suppressed because the quark pair has to be produced offshell.

Direct searches for W' bosons at the DØ experiment currently yield a lower limit on the mass of 1 TeV [30]. The LHC will extend the discovery/exclusion potential up to around 5 TeV as shown in various feasibility studies [31–34].

2.2.2 Models with Extra Dimensions and Mini Black Holes

A completely different approach to solve the hierarchy problem has been pioneered by Arkani-Hamed, Dimopoulos and Dvali [35–37] (ADD) as well as Randall and Sundrum [38, 39] (RS) (for a nice review see [40]). Motivated by string theory further dimensions are added to the well-known four dimensions of space-time (brane). While all forces except gravity are restricted to the brane, the weakness of the gravitational force is explained by its dilution into extra dimensions. Effectively this lowers the Planck scale, possibly down to scales accessible at the LHC.

[3] Due to the small mixing between the quark generations the W can mainly decay to du, sc and $l\nu$. Taking the quark colour into account one obtains $(3 \cdot 2 + 3) = 9$ different decays. A heavy W' has the additional quark-antiquark decay into tb and thus $(3 \cdot 3 + 3) = 12$ possible decays. This results in a rise of the W' width by a factor $12/9 = 4/3$.

In general one can distinguish between three common models incorporating large extra dimensions. ADD-model include several space-like but flat extra dimensions, all with the same compactification radius, while the RS-model only adds a single, warped extra dimension. In both models the SM particles are confined to the common four space-time dimensions, while gravitons are allowed to propagate into the higher-dimensional space (bulk). A third category of models are universal extra dimensions [41] where all particles are allowed to expand through the bulk. Within this work the focus is on the ADD-type model.

The gravitational potential of a particle with a mass M in a space with $d+3$ dimensions is given as

$$\phi(r) \sim \frac{1}{M_{\mathrm{f}}^{d+2}} \frac{M}{r^{d+1}} \tag{2.5}$$

introducing a new fundamental mass scale M_{f}. In case of d dimensions which are compactified to a radius R, the observer at distances $r \gg R$ will not notice the extra dimensions and measures the common gravitational potenial

$$\phi(r) \sim \frac{1}{M_{\mathrm{f}}^{d+2}} \frac{1}{R^d} \frac{M}{r} \sim \frac{1}{M_{\mathrm{Pl}}^2} \frac{M}{r} \tag{2.6}$$

One can identify the new scale M_{f} as the Planck scale M_{Pl} reduced by the volume of the extra dimensions

$$M_{\mathrm{Pl}}^2 = M_{\mathrm{f}}^{d+2} R^d. \tag{2.7}$$

This formula shows how extra dimensions solve the hierarchy problem. Due to the additional extra dimensions the relevant scale is the reduced scale M_{f} and not the Planck scale M_{Pl}. It explains the weakness of gravity due to the detection in only the few limited number of dimensions accessible to us.

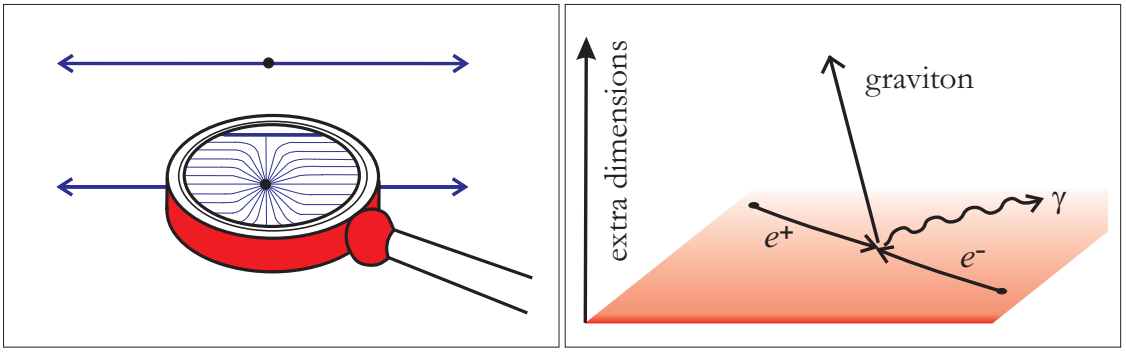

Figure 2.3: ***Left:*** *Schematic illustration of extra dimensions and of the effective weakening of gravity.* ***Right:*** *Production of a graviton and a photon from an electron-positron collision. While the high energy photon can be measured the graviton vanishes undetected in the bulk [40].*

For masses accessible at the LHC typical radii R of the extra dimensions in ADD models are of the order of 0.1 mm down to 1 pm for $2 - 7$ extra dimensions d. Direct tests of Cavendish like experiments confirm Newtonian gravity down to scales of ~ 50 μm, excluding the case $d = 1$ and disfavour $d = 2$ (see [42]). Direct searches at the Tevatron exclude scales M_{f} up to ~ 1 TeV (see e.g. [43]).

Particles which enter the bulk have quantized momenta due to the limited size of the extra dimensions. These infinite number of possible discrete and massless Kaluza-Klein excitations, enter our world as massive particles, leading to noticeable effects at energies of the new scale M_{f} and larger. In the case where the new scale is accessible at the LHC spectacular signatures of mono-jets or other particles recoiling against a graviton vanishing undetected within the bulk. Even more impressive signatures might stem from mini black holes leading to events with many high energetic, spherically distributed particles.

While astronomical black holes require an aggregation of mass of the order of the Planck scale M_{Pl}, mini black holes with dimensions smaller than the radius R of the extra-dimension only require energies of the new scale M_{f}, which might be as low as the electroweak scale. This can be seen when looking at the Schwarzschild radius of the black hole with a mass M, which marks the event horizont of the object

$$\begin{aligned} R_{\mathrm{H}} &= \frac{2M}{M_{\mathrm{Pl}}^2} \quad \text{for astrophysical black holes} \\ R_{\mathrm{H}} &= \frac{2}{d+1}\frac{1}{M_{\mathrm{f}}^{d+1}}\frac{M}{M_{\mathrm{f}}} \quad \text{for mini black holes.} \end{aligned} \tag{2.8}$$

Thus, particles which get closer than the Schwarzschild radius $R_{\mathrm{H}}(\sim 10^{-4}$ fm for $M_f = 1$ TeV) will collapse and produce a black hole. Due to the lack of fundamental understanding of quantum gravity and of the black hole and its properties, it is treated as a metastable state, which is produced and decayed through a semi-classical formalism. The partonic cross section is estimated by the classical geometric cross-section

$$\hat{\sigma} = \pi R_{\mathrm{H}}^2. \tag{2.9}$$

The total cross section is naively calculated by the folding with the parton distribution functions (PDF) summing over all possible initial state partons[4]. For a centre of mass energy of 10 TeV the black hole cross sections using the BlackMax [44] generator are given at the mass $M_{\mathrm{f}} = 1$ TeV in table 2.1. The cross section decreases with the mass of the black hole due to the missing suitable initial state partons, but also with growing number of extra dimensions.

# extra dimensions d	$M_{\mathrm{BH}} > 3$ TeV	$M_{\mathrm{BH}} > 4$ TeV	$M_{\mathrm{BH}} > 5$ TeV
2	116 pb	17.4 pb	2.06 pb
4	62.0 pb	9.17 pb	1.07 pb
6	47.9 pb	6.94 pb	0.802 pb

Table 2.1: *Production cross sections for $M_{\mathrm{f}} = 1$ TeV at a centre of mass energy of 10* TeV *obtained by the BlackMax generator [44] for different black hole mass thresholds.*

The decay of black holes can be separated in three different steps: within the balding phase the black hole radiates the multipole moments from the initial partons through gravitational radiation into a state which can be described by the three parameters mass,

[4] Of course the validity of this approach is highly questionable e.g. the PDFs might change dramatically in the regime of quantum gravity.

angular momentum and electrical charge. In the evaporation phase the black hole first looses its angular momentum through so-called Hawking radiation and later on emits thermally distributed quanta. Due to the high temperature $T = \frac{1+d}{4\pi}\frac{1}{R_{\mathrm{H}}}$ of the black hole typically of the order of several 100 GeV many Standard Model particles and gravitons are emitted with these energies either to our brane or in the bulk. This leads to events with high particle multiplicities, but also a significant amount of missing transverse energy. Finally, when the black hole reaches the new scale M_{f} it is assumed that it either decays further into some last SM particles or leaves a stable relic whose properties are quite speculative.

Depending on the mass and the number of extra dimensions the black holes leave a spectacular signature of 5 – 50 high energy particles spherically distributed within the detector. But the consequence of black holes is even more dramatic. Quarks with energies above the production threshold of black holes would end up as black holes, leading to a sharp cut of in the jet energy distribution. This would mark "the end of short distance physics" (Giddings, Thomas [45]) as no further information could be extracted from the structure of matter at smaller scales.

2.2.3 Supersymmetry

Supersymmetry [20] tries to stabilize the hierarchy between the weak and the Planck scale by introducing a symmetry between fermions and bosons. The theory predicts the existence of partner particles with the same properties as the SM particles, but a spin-difference of half a unit. The solution to the hierarchy problem can best be seen when looking at the one-loop corrections to the scalar higgs mass (see also section 2.1.2)

$$\delta m_H^2 \sim (\Lambda^2 + m_B^2) - (\Lambda^2 + m_F^2) = m_B^2 - m_F^2. \tag{2.10}$$

Due to the relative minus sign between fermion and boson corrections the quadratic divergencies are removed if the masses of the particles and sparticles are similar. It can be shown that this cancelation happens at all orders of perturbation theory and therefore provides a strong argument for *low mass* supersymmetry with SUSY particles in the TeV-range. Another appealing argument for supersymmetry at the TeV-scale is the possibility to unify the gauge couplings at a scale of 10^{16} GeV pointing towards a grand unified theory of all forces. Also supersymmetry potentially provides a candidate for cold dark matter as e.g. the stable, lightest supersymmetric particle within the minimal supersymmetric extension of the SM.

As a complete review of supersymmetry is impossible within the scope of this work, only the brief concepts will be outlined within the framework of the minimal supersymmetric extension of the Standard Model (MSSM) and discussed within two commonly used phenomenological models used as benchmark channels within this analysis. The interested reader is referred to e.g. [20].

The MSSM is the supersymmetric extension of the Standard Model with the smallest possible particle content as given in table 2.2. The bosonic fields of the gluons, W, and B-fields get gluinos ($\tilde{g}$), winos ($\tilde{W}$), and binos ($\tilde{B}$) as fermionic partners. The scalar partner to the quarks and leptons are called squarks and sleptons. In order to give mass to up- and

Standard Model Particles/Fields		Supersymmetric Partners			
		Interaction Eigenstates		Mass Eigenstates	
Symbol	Name	Symbol	Name	Symbol	Name
$q = u, d, c, s, t, b$	quark	$\tilde{q}_{\mathrm{L}}$, $\tilde{q}_{\mathrm{R}}$	squark	$\tilde{q}_1$, $\tilde{q}_2$	squark
$l = e, \mu, \tau$	lepton	$\tilde{l}_{\mathrm{L}}$, $\tilde{l}_{\mathrm{R}}$	slepton	$\tilde{l}_1$, $\tilde{l}_2$	slepton
$\nu = \nu_{\mathrm{e}}$, ν_μ, ν_τ	neutrino	$\tilde{\nu}$	sneutrino	$\tilde{\nu}$	sneutrino
g	gluon	$\tilde{g}$	gluino	$\tilde{g}$	gluino
$W^\pm$	W-boson	$\tilde{W_{1,2}}$	wino	$\tilde{\chi}^\pm_{1,2}$	chargino
H^-	higgs boson	$\tilde{H}^-_1$	higgsino		
H^+	higgs boson	$\tilde{H}^+_1$	higgsino		
B	B-field	$\tilde{B}$	bino	$\tilde{\chi}^\pm_{1,2}$	neutralino
W_3	W_3-field	$\tilde{W}_3$	wino		
H_1	higgs boson	$\tilde{H}^0_1$	higgsino		
H_2	higgs boson	$\tilde{H}^0_2$	higgsino		
H_3	higgs boson				

Table 2.2: *Particle content of the minimal supersymmetric extension of the Standard Model.*

down-type quarks two Higgs doublets with in total five mass eigenstates are introduced together with their associated spin-1/2 higgsinos.

An additional constraint utilized within the MSSM is the conservation of R-parity. This new multiplicative quantum number R-parity is defined by

$$R := (-1)^{3B+L+2S}. \tag{2.11}$$

It is 1 for all Standard Model particles and -1 for the SUSY partners. As a direct consequence of R-parity conservation, SUSY particles can only be produced in pairs. They always decay into an odd number of sparticles plus further SM particles. The sparticle decay chain stops with the lightest supersymmetric particles which is stable. It is a natural cold dark matter candidate and would typically leave a characteristic large missing transverse energy signature within collider experiments. The nature of SUSY and the LSP is highly determined by the mechanism of supersymmetry breaking.

Breaking of Supersymmetry

None of the supersymmetric partners of the Standard Model have been observed up to now, which requires that the particle masses in the SUSY sector differ from the SM ones. Thus, a realistic model, given the existence of supersymmetry as an exact symmetry, demands a mechanism of symmetry breaking. In order to do so it is required to extend the MSSM including new particles and interactions at very high energy scales. Up to now no conclusive model has been formulated. Therefore pragmatically the symmetry breaking

is performed in a so-called "hidden sector" communicated via messenger particles in an indirect or radiative manner to the observable sector of the MSSM.

On can distinguish between two popular phenomenological models of how the breaking might be mediated. In gravity or Planck scale mediated supersymmetry breaking (SUGRA) the gravitational interaction is responsible for the mass difference between the SM particles and their supersymmetric partners. The alternative model known as gauge mediated SUSY breaking (GMSB) is based on a breaking mechanism communicated by the well-known electroweak and strong interactions. In the following these two models are briefly discussed in their minimal phenomenological implementation.

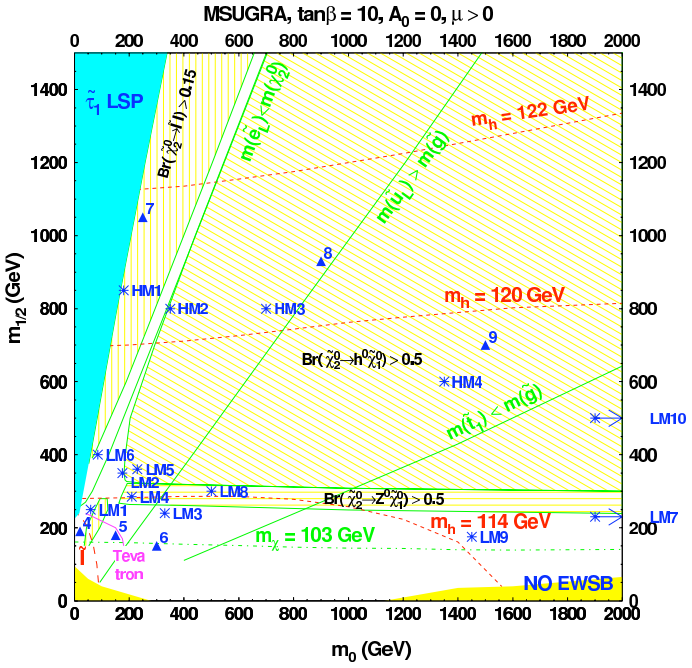

Figure 2.4: *The CMS mSUGRA benchmark points within the $m_{1/2}$ versus m_0 parameter space (for details see e.g. [34]). The low mass points (LM) correspond to regions which can be explored in the early data taking phase, while the high mass points (HM) are close to the ultimate LHC reach.*

Minimal Gravity-Mediated SUSY Breaking Model

This simplified phenomenological model, also known as Planck scale mediated SUSY breaking model or constrained MSSM, is based on a number of theoretical assumptions in order to reduce the more than hundred parameters of the MSSM to a reasonable number of five parameters. The reduction is obtained by assuming a set of boundary conditions at the grand unification scale (see e.g. [46]).

- The gaugino masses are assumed to unify to a common mass $m_{1/2}$ at the GUT scale.
- The same assumption is made for the sfermion and Higgs boson masses which are unified to m_0.
- All trilinear couplings are unified into one common trilinear coupling A_0.

- Further parameters defined at the electroweak scale are the ratio of the Higgs field vacuum expectation values $\tan\beta$ and the sign of the Higgs mixing parameter $\mathrm{sgn}(\mu)$.

Minimal Gauge-Mediated SUSY Breaking Model

Within these models the soft breaking is not mediated at the Planck scale, but at much lower masses. The breaking that occurs in the hidden sector is transmitted to the MSSM via a messenger sector by Standard Model gauge interactions (for a review see e.g. [47]). These scenarios predict the gravitino as lightest supersymmetric particle with very low mass usually below ~ 1 GeV. A minimal phenomenological model similar to the mSUGRA case can be constructed that is fully determined by six parameters.

- The effective SUSY breaking parameter Λ sets the overall mass scale of all MSSM particles, which scale approximately linearly with Λ.

- The masses of the sleptons, squarks and gauginos are generated radiatively from the gauge interactions with the N_5 generations of massive messengers (the index reflects the fact that the messenger fields form a SU(5) representation). The masses of the gauginos scale proportional to N_5, while the scalar masses depend on $\sqrt{N_5}$. For $N_5 = 1$ the next to lightest SUSY particle is the lightest neutralino $\tilde{\chi}_0^1$ typically decaying into a photon and a gravitino. Larger N_5 values determine a right-handed slepton as NSLP, which decays into a lepton and a gravitino.

- The mass scale of the messenger sector $M_m \ll M_{\mathrm{Pl}}$. The mass scale is required to be larger than Λ in order to avoid color and charge breaking in the messenger sector.

- The ratio of the vacuum expectation values $\tan\beta$ as in the mSUGRA model. Constraints are $1.5 < \tan\beta < 60$. For small $\tan\beta$ the lightest CP-even higgs approaches the LEP limits, while large $\tan\beta$ yield in a $\tilde{\tau}$ significantly lighter than all other sleptons.

- $\mathrm{sgn}(\mu)$: The sign of Higgs and Higgsino supersymmetric mass parameter μ appears in the chargino and neutralino mass matrices. For a Higgsino-like neutralino $\tilde{\chi}_0^1$ NLSP with low to moderate values of $\tan\beta$, $\mathrm{sgn}(\mu)$ is crucial in determining the relative strength of the $\tilde{\chi}_0^1$ coupling to Higgs and Z bosons through the Goldstino.

- C_{G}: The ratio of the messenger sector SUSY breaking order parameter to the intrinsic SUSY breaking order parameter controls the coupling to the Goldstino. The NLSP decay length scales like C_{G}^2.

Signatures, Benchmark Points and Limits

Within R-parity conserving supersymmetric models, the SUSY particles are always produced in pairs. At hadron colliders these are typically squark-, gluino- or squark-gluino-pairs, if their mass is within the reach of the centre of mass energy. Their decay leads to cascades of further particles, but also short decay chains might be possible. A prominent criterion which distinguishes supersymmetric events from Standard Model events is a large

Point	Λ	M_m	$\tan\beta$	N_5	$\mathrm{sgn}(\mu)$	C_G	$M(\tilde{\chi}_0^1)$	σ_LO[fb]
GM1b	80 TeV	2Λ	15	1	+	1	110 GeV	2970
GM1c	100 TeV	2Λ	15	1	+	1	139 GeV	843
GM1d	120 TeV	2Λ	15	1	+	1	168 GeV	299
GM1e	140 TeV	2Λ	15	1	+	1	197 GeV	12.4
GM1f	160 TeV	2Λ	15	1	+	1	226 GeV	5.82
GM1g	180 TeV	2Λ	15	1	+	1	255 GeV	3.11

Table 2.3: *CMS benchmark points for the searches for GMSB in the di-photon plus missing transverse energy channel. The parameters are chosen to have a short decay length of the next-to-lightest SUSY particle, which emits a photon and a gravitino.*

amount of missing transverse energy, usually caused by the lightest, stable, supersymmetric particles emitted at the end of the decay chain. This particles, which provides also a suitable dark matter candidate leaves the interaction undetected. Typically the two LSPs are accompanied by additional high energy jets and possibly also leptons/photons which might provide a cleaner signature. In general there is not *the* supersymmetry signature, but instead there are many possible regions which might result in completely different scenarios and therefore require different analysis strategies. The spectrum reaches from fully-hadronic searches, over single or di-leptonic/photon + jet + $\not\!\!E_\mathrm{T}$ analyses, up to approaches which do not rely on calorimetric $\not\!\!E_\mathrm{T}$ measurements at all.

As a complete scan of the whole sypersymmetric parameter space is not possible simplified models like mSUGRA or GMSB are used to reduce the number of free parameters. Within this reduced parameter space typically a few points with different characteristics are chosen as benchmark points. As an example the CMS mSUGRA benchmark points are shown in figure 2.4. The low mass points (LM), usually beyond the scope of the Tevatron reach, mark regions which might be accessible within the first years of data taking. Their cross sections are typically of the order of 1 – 100 pb. The high mass points (HM) have a much lower cross section and are close to the expected ultimate reach of the LHC.

Within gauge-mediated scenarios the lightest supersymmetric particle is the gravitino. The final state is characterized mainly by the nature of the next-to-lightest SUSY particle, which decays into the LSP and a SM particle. The NLSP might be a slepton, which then decays into a lepton and the LSP. In the other case where the NLSP is the lightest neutralino, the final state will consist of photons or possibly Z-bosons plus gravitino. Depending on the parameter C_G, the decay of the NLSP might happen instantaneously, after some centimetres (displaced vertices) or even far outside the detector (two heavy charged tracks or missing energy). Within this work the scenario where the neutralino is the NLSP with a very short life-time is considered as benchmark point. Table 2.3 gives a list of points considered within CMS together with their parameters and leading order cross sections. The NLSP has a mass of more than 100 GeV and decays to almost 100% into a gravitino and a photon. The decay into a Z-boson is suppressed to the per mille level due to the bino-like nature of the neutralino.

Currently the most stringent limit on GMSB models has been set by the DØ collaboration which report a lower mass limit of 125 GeV on the lightest neutralino and 229 GeV on the lightest chargino at 95% confidence level [48]. For the mSUGRA type of models the direct searches at the Tevatron state that the gluinos and squarks have a mass above approximately 310 GeV and 380 GeV, respectively [49].

Chapter 3

The CMS Detector at the Large Hadron Collider

Today's world largest particle physics laboratory, CERN, situated on the border between France and Switzerland, was founded on september 29, 1954. Since its foundation CERN made the way to breakthroughs in the understanding of fundamental particles and their interactions: the discovery of neutral currents in 1973, the discovery of the W and Z bosons in 1983, the high precision measurements of weak interactions at the Large Electron Positron Collider (LEP) experiments and lately the exploration of a new state of matter (possibly the quark-gluon-plasma) are only some of the highlights. Now the Large Hadron Collider (LHC) will join in to continue the success story. As a machine colliding protons with protons it provides a broad spectrum of centre of mass energy parton-parton interactions up to the tera-electronvolt regime. Thus, the LHC is well-suited to reveal the mechanism relevant for the electroweak symmetry breaking and to serve as a discovery machine for physics beyond the Standard Model.

Figure 3.1: *Aeroview of CERN's Large Hadron Collider with its four experiments ALICE, ATLAS, CMS and LHCb.*

3.1 The Large Hardon Collider

The Large Hadron Collider (LHC) [50] is a two-ring-superconducting-hadron accelerator installed in the existing Large Electron Positron Collider (LEP) tunnel at CERN. It is designed to collide protons (heavy ions) with a centre of mass energy of 14 TeV (5.5 TeV) up to a luminosity of 10^{34} cm^{-2}s^{-1}. This represents the next major step in the high-energy frontier beyond the Tevatron (proton-antiproton collider; centre of mass energy: 2 TeV) and the dismantled LEP machine (electron-positron collider; centre of mass energy: up to 208 GeV).

To achieve this luminosity and minimize the impact of simultaneous inelastic collisions in the detectors, collisions take place every 25 ns apart. At design luminosity this results on average in about 25 inelastic interactions per crossing. In Figure 3.2 the cross sections and the event rates at the LHC low luminosity ($\mathcal{L} = 10^{33}$ cm^{-2} s^{-1}) for various processes as a function of the centre of mass energy $\sqrt{s}$ are given. A remarkable aspect of the LHC physics is the wide cross section range of processes under investigation: While the total cross section is dominated by multi-jet production such as $qq \to qq$, $qq \to gg$ or $qg \to qg$, events from Higgs production and physics beyond the Standard Model are investigated with expected cross sections smaller by more than a factor of 10^{10}. The huge multi-jet background obfuscates the detection of a signal in final states containing only jets and thus in generally processes with leptons or photons are the preferred discovery channels. Therefore the identification and measurement of leptons especially in the high p_{T}-range is a crucial task for the LHC experiments.

3.1.1 Physics at Proton-Proton Colliders

The energy loss per revolution due to synchrotron radiation in a circular collider (radius R) is proportional to $E^4/(m^4 R)$ for a charged particle with a mass m and an energy E. This determines the LEP collider to be the last electron-positron synchrotron of these dimensions.

The use of protons with a 2000 times higher mass avoids the problem of huge radiative energy loss with the drawback of not colliding elementary particles. Instead of point-like particles the constituents of protons, namely quarks and gluons, interact with only a fraction of the protons' energy

$$\sqrt{s'} = \sqrt{x_a x_b s}. \tag{3.1}$$

x_a and x_b refer to the energy-fractions carried by the interacting partons, whereas $\sqrt{s'}$ is the centre of mass energy of the colliding partons and $\sqrt{s}$ the centre of mass energy of the protons. Thus, the centre of mass energy has to be larger compared to an electron-positron machine.

For the discovery of new particles it is not sufficient to reach a high collision energy. Also the production rate has to be large enough to produce the events of interest with a significant rate. The average number of events N_{events} per time for a special process with a cross section σ_{events} at a collider luminosity $\mathcal{L}$ is given by

$$\frac{dN_{\mathrm{events}}}{dt} = \mathcal{L}\,\sigma_{\mathrm{events}}. \tag{3.2}$$

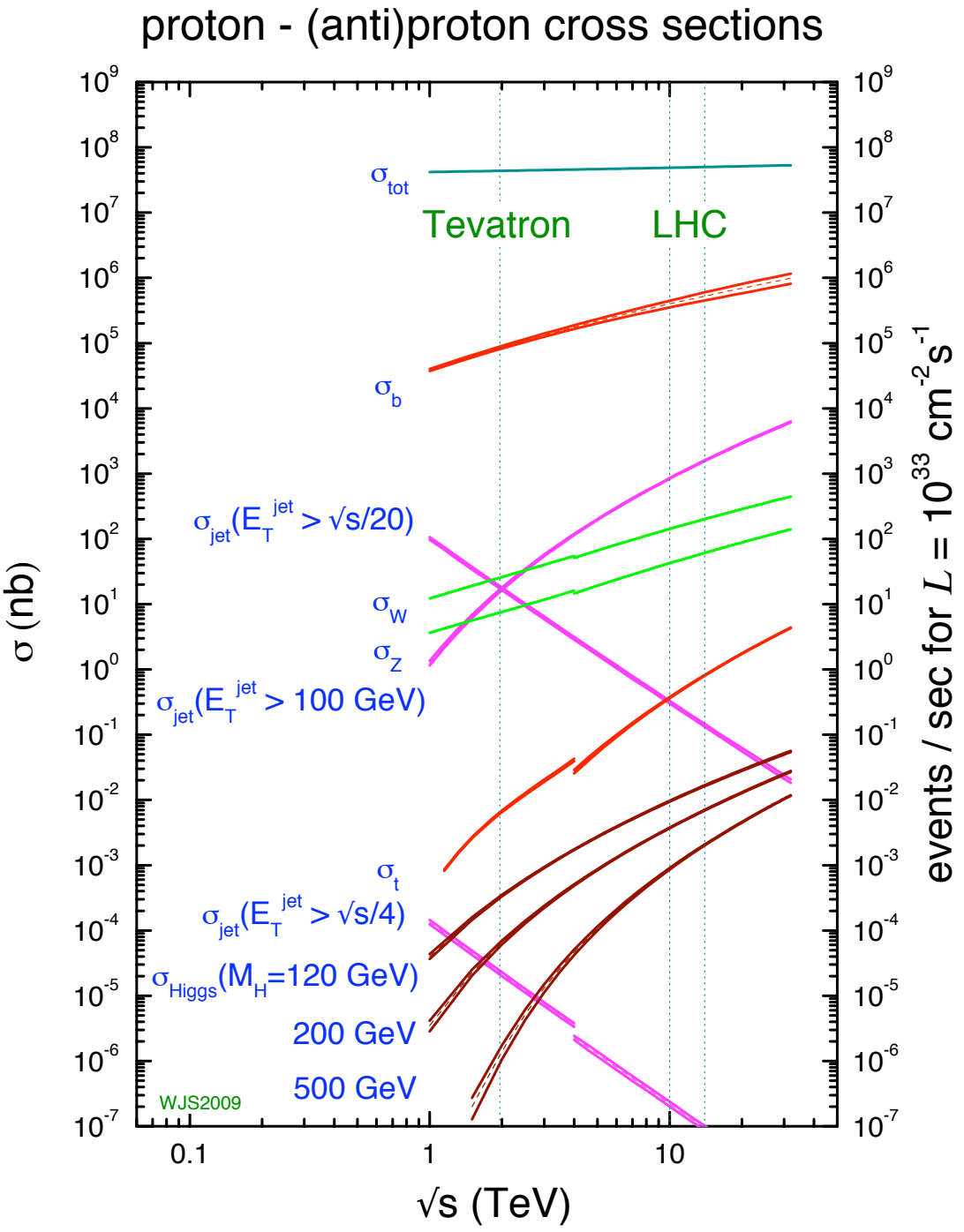

Figure 3.2: *Cross sections and event rates for different processes as function of the centre of mass energy at the Tevatron and the LHC proton-(anti)proton colliders [51].*

Assuming a Gaussian beam distribution with widths σ_x and σ_y in the x- and y-directions, respectively, the luminosity is approximately given by

$$\mathcal{L} = \frac{n_b N_b^2 f_{\mathrm{BX}}}{4\pi\sigma_x\sigma_y}. \tag{3.3}$$

N_b yields the number of particles per bunch, n_b the number of bunches per beam and f_{BX} the revolution frequency. All these parameters have to be tuned in order to achieve the highest possible luminosity and thus the best capability for new discoveries.

The cross section of a special partonic process depends on the cross section $\hat{\sigma}$ of the partons inside the proton (partonic cross section), graphically modelled by Feynman graphs. Since only two partons interact directly within a pp-collision the cross section also depends on the parton distribution functions (PDF) inside the proton,

$$\sigma(pp \to X) = \sum_{i,j} \mathrm{PDF}_{i,p}(x_1, f_1, Q) \otimes \mathrm{PDF}_{j,p}(x_2, f_2, Q) \otimes \hat{\sigma}_{ij\to X}(Q') \tag{3.4}$$

The value $\mathrm{PDF}_i(f_i, x_i, Q)$ equals the probability to find a parton with flavour f_i inside the proton carrying the momentum fraction x_i at the energy scale Q (factorization scale). The partonic cross section $\hat{\sigma}$ depends on a scale Q' referred to as renormalization scale. Due to the limited knowledge of the perturbation expansion (LO, NLO, ...) possible divergent terms might arise. Since these divergencies are unphysical, they need to be canceled by a suitable renormalization of the physical constants (e.g. couplings) at a certain (unphysical) scale, the renormalization scale Q'.

In hard scatterings the interaction energy and thus the rest frame is not known, because the proton remnants, which carry a sizable fraction of the protons' energy, escape undetected at small angles mainly through the beam pipe. Thus, only energy and momentum conservation in the transverse plane can be used to reveal the presence of non-interacting particles such as neutrinos.

Since the LHC is aiming for rare events the luminosity and thus the number of particles per bunch are chosen as large as possible. This has the drawback of having several interactions in one beam crossing. For the high luminosity phase ($\mathcal{L} = 10^{34}\ \mathrm{cm}^{-2}\,\mathrm{s}^{-1}$) of the LHC there are on average 25 simultaneous interactions, mainly multi-jet events (minimum bias). For the detectors this results in an extreme challenge identifying interesting physics processes out of the enormous amount of collisions.

The proton with its quark-gluon substructure enlarges the challenge. Since most of the events are created by two interacting partons colour charged fractions of the two protons leave the interaction point and produce additional jets. Since these particles carry small transverse momenta they vanish mainly through the beam pipe (beam remnants).

3.1.2 The LHC Design

With an expected centre of mass energy of 14 TeV and a design luminosity of $\mathcal{L} = 10^{34}\ \mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ a number of 2808 bunches of $1.15 \cdot 10^{11}$ protons each will be accelerated in the 27 km long former LEP tunnel 45 - 170 m below the surface. Bunches of protons will col-

lide every 25 ns at four interaction points where the experiments ALICE[1], ATLAS[2], CMS[3] (plus TOTEM[4]) and LHCb[5] are located. The two multi-purpose detectors, ATLAS and CMS, aim at rare events at the highest luminosities ($\mathcal{L} = 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$), whereas the low luminosity experiments LHCb ($\mathcal{L} = 10^{32}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$) and TOTEM ($\mathcal{L} = 2 \cdot 10^{29}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$) are investigating B-physics and protons from elastic scattering at small angles. Due to the general layout of the accelerator it is also possible to operate the LHC with heavy ion beams. In addition to the ATLAS and CMS experiments the LHC has one dedicated heavy ion experiment ALICE aiming at a peak luminosity of $\mathcal{L} = 10^{27}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ for $Pb-Pb$ collisions.

Parameter	Value	Unit
Momentum at Collision	7	TeV
Dipole Field at 7 TeV	8.33	T
Quadrupole Gradient	220	T/m
Circumference	26659	m
Design Luminosity	10^{34}	$\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$
Number of Bunches	2808	-
Particles per Bunch	$1.1 \cdot 10^{11}$	-
DC Beam Current	0.56	A
Stored Energy per Beam	362	MJ
Ultimate Dipole Field	9	T
Injection Dipole Field	0.4	T
Ramp Time	20	min
Distance between Beams	194	mm

Table 3.1: *Excerpt of the LHC design parameters [52].*

The LHC is designed as a superconducting collider accelerating two beams of equally charged particles with separate magnet dipole fields and vacuum chambers in the main arcs. The beams share common sections only at the four interaction points and at the insertion region. To allow an operating magnetic field of 8.4 T the 1232 dipole magnets are cooled with superfluid helium to a temperature of 1.9 K. A highly sophisticated system of magnets is used to focus the beam and thus to guarantee a continuous operation. The accelerator is divided into 8 parts from which only one octant serves for the beam acceleration via the radio frequency system. Further insertions apart from the four experiments are used for the beam cleaning system (twice) and the beam extraction system each in a separate octant. The injection of the two beams occurs in the octants shared with the ALICE and LHCb experiment.

The high centre of mass energy of 14 TeV can only be achieved by accelerating the bunches of particles stepwise using several already existing CERN pre-accelerator facilities

[1] **A** **L**arge **I**on **C**ollider **E**xperiment
[2] **A** **T**oroidal **L**HC **A**pparatus
[3] **C**ompact **M**uon **S**olenoid
[4] **TOT**al and **E**lastic **M**easurement
[5] The **L**arge **H**adron **C**ollider **b**eauty experiment

(see Figure 3.3). The upgraded Linac 2 will deliver protons of 50 MeV energy with an intensity of 180 mA and pulses of about 20 μs to the PS[6]. The modified PS with its two new radiofrequency systems, will feed the SPS[7] with bunches of 25 ns spacing and an energy of 26 GeV.

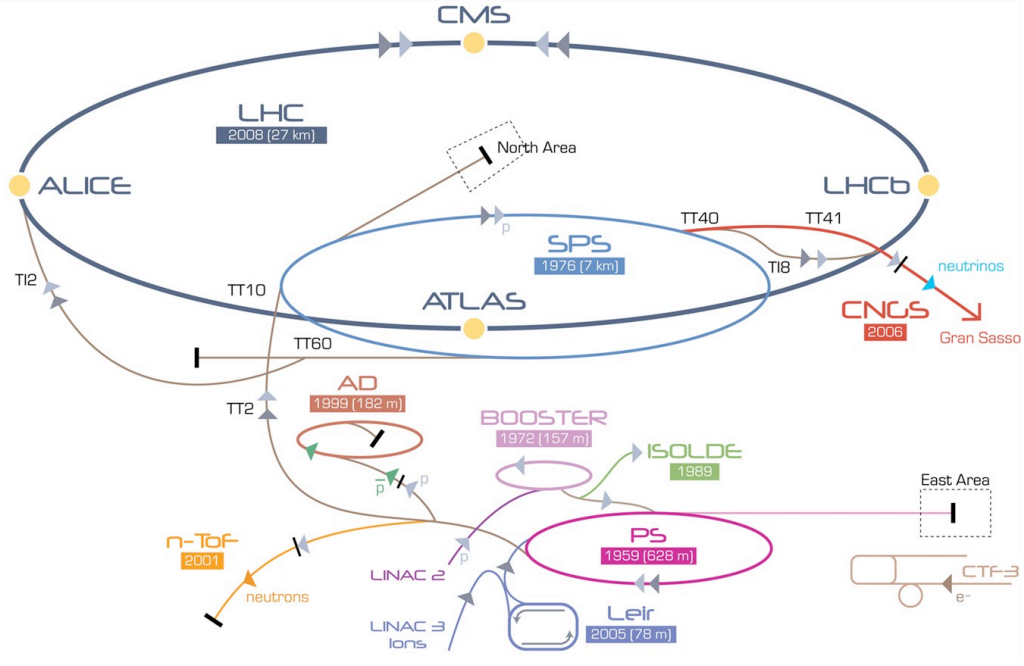

Figure 3.3: *Overview of CERN's accelerators and its chains into the LHC [53].*

The SPS itself, upgraded with a new superconducting radio frequency system, will accelerate the protons to an energy of 450 GeV and will finally fill the LHC. One full injection of the LHC requires twelve cycles of the SPS synchrotron and each SPS fill requires three or four cycles of the PS synchrotron. Counting 21.6 s for every SPS and 3.6 s for every PS cycle with some additional injection and machine adjustment cycles the minimum LHC injection time is 16 minutes. Further 20 minutes are needed for ramping the 2808 proton bunches in the LHC from 450 GeV to 7 TeV. Thus after a total time of about 40 minutes the LHC is ready for collisions at the highest centre of mass energies.

Due to interactions of the beams with their environment the luminosity lifetime is expected to be about 15 h. The anticipated time of data taking is around 6 to 12 hours per fill due to the luminosity decrease from collisions. With these parameters the maximum total integrated luminosity per year is expected to be between 80 fb^{-1} and 120 fb^{-1} depending on the average operating time of the machine.

3.1.3 The Current Machine Status

The mechanical construction of the LHC finished in November 2007 with the connection of the last two magnets. It took until September 2008 to cool down the ring, test the

[6] **P**roton **S**ynchrotron
[7] **S**uper **P**roton **S**ynchrotron

electrical connections and safety systems and commission the magnets up to 4 TeV beam energy. First partial injections of single beams already occurred in August leading to the first beam related events within the detectors (see section 3.2.9). Almost 20 years after the first workshop the LHC was launched on September the 10^{th} 2008. First beams were circulating within only one hour. In the following weeks further tests were made leading to a synchronization of the beam with the radio frequency and in total about 40 hours of circulating beams (see Figure 3.4).

On the 19th of September a fatal incident happened [54]: during the commissioning of the last sector to 5 TeV a faulty electrical connection in a region between two of the accelerator's magnets lead to an electric arc, which resulted in mechanical damage (see Figure 3.5) and the release of 6 tons of helium from the magnets' cold mass into the tunnel. Around 50 magnets had to be replaced. Further electrical connection tests revealed two other dipoles with faulty connections. Additional monitoring systems and measures to prevent a similar incident in the future are currently installed so that a restart of the LHC in 2009 can be envisaged.

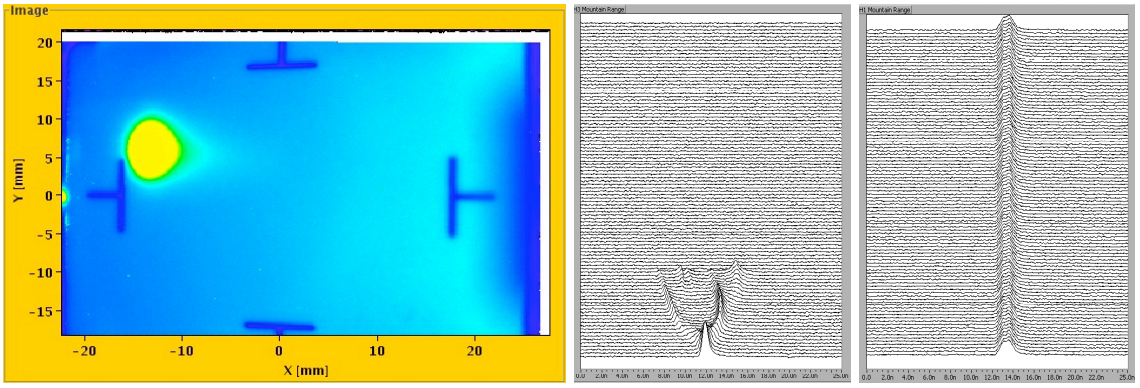

Figure 3.4: ***Left:*** *One of the first circulating beams captured by a beam monitor.* ***Right:*** *Beam monitor which shows the trials to synchronize the beam with the radio frequency [55]. On the left plot the phase is maximally wrong and thus the beam gets diluted, while on the right plot the capture succeeds and the beam could be driven for several hundred turns.*

Figure 3.5: *Mechanical damages caused by the LHC incident [56].*

3.1.4 LHC Physics Run at 10 TeV

The new LHC schedule has been fixed at the February 2009 LHC experiments committee workshop in Chamonix. It foresees first beams circulating inside the LHC at the end of September, with collisions following in late October. After a short stop over the Christmas period the LHC is planned to be steered till autumn 2010 to collect adequate data to carry out the first physics analyses. First beams will be injected and collided at the SPS centre of mass energy of 900 GeV. The beam energy will be increased in incremental steps, with tests at each stage, eventually reaching 5 TeV for the physics run. The goal is to record an integrated luminosity of more than 200 pb^{-1} at an operating energy of 5 TeV per beam. This thesis will therefore concentrate on the expected centre of mass energy of 10 TeV.

The impact of the reduction of the initial centre of mass energy from 14 TeV to 10 TeV can be illustrated by looking at the parton luminosity of two colliding partons a and b i.e. the available parton densities to create an object of a certain mass M_X.

$$\frac{d\mathcal{L}_{ab}}{dM_X^2} = \frac{1}{s}\int_\tau^1 \text{PDF}(x, f_a, M_X)\cdot\text{PDF}(\tau/x, f_b, M_X) \qquad \text{with} \qquad \tau = \frac{M_X^2}{s} \tag{3.5}$$

Figure 3.6 shows the ratio of this luminosity at 10 TeV compared to 14 TeV as a function of the mass M_X. The ratio is given for gluon-gluon and quark-anti-quark ($\sum_{q=u,\ldots,b} q\bar{q}$) initial states. One can estimate that the production rate of a hypothetical Z' with a mass of 1 TeV, which is mainly produced via quark-anti-quark partons, is reduced by roughly 50%. One should also notice that by the reduction of the centre of mass energy not only the cross section is reduced, but also the relative composition of the initial state partons (e.g. $t\bar{t}$-production at 14 TeV: gg: 90%, $q\bar{q}$: 10%; at 10 TeV: gg: 80%, $q\bar{q}$: 20%).

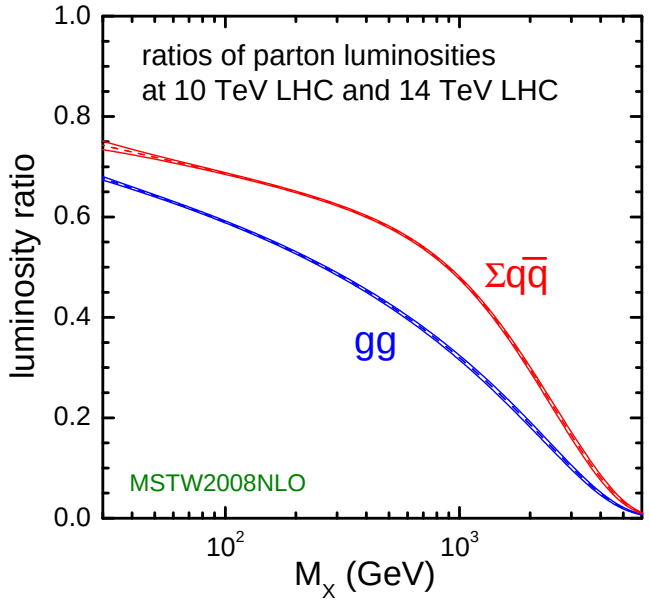

Figure 3.6: *Ratio of the LHC gluon-gluon and quark-anti-quark parton luminosities at 10 TeV compared to 14 TeV as a function of the invariant mass of the two colliding partons [57]. The ratio gives an estimate of the expected rate decrease due to the reduced centre of mass energy at the start-up of the LHC.*

3.2 The CMS Detector

CMS is a general-purpose detector which is built from various components to measure the particles which are directly or indirectly created within a pp-collision. The subdetectors are placed shell-like around the interaction point ordered by their tasks and with increasing material budget. Elements close to the beam line are built with as little material as possible to suppress multiple scattering and absorption of particles before their identification in the dedicated detector parts.

A first proposal of the CMS detector has been presented during an LHC workshop [58], which took place in Aachen in 1990. The proposal is based on a solenoid magnet with a highly performant muon system and a compact design.

Since then much effort has been spent on the research and development of the whole detector. Today's layout as shown in Figure 3.7 and 3.8 consists of a 4 Tesla solenoidal superconducting magnet, 13 m long with an inner diameter of 5.9 m. The view of the detector is dominated by the iron return yoke surrounding the magnet with five so-called wheels and two endcaps made of three discs each. In total CMS has a length of 21 m and an outer diameter of 15 m resulting in a weight of around 12500 t.

The detector is equipped with an all-silicon inner tracker to achieve a good spatial resolution of tracks within an environment of high particle fluxes. The high quality silicon strip tracker provides robust track and detailed vertex reconstruction measuring the momentum of charged tracks. A pixel vertex detector is mounted close to the beam pipe to allow for a precise vertex reconstruction and to identify secondary vertices arising for example from B-mesons and τ-leptons.

The electromagnetic calorimeter made of lead tungstate crystal and the brass-scintillator hadronic calorimeter will measure electromagnetic and hadronic showers from electrons, photons and jets, respectively. As the calorimeters are contained inside the magnet coil their performance is not affected by the coil and a high intrinsic resolution is guaranteed. In addition the strong magnetic field reduces the arrival of soft charged hadrons and other low energetic particles in the calorimeter.

CMS is completed by a redundant muon system embedded in the return yoke of the magnet. With its three different technologies and a nearly hermitic solid angle coverage up to $|\eta| = 2.4$ it is designed to identify and measure muons up to the TeV-energy region.

In order to reduce the event rate from the LHC bunch crossing rate of 40 MHz to about 100 Hz, which can be permanently stored on tape, a two-folded trigger system is arranged to filter interesting events. The short time between the bunch crossings requires a sophisticated read-out and Level-1 trigger system based on custom hardware and a high level trigger farm consisting of commercial PCs.

In the following sections the CMS detector is briefly described, starting from the innermost part and following a particle track to the outermost instruments. Details can be found in [59–67].

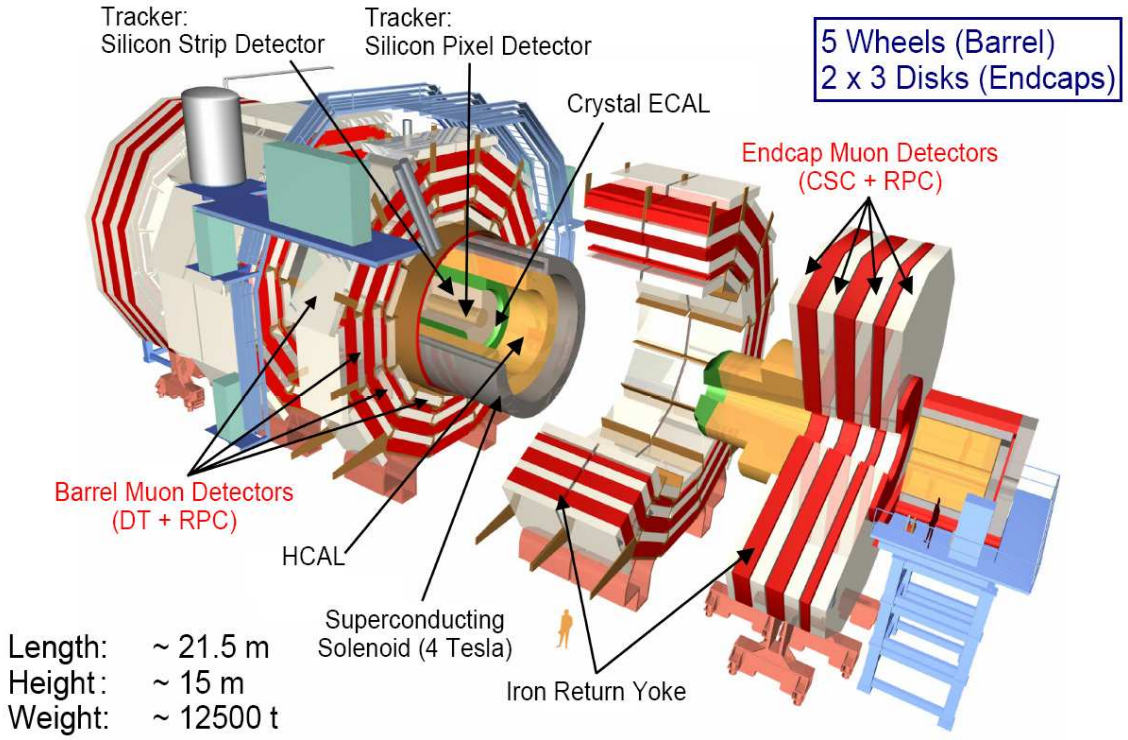

Figure 3.7: *Exploded view of the Compact Muon Solenoid [68].*

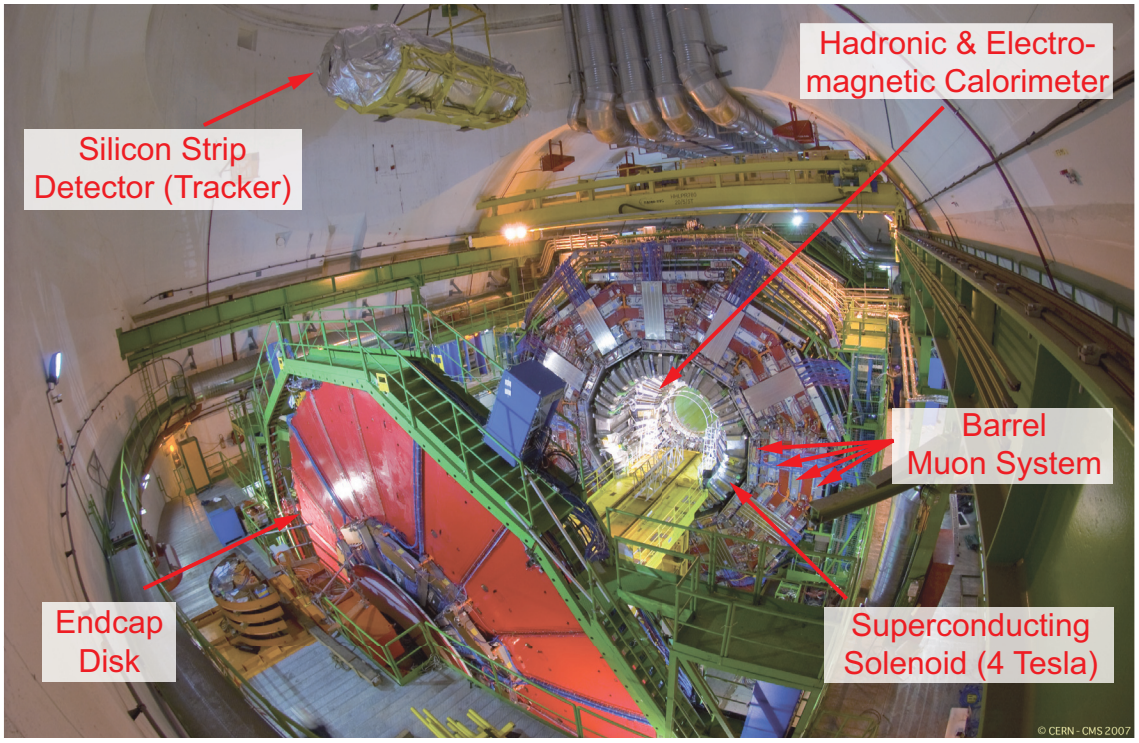

Figure 3.8: *The Compact Muon Solenoid at the final phase of its construction. The picture shows the lowering of the silicon strip tracker in December 2007 [53].*

3.2.1 The Silicon Pixel Detector

Several interesting events at the LHC are likely to contain secondary vertices, e.g. from b- or c-quarks or from τ-leptons. These particles are created at the pp-collision point, but travel a few millimeters before they decay at a secondary vertex. To allow for an efficient observation of these decays a high-resolution pixel detector is mounted as close as possible to the interaction point. Due to the close neighbourhood to the beam the detector is exposed to high particle fluxes resulting in a limited lifetime.

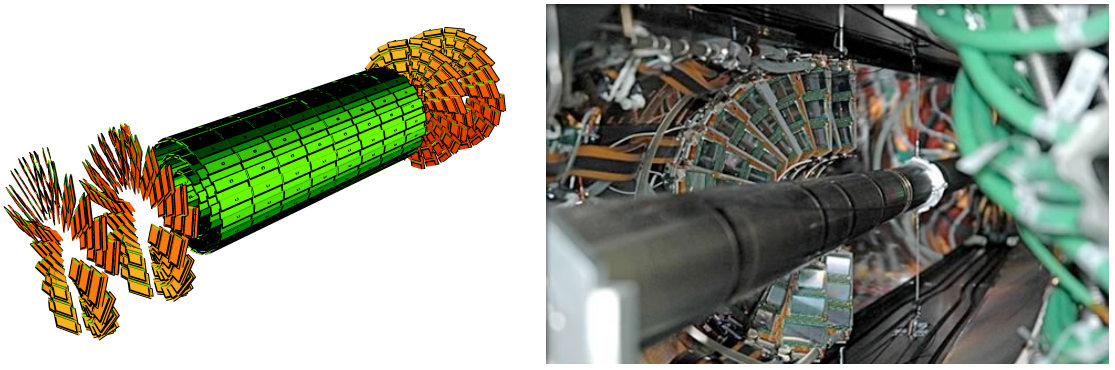

Figure 3.9: ***Left:*** *Schematic view of the pixel detector. The forward detectors are tilted by 20° in a turbine-like geometry to induce the charge-sharing necessary for a spatial resolution smaller than the size of a single pixel.* ***Right:*** *Installation of the endcap pixel detector at its final position close to the beam pipe and within the tracker.*

The pixel detector is expected to provide space point information with a high resolution and a minimum of two pixel hits per track to improve the ability to distinguish secondary vertices originating from long-lived objects against jets arising from light quarks and gluons. Therefore the CMS pixel system (see Figure 3.9) consists of three barrel layers and two pairs of forward and backward end discs.

The 53 cm long inner barrel layers reside at 4.4, 7.3 and 10.2 cm away from the nominal beam axis. The endcap discs with a radius from 6 –15 cm are placed at ±34.5 cm and ±46.5 cm in z-direction. The arrangement as shown in Figure 3.9 gives at least two pixel hits over almost the full geometrical coverage range of $|\eta| \leq 2.5$ for tracks originating from the centre of the interaction region. The radiation environment close to the interaction region will cause damage to the pixel sensors and readout chips and hence limit their lifetime to several years of LHC operation. The silicon detector is a good compromise between radiation hardness, cost, occupancy and achievable space point resolution. Under the assumption of an overall alignment precision within 10 μm a hit spatial resolution of about 15 – 20 μm can be obtained with a pixel size of 150 μm × 150 μm .

The readout is performed in an analog way to profit from effects of charge sharing among the pixels due to the 4 T magnetic field. Only via the use of charge interpolation among several pixels a hit resolution almost ten times smaller than the pixel size is obtained. To minimize the effect of radiation damages within the silicon the 48 million barrel and 18 million endcap pixels, covering in total an area of roughly 1 m^2, are operated at a temperature of $-10°$C.

The pixel detector allows a fast and efficient track seed generation from which the track reconstruction can start to extrapolate the particles into the silicon strip detector and further on.

3.2.2 The Silicon Strip Tracker

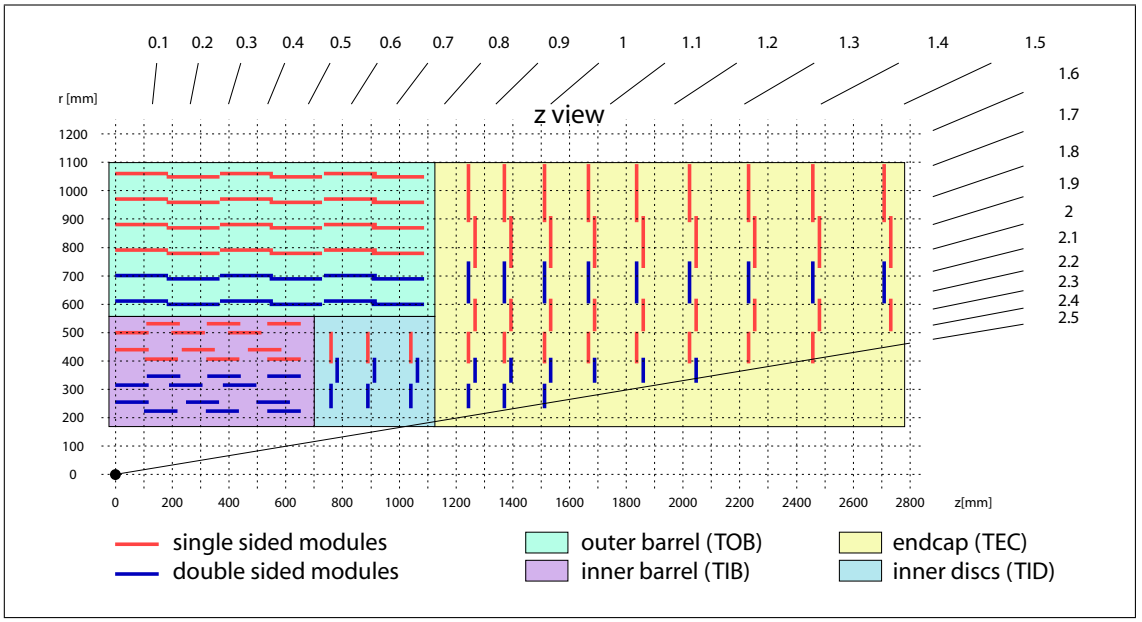

Figure 3.10: *Cross section of one quarter of the CMS silicon tracker [59].*

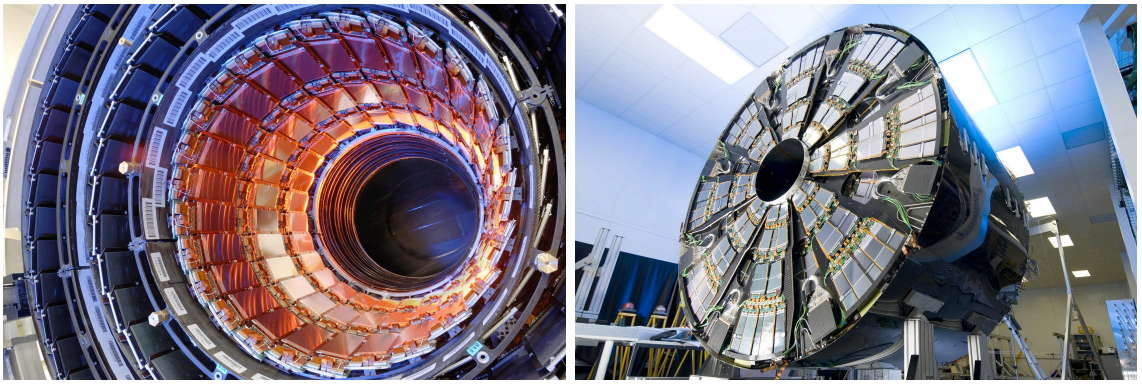

Figure 3.11: *Assembly of the tracker barrel (left) and the tracker endcap. The highly reflecting structures in both photos represent the silicon strip detector chips [53].*

The silicon strip tracker is designed to measure the transverse momentum of charged particles up to the TeV regime with high precision and efficiency. This is achieved by a high point resolution of the tracks bent in the magnetic field and a large number of measurements along the track. In conjunction with the pixel detector the tracker improves the impact parameter resolution with a sophisticated pattern recognition.

The tracker (see Figure 3.10) covers a cylindrical volume with a length of about 5.8 m and a radius between 0.2 and 1.2 m. An active area of approximately 200 m^2 of silicon detectors is mounted onto ten barrel layers and nine discs in each outer end-cap plus three mini-discs. They are arranged as shown in Figure 3.10.

The high rate of underlying events in one collision and a bunch crossing every 25 ns results in a very high charged particle flux in the tracker. Due to the strong magnetic field charged particles with less than a few GeV transverse momentum cannot leave the tracker and spiral until they are absorbed. At a radius of 22 cm still 10^6 charged particles penetrate the detector per square centimeter and second. Thus the tracking system requires a high granularity to separate close tracks and a fast response for the correct bunch crossing assignment. It must be radiation hard, but should consist of as little material as possible to e.g. reduce the conversion of photons before reaching the calorimeter.

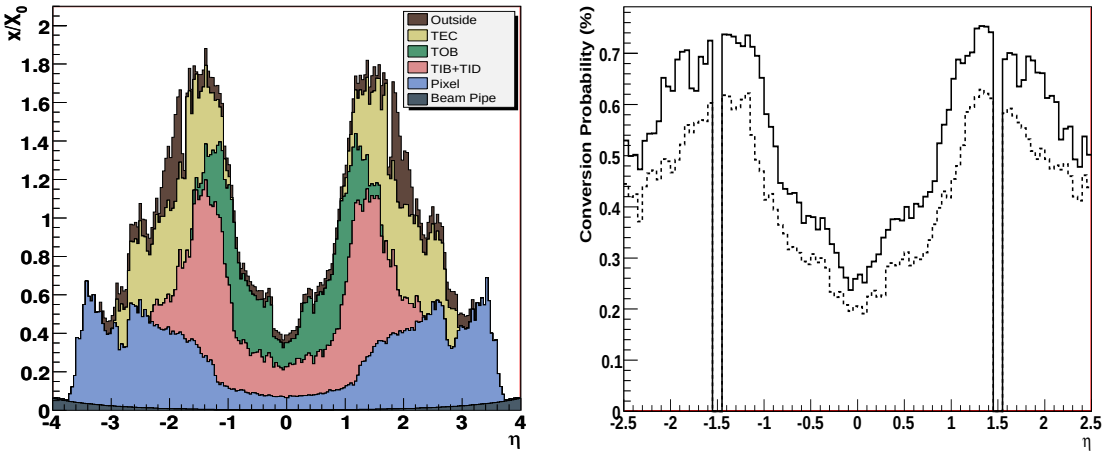

Figure 3.12: ***Left:*** *Material budget of the tracker in radiation lengths as a function of pseudorapidity.* ***Right:*** *Probability for a photon with an energy of 20 – 150* GeV *to convert into an electron-positron pair before reaching the electromagnetic calorimeter [69]. A dedicated reconstruction algorithm is required for converted photons.*

The tracker covers an $|\eta|$-range up to 2.5, in which electrons and muons up to several 100 GeV transverse momentum are reconstructed with an efficiency larger than 98%, a track fake rate below 1%, and an expected momentum resolution, which is for isolated charged leptons approximately given by [70]

$$\frac{\Delta p_{\mathrm{T}}}{p_{\mathrm{T}}} = 0.15\,\frac{p_{\mathrm{T}}}{\mathrm{TeV}} \oplus 0.5\%. \tag{3.6}$$

As shown by detector simulations a good determination of the track parameters with only 4 – 6 hits allows fast and clean pattern recognition. The whole tracker has to be kept at -10°C to ensure that the silicon survives the harsh radiation environment of the LHC.

3.2.3 The Electromagnetic Calorimeter

The electromagnetic calorimeter (ECAL) [60] measures the energy and the direction of electromagnetically interacting particles like electrons, photons or parts of the electromag-

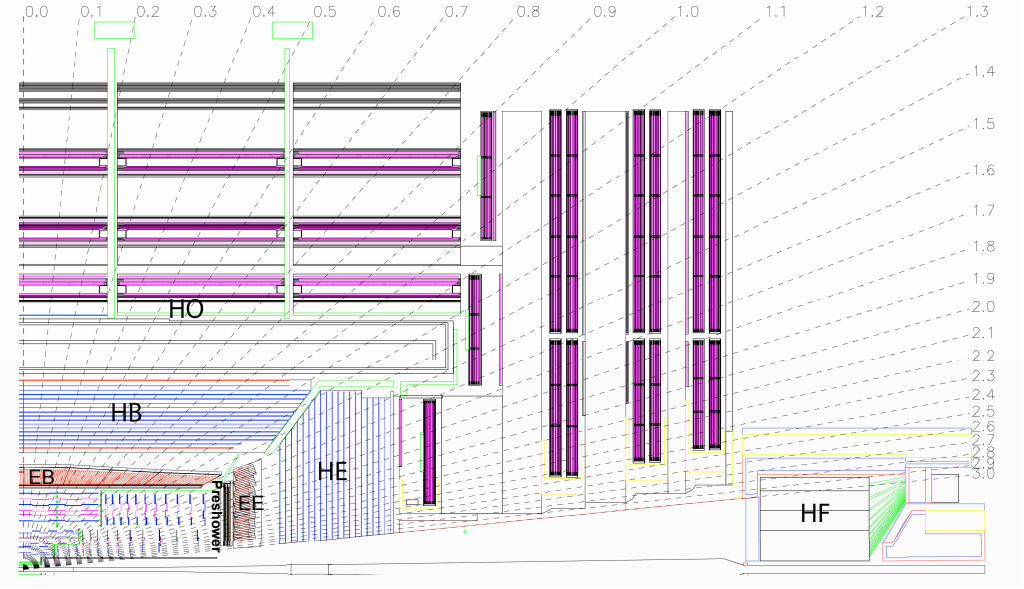

Figure 3.13: *One quadrant of the CMS calorimeters [60]. The tracker is surrounded by the barrel electromagnetic (EB) and hadronic barrel calorimeter (HB). In the direction of the beam line the calorimeter is completed by the electromagnetic (EE) and hadronic endcap calorimeters (HE).*

netic fraction of jets with high precision by absorbing these particles inside scintillating crystals. To meet the LHC requirements of radiation hardness and to achieve a high energy resolution $PbWO_4$ has been chosen as scintillator. It has a high density and therefore a short radiation length[8] X_0 and a small Moliere radius[9] of 22 mm. This allows a very compact electromagnetic calorimeter contained within the solenoid which fits into the design of CMS.

Special efforts have been made for the development of crystals, photodetectors, electronics and software to meet the challenging LHC requirements of an average of 1000 charged tracks penetrating the ECAL every 25 ns. The readout is done by special avalanche photodiodes in the barrel and vacuum photo triodes in the endcaps, which are both insensitive to high magnetic fields. They amplify the light gained from the crystals and measure the energy deposit.

Because of the strong temperature dependence of the crystal light yield and of the diode gains, the temperature inside the calorimeter has to be kept constant within 0.1°C to guarantee a precise operation of the ECAL [60].

The ECAL is built of a cylindrical barrel with a length of around 6 m, an inner radius of 1.3 m and an outer radius of 1.8 m (see Figure 3.13 and 3.14). Endcaps are located in forward and backward direction at ± 3.2 m with an extension of 0.7 m along the z-direction. With these dimensions the crystals hermetically cover an $|\eta|$-range up to 3.0. The precision of the energy measurement for electrons and photons is limited by the amount of pileup energy deposited and the tracker coverage up to $|\eta| = 2.5$. The shape of the

[8] The energy of a high-energetic electron ($E \gg 1$ MeV) has dropped to 1/e - on average - after passing the distance of one radiation length X_0.

[9] In a cylinder with a radius of a Moliere radius on average 95 % of the electromagnetic shower energy is contained.

approximately 60000 barrel and 20000 endcap crystals is chosen so that their front face (22 x 22 mm^2) points to the interaction region (pseudo-projective geometry). This corresponds to a granularity of $\Delta\eta \times \Delta\phi = 0.0175 \times 0.0175$ in the ECAL barrel which grows progressively with η to a maximum of $\Delta\eta \times \Delta\phi \approx 0.05 \times 0.05$. The typical crystal depth of 230 mm equals 26 radiation lengths X_0. For trigger purposes arrays of 5×5 crystals are grouped to one ECAL trigger tower which coincides with the HCAL tower granularity.

The neutral pion and photon separation is improved by an endcap preshower detector installed in front of each ECAL endcap [71]. It covers a pseudorapidity range from $1.65 < |\eta| < 2.61$ and consists of a lead absorber to initiate photon showers. Its thickness of 2.8 X_0 is well-adapted to guarantee a 95% conversion probability and to prevent a degradation of the excellent crystal calorimeter energy resolution. The readout is performed by silicon sensors which act as energy sampling devices. The preshower detector improves the π^0/γ but also the $e^{\pm}/\pi^{\pm}$ separation and enhances the spatial resolution of the calorimeter.

Using the notation $a \oplus b := \sqrt{a^2 + b^2}$, the energy resolution of a calorimeter can be described by

$$\frac{\sigma(E)}{E} = a \cdot \frac{\sqrt{\text{GeV}}}{\sqrt{E}} \oplus b \cdot \frac{\text{GeV}}{E} \oplus c. \tag{3.7}$$

The term a, called stochastic term, reflects the shower fluctuations, the photon-statistics and the fluctuation of the transverse leakage of the produced shower in the calorimeter. The value of a determined within test beams is approximately 2.1% for the barrel and 5% for the endcap calorimeter [66]. The so-called noise term b comprises the electronic noise including dark currents and pileup of overlapping events. The noise term corresponding to a cluster of 5×5 crystals is expected to be about 150 MeV (210 MeV) for the barrel and 205 MeV (245 MeV) for the endcaps at low (high) luminosity. The constant term c of about 0.3% results from intercalibration errors, crystal non-uniformity and shower leakage [66].

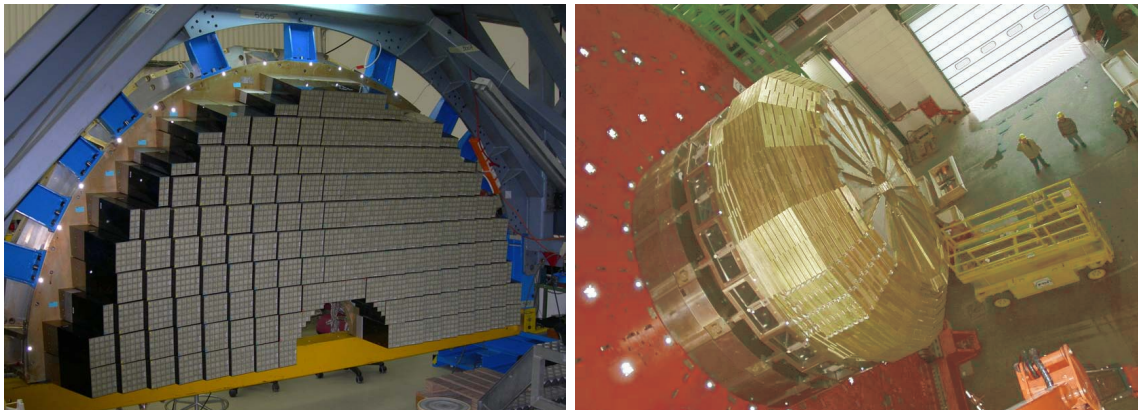

Figure 3.14: ***Left:*** *One half of the electromagnetic calorimeter endcap (a so-called Dee) with groups of crystals.* ***Right:*** *Shining brass of one of the partially assembled hadron calorimeter endcaps [66].*

3.2.4 The Hadronic Calorimeter

The CMS detector is equipped with four different hadronic calorimeters [61], featuring a good segmentation, a moderate energy resolution and a full angular coverage up to $|\eta| = 5$. As displayed in Figure 3.13 the barrel hadronic calorimeter (HB) is contained within the magnet coil and surrounds the electromagnetic calorimeter up to a pseudorapidity of $|\eta| = 1.3$. It is completed by two endcap hadron calorimeters (HE), $|\eta| \leq 3$, also located inside the solenoid and extended by the two (very) forward calorimeters (HF), surrounding the beam pipe 11 m away from the interaction point. In addition the central shower containment is improved with an array of scintillators located outside the magnet labeled as outer hadronic calorimeter (HO).

The HCAL measures the hadronic component of jets and other hadronic particles. Due to the hermetic layout of both, the electromagnetic and hadronic calorimeters, the transverse component of the energy imbalance can be calculated. Thus, neutrinos or other particles not interacting inside the detector, can be seen indirectly.

Hadronic Calorimeter: Barrel and Endcaps

For the HB and HE placed inside the magnet the collaboration decided to use a sampling calorimeter made of brass and plastic scintillators, which are read out by wavelength-shifting plastic fibres. The HB is divided into two cylindrical sections segmented into 18 identical wedges. Each wedge, aligned parallel to the beam axis, consists of alternating 17 layers of 5 – 8 cm brass and readout scintillators divided into $\Delta\eta \times \Delta\phi = 0.087 \times 0.087$ segments. It is sandwiched by stainless steel for structural strength.

The HE consists of 18 20°-modules, each made of 19 layers of brass and scintillator with the same transverse segmentation as the HB to match the trigger tower granularity of the ECAL. While the HB has a minimum depth of 5.8 nuclear interaction lengths[10] λ_I, the HE consists of at least 10 interaction lengths λ_I.

The Forward Calorimeters

The HF calorimeters (1.65 m length, 1.4 m radius) are made of steel absorbers and embedded radiation hard quartz fibres, which provide a fast collection of Cherenkov radiation by photomultipliers. With a depth of roughly 9 λ_I it is a crucial tool to improve the missing energy detection and also useful to tag forward jets to reduce backgrounds in signal reactions without associated jet production in forward direction.

Charged particles entering the HF produce particle showers in which only electrons and positrons are fast enough to produce Cherenkov light. Thus the calorimeter is mainly sensitive to the electromagnetic component of showers, providing a very clean and fast signal. In addition it is used for luminosity monitoring.

[10] On average a hadronic interaction occurs at one nuclear interaction length λ_I.

Outer Hadronic Calorimeter

In the barrel region a particle has to pass about 8 nuclear interaction lengths until it reaches the magnet. That means, that for a 300 GeV pion 5% of the energy would be deposited beyond the outer limits of the HB. To improve the shower containment two layers of scintillators attached to a 20 cm thick piece of iron are located outside the solenoid but in front the first muon station. This extends the total depth of the HB to 11.8 λ_{I} with an improvement in linearity and resolution.

The overall resolution for pions using the complete calorimeter system including both, the electromagnetic and the hadronic calorimeter, has been determined in test beams [72] to

$$\frac{\sigma(E)}{E} = \frac{0.7\ \sqrt{\mathrm{GeV}}}{\sqrt{E}} \oplus \frac{1\ \mathrm{GeV}}{E} \oplus 8.0\%. \tag{3.8}$$

The very forward calorimeters CASTOR (Centauro And Strange Object Research) and ZDC (Zero Degree Calorimeter) with a coverage up to $|\eta| \approx 10$ complete the CMS physics programme with diffractive and low-x physics within proton-proton but also heavy ion collisions. The dedicated TOTEM experiment [73] is also placed in the forward direction. Its main task is the determination of the total proton-proton cross section (see section 3.2.8).

3.2.5 The Superconducting Solenoid

The CMS detector is equipped with a superconducting solenoid [62] bending the tracks of charged particles and thus allows to measure their transverse momentum. The superconducting coil with a length of 13 m and a diameter of about 5.9 m is located inside the barrel wheels, which constitute the return yoke (see Figure 3.7). The magnet is cooled with liquid helium 4 K. As shown in Figure 3.15 the magnetic field is designed to reach up to 4 T. Especially in the endcaps the magnetic field is quite inhomogenous. Fully powered the magnet stores an energy of 2.7 GJ.

3.2.6 The Muon System

As implied by the name of the detector, CMS is specially focused on triggering and reconstruction of muons, which give clear signatures for a variety of physics processes. Muons appear for example within the "golden channel" for the Standard Model Higgs searches $\mathrm{H} \rightarrow \mathrm{ZZ} \rightarrow 4\mu$, within the decay of new hypothetical heavy gauge bosons $Z' \rightarrow \mu\mu$ or supersymmetric events. Apart from the identification the muon system determines the momentum as well as the charge of the muons by measuring the track bending due to the magnetic field with three different types of gaseous detectors.

The choice of the detector technology is driven by the very large surface to be covered, the magnetic field, the precision needed, and the different radiation environments. Beside the crucial features of muon identification and bunch crossing assignment, the p_{T} measurement especially for high momentum muons is performed by the muon system. It has a spatial resolution of the order of 100 μm. Due to the multiple scattering of the muons in the iron

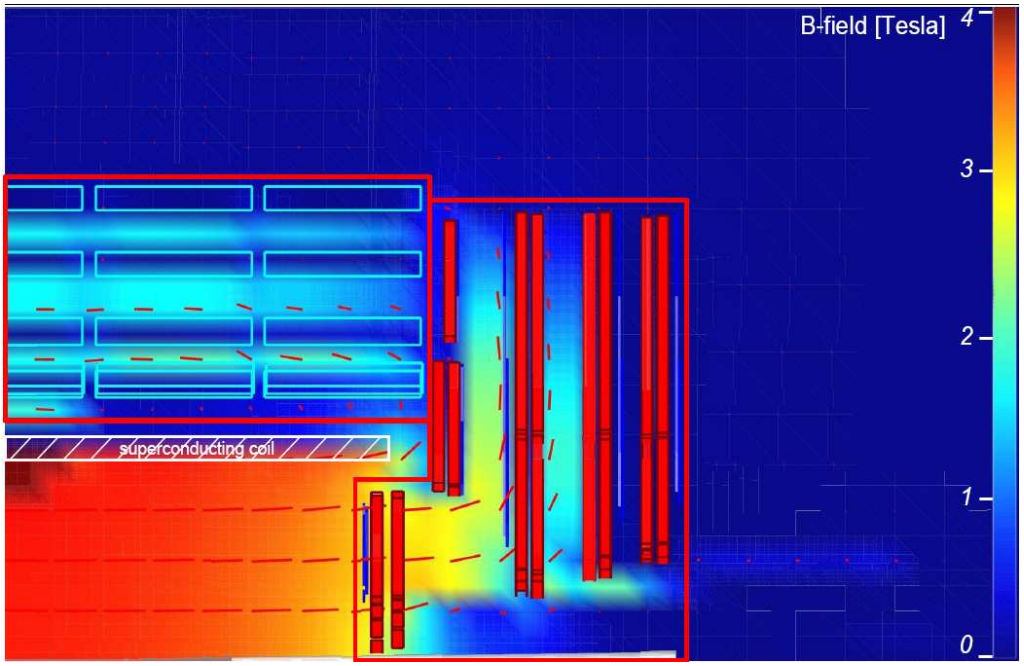

Figure 3.15: *The magnetic field within one quarter of the CMS detector [74].*

of the return yoke the overall p_T resolution for low momentum muons ($p_T < 200$ GeV) is determined by the tracker.

The muon system (see Figure 3.16) is embedded in the iron return yoke of the magnet. It consists of four stations, arranged as concentric cylinders around the beam pipe in the barrel region and as discs perpendicular to the beam line in the endcaps. The 10 interaction lengths before the first muon station and another 10 from the iron yoke before the last station, guarantee that no other particles than muons (with an energy of more than 5 GeV) and neutrinos pass the muon system. This ensures a muon identification efficiency above 95%.

Three different technologies are employed in the almost hermetic muon system: in the barrel drift tubes (DT) are installed, where the occupancy, the background noise and the residual magnetic field are relatively low compared to the endcaps. Here cathode strip chambers (CSC) are used. In both regions resistive plate chambers (RPC) provide an additional independent measurement for trigger purposes with a superior time, but a lower spatial resolution. The muon system covers regions up to $|\eta| = 1.2$ for DTs, $|\eta| = 2.4$ for CSCs and RPCs.

The Drift Tube Chambers

In the barrel region of the CMS muon system, drift tube chambers face a moderate environment: the pollution from radiation and charged particles is one of the lowest inside CMS. Due to the flux containment inside the iron yoke (see Figure 3.15) the almost uniform magnetic field inside the chambers has a strength less than 1 T.

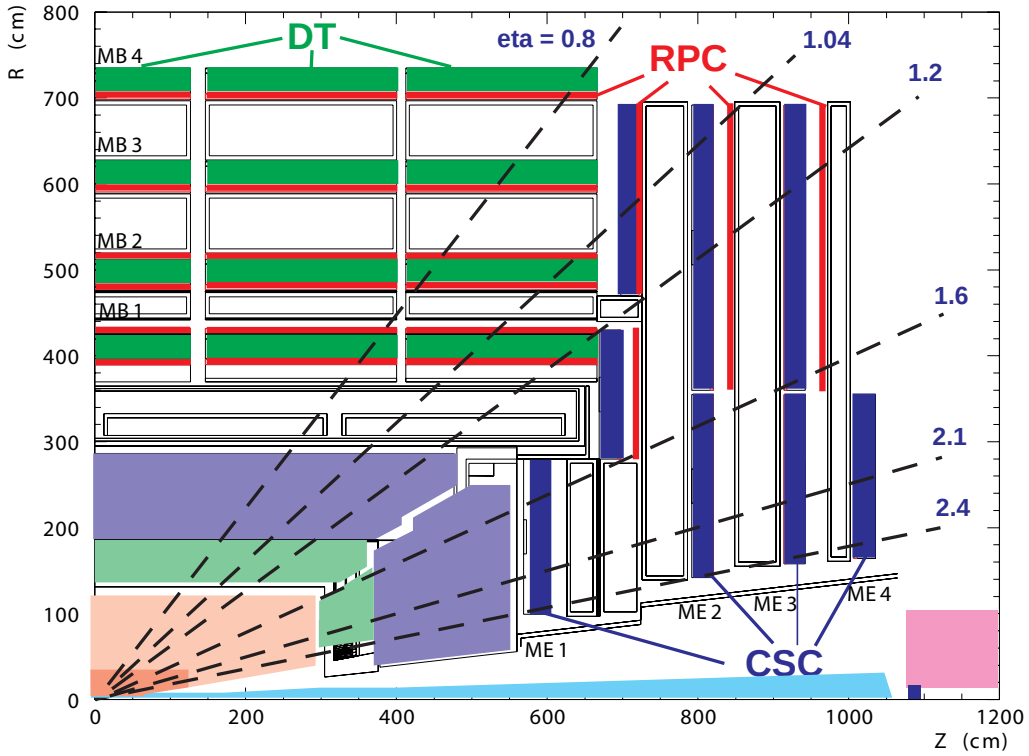

Figure 3.16: *Layout of one quarter of the CMS muon system comprising three technologies: drift tubes (DT) in the barrel, cathode strip chambers (CSC) in the endcap and resistive plate chambers (RPC) in both regions [72].*

The drift tube system consists of four concentric cylinders with growing diameter centered around the beam pipe. Each wheel is divided into 12 azimuthal sectors which cover approximately 30° each. One segment, a single "chamber", is the basic unit in the DT system, which in total consists of 250 chambers. With a length of 2.5 m the chambers follow the segmentation of the return yoke wheels they are mounted at (see Figure 3.17).

A chamber is made of three "superlayers" (SL)[11], where the inner and outer so-called ϕ-SL are separated maximally to increase the lever arm for the track measurement in the $r\phi$-bending plane. To be able to reconstruct a $3D$-segment within a single chamber, the middle superlayer is rotated by 90° to provide a measurement of the z-coordinate (Θ-SL[12]). Both types of superlayers have the same substructure of four layers of stacked drift tube cells.

A basic cell (see Figure 3.18) has a width of 42 mm and a height of 13 mm. The length depends on the superlayer type: it is roughly 2.5 m for the ϕ-SL and increases for the Θ-SL with the distance from the beam line. Along the centre of a drift cell a 50 μm gold plated steel wire serves as an anode. Within a single cell only the absolute distance to the wire can be measured. This results in a so-called "left-right ambiguity" which is resolved by staggering the four layers of cells inside a superlayer by half a cell width.

[11] The outermost chambers consist of only two superlayers.

[12] This kind of SL is missing in the outermost stations. Thus a $3D$-segment reconstruction within this chambers is not possible. The r-coordinate is always given by the location of the detector component.

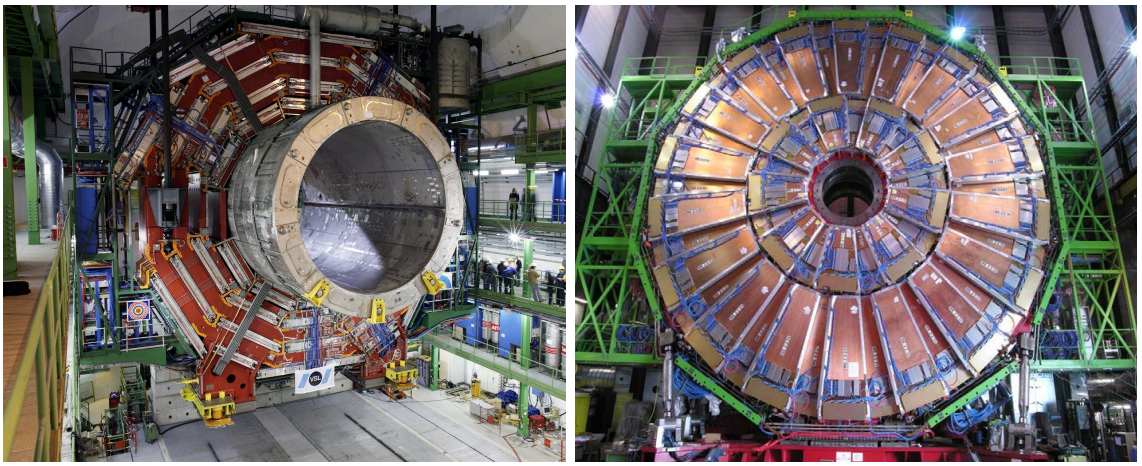

Figure 3.17: ***Left:*** *Lowering of the CMS central barrel wheel with the magnet coil. The muon drift tubes (silver) are inserted in the iron return yoke (red) [53].* ***Right:*** *End cap disc with attached cathode strip chambers [66].*

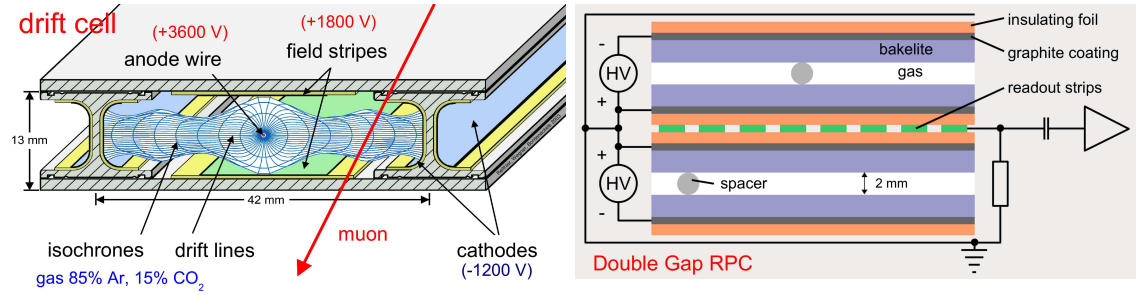

Figure 3.18: ***Left:*** *Cross section of a CMS drift cell with drift lines of electrons and isochrones [75].* ***Right:*** *Cross section of a double gap resistive plate chamber [67].*

The cathodes located at the edges of the cell are mounted at "I"-shaped aluminium beams, which isolate one cell from the other. Field shaping electrodes at the top and bottom of a cell improve the linearity of the space-drifttime-relation. The cells are flushed with a gas mixture of 85% Ar and 15% CO_2, which provides good quenching properties and a drift velocity of about 55 μm/ns.

This results in a maximum drift time of about 380 – 400 ns, which equals the time of $\approx$16 bunch crossings. A hit inside a cell can be measured with a precision of approximately 190 μs and an efficiency of more than 99% [76].

Cathode Strip Chambers

The cathode strip chambers are located in an environment of a highly non-uniform magnetic field (up to 3.1 T, see Figure 3.15), a high flux of charged particles and an intense rate of neutron background (up to 1000 Hz/cm^2). The CSC system is arranged in four discs per endcap yoke in a plane perpendicular to the beam.

Beside the innermost station, which is divided into three concentric rings of chambers, all other stations consist of two rings. These rings are segmented into 18 trapezoidal chambers

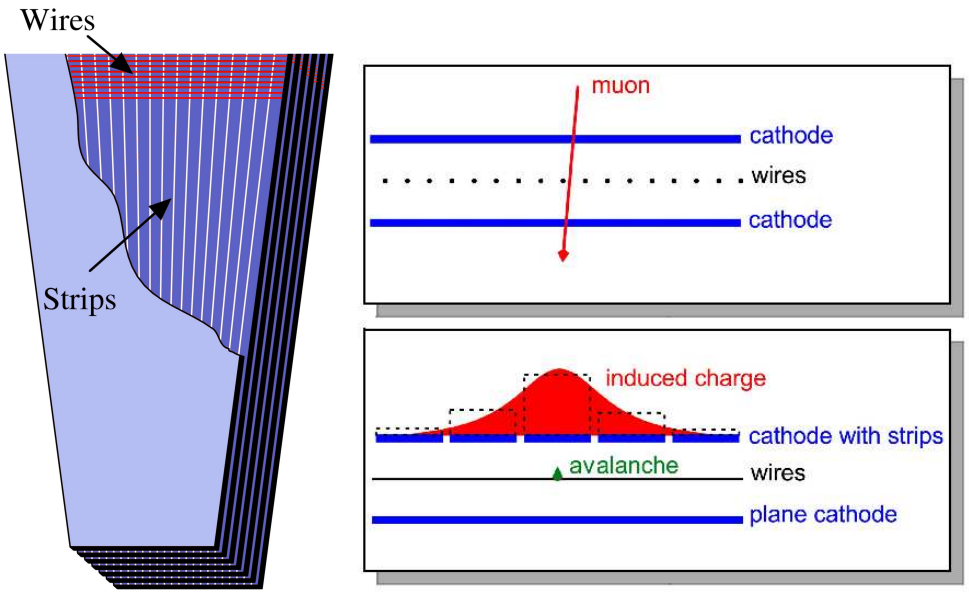

Figure 3.19: *Sketch of a muon cathode strip chamber (CSC) and its functional principle [63].*

for the inner rings and into 36 chambers for the other rings. Apart from the outermost ring of the first station all chambers have overlaps in the $r\phi$-plane to avoid dead regions.

The CSC system is constructed to achieve a high muon detection efficiency. It provides a robust and background rejecting pattern recognition and improves the bunch crossing assignment. A single chamber is composed of six equal layers of active volume. Each layer is a multi-wire proportional chamber (see Figure 3.19) defined by an array of 50 μs anode wires sandwiched between two parallel cathode planes, which are separated by a 9.5 mm gas gap (filled with a mixture of 30% Ar, 50% CO_2 and 20% CF_4). The cathodes are segmented into strips, which are aligned perpendicular to the wires in radial direction. Their width is chosen to cover a $\Delta\phi$-slice between 2 and 5 mrad and thus are also trapezoidal. A voltage of 4.1 kV is applied.

The chambers of the ring with the closest distance to the interaction point show minor differences in their construction. Due to the high magnetic field of about 3 T oriented along the z-direction and the resulting skewed drift of electrons, the gas gap is only 6 mm wide, the high-voltage counts roughly 3 kV and the wires, having a diameter of 30 μm are strung at a 25° angle in the chamber plane.

Since the signals are read out from the strips as well as from the wires, the CSCs are fast detectors suitable for triggering. Electrons from the gas ionisation along a muon track drift to the array of wires and develop an avalanche due to the increasing electric field. The moving charges induce a signal on several of the strips in the cathode plane. The interpolation of induced charges between adjacent strips results in a very fine spatial resolution in the $r\phi$-plane of about 50 μm at normal muon incidence [63]. Simultaneously, the signal on the wires is read out to gain a measurement of the radial coordinate with a coarse precision of a few mm.

Resistive Plate Chambers

The resistive plate chamber system is complementary to the other muon detectors. With their reasonable spatial resolution, but excellent time resolution of a few nanoseconds, they are specifically designed for trigger purposes and add robustness and redundancy to the muon system.

In the barrel region the RPCs are directly attached to the DT chambers. The first two DT stations are sandwiched by RPCs to provide at least four measurements even for lower energetic muons, while only one RPC is attached to each of the outer two stations. In the endcaps, trapezoidal shaped resistive plate chambers are combined with the CSC system, resulting in four discs which cover a range up to $|\eta| = 2.4$.

A single RPC chamber is made of a pair of parallel bakelite plates, separated by a 2 mm small gap filled with a gas mixture of 96% $C_2H_2F_4$, 3.5% i-C_4H_{10} and 0.5% SF_6 (for streamer supression). For an improved efficiency per station double gap RPCs are used as shown in Figure 3.18. The highly resistive plates are coated with graphite electrodes to apply the high voltage of 9.5 kV. Insulated aluminium strips are placed between the two gas gaps as a common readout.

This double-gap layout is chosen to compensate the weaker induced signal caused by the operation of the RPCs in the "avalanche" mode rather than in the more common "streamer" mode, to sustain higher rates. However, the gas amplification is reduced and an improved electronic gain is required.

In the barrel the RPC readout strips, with a length of 80 or 120 cm, are aligned parallel to the beam line while the strips in the endcaps, with a length of 25 to 80 cm, are orientated perpendicular to the beam line. The width is chosen to cover always $(5/16)^\circ$ in the ϕ-coordinate and thus increases with the distance to the beam. By signal interpolation of adjacent strips this coordinate is measured, while the position parallel to the strip is only constrained by the strip length.

A critical point in the operation of the RPCs is the flatness of the bakelite surface. Local bumpiness results in an increase of the electric field and thus to intrinsic noise. A solution for surface smoothing is the treatment of the bakelite electrodes with linseed oil, which also absorbs UV quanta from avalanches. CMS has made the choice of oiling all barrel and endcap RPCs up to $|\eta| = 1.6$. The remaining RPCs are supposed to be non-oiled to avoid potential aging effects, which might be related to the degeneracy of the oil in this region due to very high particle fluxes [77].

The momentum resolution $\Delta p_T/p_T$ of the muon system stand-alone is expected to be 8 – 15% (20 – 40%) for muons with transverse momenta of 10 GeV (1 TeV) depending on $|\eta|$. In combination with the tracker the resolution can be improved to 1 – 1.5% (6 – 17%).

3.2.7 The CMS Trigger and Data Acquisition

The LHC environment presents challenges to the trigger and data acquisition system [64, 65] much more demanding than those encountered at past and present experiments worldwide. The bunch crossing rate of 40 MHz and an average of 25 interactions per bunch crossing plus additional overlapping events result in approximately 10^9 interactions

per second. CMS has more than 10^8 readout channels resulting in a data rate of the order of 10^{15} bits per second at full operation. After zero suppression still 1.5 MB of data per event will emerge from the high level trigger farm (HLT) and will be stored permanently. Since today's permanent storage devices such as tape drives are only able to cope with a data rate of up to 300 Hz, only events containing "interesting" physics are sorted out and written to tape. Thus the number of events has to be reduced by a factor of 10^7.

HLT Path	L1 Condition	HLT Threshold [GeV]	Rate [Hz]
Single Isolated μ	A_SingleMu7	11	18.3 ± 2.2
Single Relaxed μ	A_SingleMu7	16	11.4 ± 0.8
Double Relaxed μ	A_DoubleMu3	(3, 3)	10.8 ± 1.3
Single Isolated e	A_SingleIsoEG12	15	17.1 ± 2.3
Single Relaxed e	A_SingleEG15	17	1.8 ± 0.2
Single Isolated γ	A_SingleIsoEG12	30	8.2 ± 0.7
Single-Jet	A_SingleJet150	200	8.8 ± 0.1
Double-Jet	A_SingleJet150 A_DoubleJet70	150	4.3 ± 0.0
Triple-Jet	many	85	4.4 ± 0.1
$\not{E}_{\mathrm{T}}$	A_ETM40	65	3.5 ± 0.4
Double τ	A_DoubleTauJet40	15	4.7 ± 0.6
μ + Jet	A_Mu5_Jet15	(7, 40)	4.0 ± 0.4
e + Jet	A_IsoEG10_Jet30	(12, 40)	6.4 ± 0.6
Minimum-bias	A_MinBias_HTT10	-	1.5 ± 0.0

Table 3.2: *Excerpt of the high level trigger paths which are expected to predominantly contribute to the total trigger band width at the early stage of data taking at a luminosity of up to* $\mathcal{L} = 10^{32}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. *The estimate uses a safety factor of 2 and thus the total trigger rate sums up to a total of 150* Hz *[78].*

The CMS level-1 trigger is designed to reduce the initial bunch-crossing rate of 40 MHz to 100 kHz. Using only coarse detector data from muon detectors and calorimeters the first level trigger generates dead time free decisions every 25 ns with the thresholds and rates given in Table 3.2. Due to the limited storage capacity of detector readout buffers the decision must be available 3.2 μs after the corresponding bunch crossing.

The reduction of the rate is performed in several steps, which form a series of progressively more complex, but also time consuming levels. The first level (level-1 trigger) lowers the rate of events from 40 MHz to at least 50 kHz. The following levels comprised as high-level trigger have more time for the decision and further reduce the rate to finally less than 300 Hz. The first level is based on custom pipelined hardware processors, whereas the HLT is based on standard computer systems.

If an event is accepted at level 1 the full detector information is read out and passed to the high-level trigger online farm of about 1000 commercial CPU's. Highly sophisticated algorithms are used to reconstruct the event. Finally events containing "interesting" physics are written to tape with a rate of up to 300 Hz.

3.2.8 Luminosity Monitoring

The luminosity relates the cross section σ to the event rate according to equation (3.2). Therefore it is the most important parameter of the LHC apart from the centre of mass energy. Its precise determination and monitoring is necessary during the whole operation of the LHC. There are several methods to provide such a measurement. Two of them are discussed here.

Direct measurements

By the measurement of the beam parameters, such as the bunch geometry and the particle density within the beam, the luminosity can be obtained from equation (3.3). This method does not result in a very precise luminosity measurement ($\Delta\mathcal{L}/\mathcal{L} \approx 10\%$) because an accurate measurement of the beam currents and especially of the beam size at the interaction point is difficult.

A second direct method is based on equation (3.2). If the rate of a special process can be measured precisely and its cross section is well-known from theoretical calculations, the luminosity is given as the ratio of both. Ideal candidates are W- and Z-production. Due to their high production rate they allow an instantaneous luminosity measurement within minutes. Even $t\bar{t}$-production might yield as a standard candle especially as a reference for processes with gg initial states. The precision is limited by experimental corrections to the rate, like detector acceptance and efficiencies. The precision of the luminosity measurement, which can be achieved with this method, is comparable to the first method.

If one measures the total inelastic and diffractive cross section (see next paragraph) of roughly $\sigma = 80$ mb, one can count the number of interactions per bunch crossing and obtains a Poisson distribution with a mean of

$$\mu = \frac{\sigma L}{f_{\mathrm{BX}}} \tag{3.9}$$

This mean can be elegantly measured by performing "zero counting" within the hadronic forward calorimeter i.e. one counts how often no interaction is seen within the detector: $\mu = -\ln p(0)$). The method requires an absolute calibration and a not too high luminosity still having enough crossings with zero interactions. The relative luminosity precision is expected to be 5%.

Measurement via the Optical Theorem

Using the TOTEM detector [73] the luminosity will be determined through the measurement of the total cross section which will be used to arrive at an absolute luminosity normalization. It is based on the simultaneous measurement of small angle elastic scattering and of the total inelastic rate. The total cross section σ_{tot} can be expressed in terms of the number of elastic and inelastic interactions N_{el} and N_{inel} within an integrated luminosity $\mathcal{L}_{\mathrm{int}}$ by

$$N_{\mathrm{el}} + N_{\mathrm{inel}} = \mathcal{L}_{\mathrm{int}}\, \sigma_{\mathrm{tot}}. \tag{3.10}$$

Taking the optical theorem into account, which relates the total cross section σ_{tot} to the imaginary part of the forward scattering amplitude $F(0)$,

$$\sigma_{\mathrm{tot}} = \frac{4\pi}{p^*} \operatorname{Im}(F(0)) \tag{3.11}$$

one can transform the differential elastic scattering at zero angle,

$$\left(\frac{d\sigma_{\mathrm{el}}}{d\Omega^*}\right)_{\theta=0^\circ} = |F(0)|^2 = (\operatorname{Re}(F(0)))^2 + (\operatorname{Im}(F(0)))^2 \tag{3.12}$$

into

$$\left(\frac{d\sigma_{\mathrm{el}}}{d\Omega^*}\right)_{\theta=0^\circ} = (1+\rho^2)(\operatorname{Im}(F(0)))^2 = (1+\rho^2)\left(\frac{p^*\sigma_{\mathrm{tot}}}{4\pi}\right)^2. \tag{3.13}$$

p^* is the momentum of the scattering particles in the rest frame and ρ has been defined as ratio $\rho = \operatorname{Re}(F(0))/\operatorname{Im}(F(0))$.

Replacing the differential cross section per rest frame solid angle Ω^* by the differential cross section per momentum transfer t related by

$$\left(\frac{d\sigma_{\mathrm{el}}}{dt^*}\right)_{t=0} = \frac{\pi}{p^{*2}} \left(\frac{d\sigma_{\mathrm{el}}}{d\Omega^*}\right)_{\theta=0^\circ} \tag{3.14}$$

one obtains

$$\left(\frac{d\sigma_{\mathrm{el}}}{dt^*}\right)_{t=0} = \frac{\sigma_{\mathrm{tot}}^2}{16\pi}(1+\rho^2). \tag{3.15}$$

Replacing the cross sections partly by event rates results in

$$\left(\frac{dN_{\mathrm{el}}}{dt}\right)_{t=0} = (1+\rho^2)\,\sigma_{\mathrm{tot}} \left(\frac{N_{\mathrm{el}} + N_{\mathrm{inel}}}{16\pi}\right) \tag{3.16}$$

thus,

$$\sigma_{\mathrm{tot}} = \left(\frac{dN_{\mathrm{el}}}{dt}\right)_{t=0} \frac{16\pi}{N_{\mathrm{el}} + N_{\mathrm{inel}}} \frac{1}{1+\rho^2}. \tag{3.17}$$

The TOTEM experiment will measure dN_{el}/dt at small t and N_{el} with its so-called Roman Pots, while simultaneously measuring N_{inel} with a forward inelastic detector (also part of TOTEM) and the CMS hadronic forward calorimeter. Using equation (3.17) the total cross section σ_{tot} can be measured with a precision of $\sim 1\%$. Since TOTEM also allows a separate measurement of the elastic and inelastic contributions to the total cross section, one can use the result for the calibration of the methods described above via formula (3.10) to obtain the luminosity.

Since TOTEM will only operate at low luminosity and with different machine optics a systematic uncertainty is introduced in the calibration of the (real-time) methods at design luminosity. Still, using this method the luminosity is expected to be determined with an error smaller than 5% [73].

3.2.9 The Detector Status in Summer 2009

After almost 20 years of design and construction, CMS is ready since fall 2008 to record LHC collisions. Most of the installations and also first tests have been performed at the surface before lowering CMS in parts into the cavern. The lowering of the heavy elements began in November 2006 starting with the forward calorimeters and shortly thereafter by parts of the endcap steel disks and barrel wheels. Piece by piece the other parts followed and by January 2008 the last heavy element (an endcap disc) was lowered. In spring 2008 the beam-pipe was installed and baked out, followed by the insertion of the pixel detector into the previously installed silicon strip detector. In summer 2008 the two ECAL endcaps joined the barrel of the ECAL which is already fully operational since autumn 2007. The forward hadron calorimeter was raised to its final position just before the arrival of the first beam and thus completed the HCAL, one of the first sub-detectors being operational. Finally the solenoid previously tested in 2006 in the surface assembly hall was ramped up to almost 4 T. The only missing sub-detector is the preshower, to be located in front of the ECAL endcaps which is currently being installed.

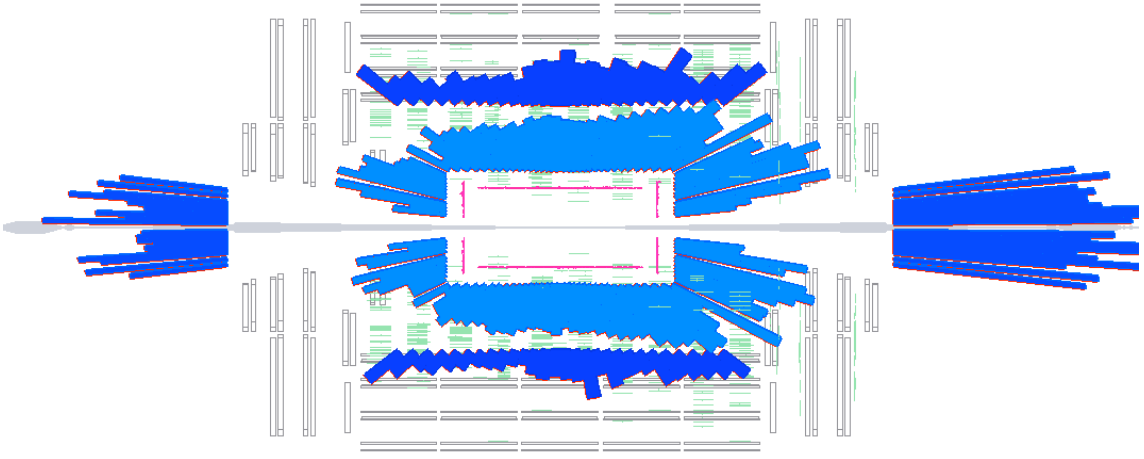

Figure 3.20: *First beam related events within CMS: single shots of one LHC beam onto a collimator placed 150* m *upstream of CMS provided millions of muons. The event display shows the energy deposits in the electromagnetic calorimeter. [79].*

The muon system was already extensively commissioned and integrated with other detector subsystems during the so-called Magnet Test and Cosmic Challenge (MTCC) in 2006. The challenge provided important commissioning and operational experience. It was performed above the ground and involved several cycles of magnet tests including the mapping of the magnetic field. In addition approximately 200 million cosmic muon events were recorded for purposes of calibration, alignment, and detector performance studies (see also section 4.5).

Since spring 2007 every month at least one week has been devoted to global commissioning activities using the installed detectors and electronics in its final layout and location. Subsystem by subsystem joined until summer 2008. Millions of cosmic muon events were taken and fed through the full data acquisition chain, the high level trigger and finally the data were released for analysis in the world-wide LHC computing grid. Upon the start-up

of the LHC in September 2008, the closed CMS detector, including all sub-detectors, has taken almost 300 million cosmic events with magnetic field on and about 30 million cosmic events with field off. All subdetectors have demonstrated that they are operational, including data acquisition, trigger and computing.

First beam related events were recorded in September 2008. Single shots of one LHC beam onto a collimator placed 150 m upstream of CMS provided millions of muons which were used to synchronize the CMS beam monitoring system to the beam timing. With the usage of the beam monitoring system as trigger, CMS took data with all sub-systems except the inner tracker, which was shut down for safety reasons. These "splash events" as shown in Figure 3.20 deposit several hundred TeV of energy within the calorimeter and allowed for example the alignment of the ECAL channels in time with a precision of 1 ns.

With the LHC beam traversing CMS, beam halo events were observed and reconstructed with the help of the CSC chambers until the LHC incident happened.

Chapter 4

The World Wide LHC Computing Grid and CMS

The analysis of data via a model-independent approach would not be possible without the enormous developments in computer science concerning both processing speed, as well as the fast interconnection of computing centres enabling decentralized distributed computing. This chapter introduces the need and the basic principles of grid computing [80–82] and explains its paramount importance to CMS. The CMS tools embedded in such an environment and its application are demonstrated.

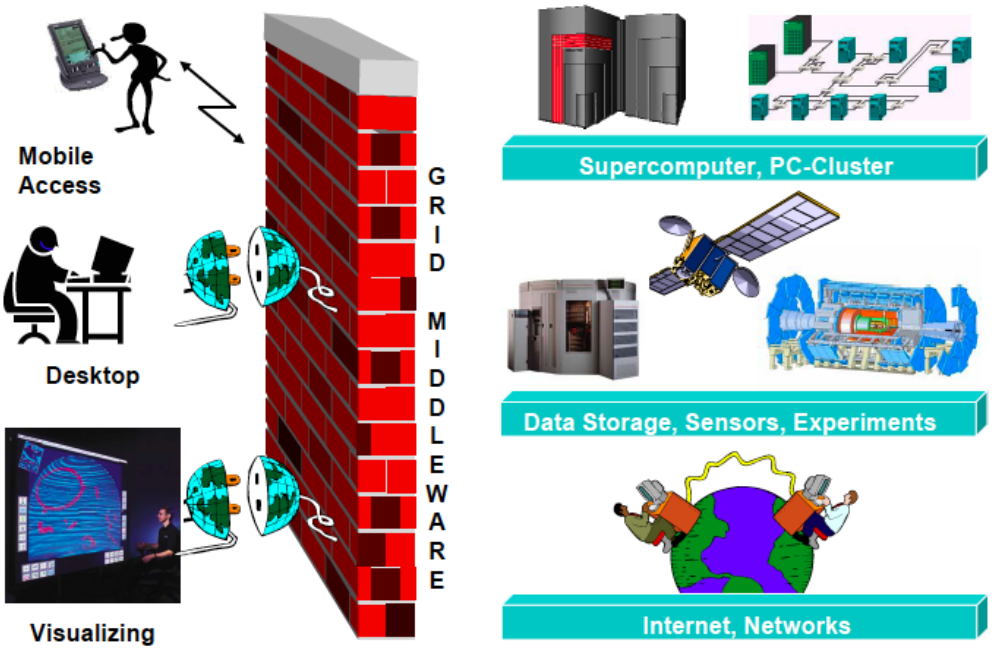

Figure 4.1: *The Grid Vision [83]: Consumption of computing and storage like electric power. Equipped with an internet connection and suitable software, enormous computational, storage and information resources can be accessed.*

4.1 The World-Wide LHC Computing Grid (WLCG)

The Large Hadron Collider and each of its experiments are expected to deliver data of the order of 10 PetaBytes (PB[1]) annually, which will be accessed and analysed by thousands of scientists around the world. Considering the unprecedented amount of computing and storage resources required, it is clear that this cannot be funded at one central place. The LHC experiments adopted the solution of distributed computing, utilizing the storage and computing power of national and regional computing facilities. The goal is to build and maintain a data storage and analysis infrastructure for the entire high energy physics community that will use the LHC. Similar to the evolution of the World-Wide Web in the early 90's, the World-Wide LHC Computing Grid [84] was founded to meet the needs of the physicists. The basic idea of the grid in general, visualized in figure 4.1 is to provide storage, information and computing capacities like the electrical grid delivers power: the end-user should not worry about the internals, but just gets the product. The rapid evolution of wide area networks with its increasing capacity and bandwidth coupled with the decreasing hardware costs make the grid solution realizable and attractive for the LHC use case.

While in the past computing demanding tasks were the dedicated working area of expensive super-computers, the trend in the grid sector is adverse: large amounts of cheap customized hardware with more and more processing cores per central processing unit (CPU) similar but more reliable to those used as desktop personal computers are installed as the working horses of a grid site. With a similar approach many relatively cheap disks are grouped together as one logical unit by tertiar storage systems serving as the main storage of a site. Together with the information systems publishing information about the status of the site and the interfaces provided by the middleware, the storage and computing resources are the main building blocks of a grid site.

In general one can distinguish several types of grids by their focal point:

- The **computing or computational grid** is the prototype of a grid. It allows to share large-scale computing resources within the participating groups. In the late 90s Foster and Kesselman, the godfathers of the grid idea, defined it in a more rigourous way as "a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities" [81].

- **Data grids** focus on providing storage capacities for large amounts of data and their transparent access to the customer.

- **Information or application grids** aim to provide information and data exchange, using well-defined standards and web services or allow application sharing such as in gaming grids. In general there is a trend towards the terminal like access of resources as storage, software and computing power: the end-user has basically a screen and a keyboard connected to all kind of on-demand services via a broad-band internet connection.

[1] 1 PetaByte = 1024 TeraByte; 1 TeraByte = 1024 GigaByte

The common idea of the grids is to share globally distributed resources within so-called virtual organizations (groups of humans and their resources within the grid) and to provide transparent access to information, data, and computing cycles. As such, the grid infrastructure consists of services to access the resources, the so-called middleware, and of the resources itself. In contrast to distributed computing, the grid resources are not centrally controlled, but are maintained and operated by the national and local institutes and universities. Therefore the usage and development of standard, open, general-purpose protocols and interfaces is mandatory. A grid must deliver high quality services and needs the ability to recover from failures e.g. by relocating a job which failed at a certain site. However the general grid infrastructures are generic without any dependencies of the applications/experiments, although the grid used at the LHC has some specialization accommodating the physicist's requirements.

The WLCG is a mixture of a data and computing grid. Therefore it has to deal with a large amount of data as well as it has to provide sufficient computational resources to process the data. Logically and technically one can distinguish the different grids under the hood of the WLCG by the different operational grid organizations and by its middlewares: EGEE (Enabling Grids for E-SciencE [85]) in Europe and OSG (the Open Science Grid [86]) in the United States, but also several national and regional grid structures such as GridPP [87] in the UK, INFN Grid [88] in Italy, and NorduGrid [89] in the Scandinavian region. The WLCG operates a grid distributed over more than 200 sites around the world, with more than 100,000 CPUs and 100 PB of data storage. The status of the grid sites and their utilization can be seen from various monitoring pages such as shown in figure 4.2.

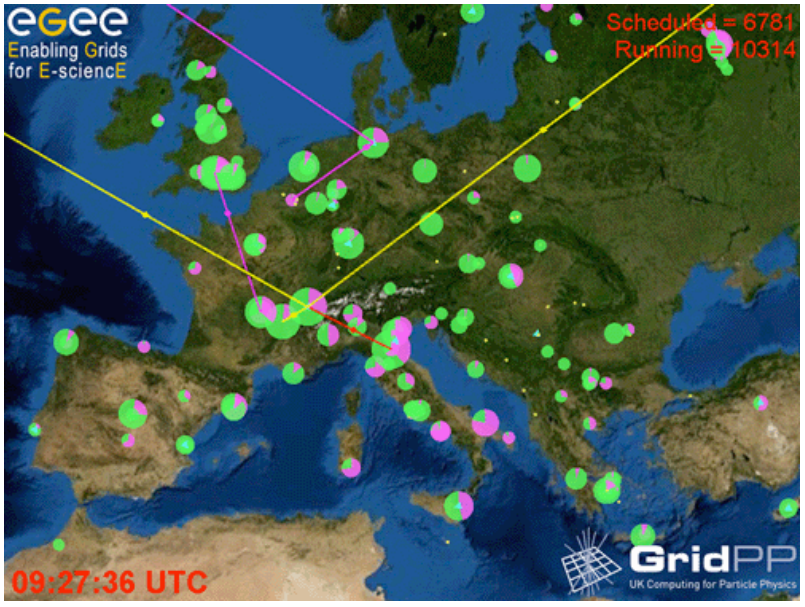

Figure 4.2: *Grid real-time monitoring.*

The advantages of such a distributed concept are:

- Reduction of single points of failure.
- Distribution of costs, operation, and maintenance.

- Data analysis independent of the geographical location.
- Optimal usage of resources.
- Distribution of computing centres and experts among time-zones allows 24/7 monitoring and support.
- Flexible evolution of the global system, easily adoptable to the needs of the LHC and its experiments.
- Adoptable to new technologies that may appear and that offer improved usability, cost effectiveness, or energy efficiency.

Such a huge distributed environment has never been set up before so that there are many challenging tasks to be solved before physicists are able to work reliably with such a system. The requirements for the LHC experiments are broad and can be sumarized as:

- Reliable and automatized placement of large volumes of data around the grid.
- Administering of the storage space at each of the sites.
- Keeping track of the tens of millions of files generated by thousands of physicists as they analyse the data.
- Ensuring adequate network bandwidth: optical links between the major sites, but also good reliable links to the most remote locations.
- Guaranteeing security across a large number of independent sites while minimizing red tape and ensuring easy access by authenticated users.
- Maintaining coherence of software versions installed in various locations.
- Coping with heterogeneous hardware.
- Providing accounting mechanisms so that different groups have fair access, based on their needs and contributions to the infrastructure.

In summary the grid has been deployed to meet the vast resource requirements (not only of the LHC) on a global scale, providing huge amounts of resources to a single user at the price of a certain overhead. It allows to couple local resources at many places without giving up their political independence. Since even the local resources appear as remote for the user, grid computing requires a new view to computing.

4.2 The Physical Grid Building Blocks

The computing centres within the WLCG, which are distributed all around the world, are arranged in a hierarchical structure. They are classified into different so-called Tiers depending on their role within the computing model. This is reflected by the services the site provides, but also their amount of resources in terms of storage, computing power,

and network connectivity. The single Tier-0 located at CERN provides resources for only central and time-critical tasks like raw data processing, archiving, and distribution of data to the Tier-1s. Those centres store a second copy of the raw data and are responsible for the reprocessing of the data with updated calibration and alignment constants and the extraction of a reduced data format (AOD) for analysis purposes. These reduced data are distributed to the Tier-2s, where the user analyses are performed or Monte Carlo events are produced. Finally Tier-3 centres do not have to fulfill central tasks, but provide additional resources for the local physics community. The individual tasks and services which have to be provided by the different Tiers vary from experiment to experiment. The CMS Model is described in section 4.4.

A single site within the WLCG physically consists at least of a computing element, some worker nodes and an attached storage element.

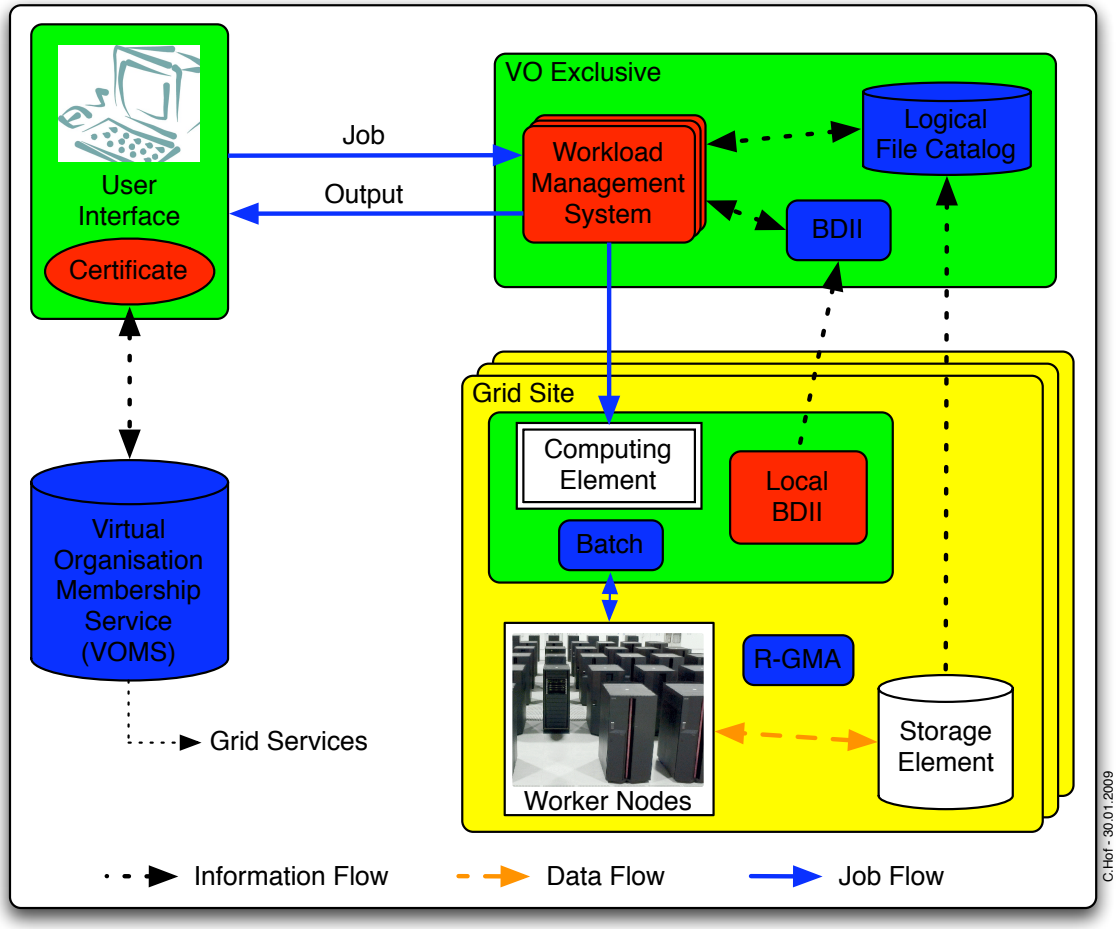

Figure 4.3: *Interplay of the grid building blocks and the middleware services.*

Computing Element

The Computing Element (CE) subsumes the computing resources localized at a grid site. Technically it reduces to one or a set of machines acting as an entry point (grid gateway)

for jobs sent via a Workload Management System. The CE hands over the job to a batch system (local resource management system), which is responsible for the scheduling and the execution of the jobs on the local worker nodes. After the processing of the job the CE returns the output through the Workload Management System back to the user.

Worker Nodes

The worker nodes are the place where finally the job processing is performed. For the grid-users the worker nodes are hidden. The communication occurs only via the workload management system or the computing element. The worker nodes have direct access to the data stored at the site's storage element, allowing for a high input/output rate.

Storage Element

The Storage Element (SE) is the grid storage space associated to a site. It is world-wide visible and provides the grid interfaces (SRM, gridFTP) for interactions such as file listing and replication. In addition it needs to provide authorization mechanisms for the virtual organizations. For Tier-2 sites the storage element consists of disk only storage which is grouped together by tertiar storage systems to appear as a single logical unity. At the Tier-0 and the Tier-1s the disk space is operated as a cache or front-end for the tape-based Mass Storage Systems (MSS) as back-end. Software as the CERN Advanced STORage system (CASTOR [90]) or the FNAL-DESY development dCache [91] provide the possibility of such a taped-backend disk system. At the Tier-2s one can also find DPM (Disk Pool Manager [92]), BestMan [93] or StoRM [94] as grid storage system implementations.

These elements need to be connected to each other and to the user via a set of software packages and services, subsumed as the grid middleware.

4.3 The Logical Grid Building Blocks

The logical layer of software which connects all elements is the so-called middleware. It is grid-specific and the description here is restricted to the gLite [95] incarnation. The middleware implements the grid services and client software, while trying to hide much of the complexity of this environment from the user, giving the impression that all of these resources are available in a coherent virtual computer centre. The following sections describe the middleware components relevant for an end-user and their relation as sketched in figure 4.3.

Virtual Organizations

A virtual organization (VO) is a dynamic collection of individuals, institutions, and resources which is defined by certain sharing rules. In that sense a VO might represent an experiment collaboration as in the case of the WLCG. A single user asks for a grid certificate through a Certification Authority (CA) which issues a personal grid certificate (X.509 certificate). With this certificate a single user can request the membership to a

certain virtual organization like CMS. This certificate is then the key (authentication and authorization) to all resources belonging to the virtual organization. For security reasons so-called proxy certificates, which are temporary copies of the certificate with a limited life-time of typically some hours or days, are delegated across the grid. For example they can be attached to the grid job for authorization and authentication. Following the grid principles all users within a certain virtual organization are equal and share the resources on a fair basis. However, authorized users may equip themselves with different roles within a VO such as software manager or Monte Carlo production operator. Also VO sub-groups are supported, which for example allow users affiliated with a German university or laboratory (VO: cms; subgroup: dcms) to obtain higher priorities for processing at the German grid sites.

The User Interface

The access point to the WLCG grid is the User Interface (UI). This can be any machine where the appropriate software and the user's certificate is installed. It can be compared to the web browser as an interface to the world-wide web, although the UI for the WLCG is still at a level where most interaction is performed via command line tools instead of a graphical user interface. The UI provides access to the functionalities offered by the information, workload and data management systems, such as:

- Discovery of all resources suitable for the execution of a given job.
- Job submission and cancelation.
- Status checks for submitted jobs.
- Output retrieval for finished jobs.
- Access to logging and bookkeeping information of jobs for debugging purposes.
- Copy, replication, and deletion of files from/to the grid storage elements.
- Retrieval of the status of different resources from the information systems.

The Information System

The information system is a critical part of the grid infrastructure. It allows users and services to discover which resources and services are available within the grid or at a certain site. The precision and up-to-dateness of the information determine the quality of service of the whole grid.

At a grid site the computing and the storage elements are equipped with so-called information provider software, which generate data about the resource (e.g. general availability/status, free/used storage space/batch slots). The data of the different information providers are aggregated by a local/site-level BDII (Berkeley Database Information Index). This database stores and publishes the data. Finally a top-level BDII polls the data from all available sites within the specific grid. Effectively the top-level BDII defines a view

of the overall grid resources and serves for example as an input source for the workload management systems.

A different source of information is the R-GMA (Relation-Grid Monitoring Architecture). While the Berkeley Database Information Index (BDII) is an LDAP-based[2] information system, R-GMA provides data as a global distributed relational database. R-GMA is currently used for accounting and both system- and user-level monitoring.

The Workload Management System

The Workload Management System (WMS) acts as job distributor and load balancer within the grid. Its task is to accept jobs and to assign them to the most appropriate computing element. The WMS regularly checks the status of the jobs and retrieves the output upon the end of each job. By calls to the WMS via the user interface the user can get information about the jobs.

The user can specify certain requirements within the jobs, such as the operating system, the closest storage element, needed input files, or time requirements. Upon the submission of a job into the grid it is handed over to one of the independent WMSes of the VO. Among all available computing elements, which fulfill the requirements expressed by the user, the WMS passes the job to the CE with the best ranking. The ranking is based on quantities derived from the CE status information expressing the quality of the CE (typically a function of the numbers of running and queued jobs). In addition to the submission of single jobs the latest implementation of the WMS allows to submit a collection of jobs in bulk. This allows for a much more efficient job submission and improves the limit of jobs/day hit within the CMS Computing, Software, and Analysis Challenge 2008 [96].

Monitoring and User Support

A key component of every evolving and still error-prone system is a detailed and consequent monitoring. Apart from the site and experiment specific monitoring which is described in section 4.4.4, central WLCG/EGEE control the basic functionality of all grid sites by e.g. submitting test jobs. Only sites which pass these so-called Site Availability Monitoring (SAM) tests [97], are visible in the top-level BDII and thus are available for the users. These tests do not only spot problems, but equip the grid with a robustness against failures: unstable sites are flagged and the jobs are routed to more reliable clusters.

The Global Grid User Support (GGUS) [98] provides centralised support for WLCG sites and users. The service consists of a ticket system for an efficient solution of problems by the direct involvement of grid site administrators and grid experts. In addition known bugs are tracked, lists of frequently asked questions and documentation are maintained. The GGUS portal is supposed to be the key entry point for grid users looking for help.

[2] Lightweight Directory Access Protocol.

4.4 The CMS Computing Model

The CMS distributed computing and analysis model is well-integrated within the World-Wide LHC Computing Grid. The model is designed to serve, process, and archive the large amount of data taken with the CMS detector. Therefore CMS uses a number of event data formats, starting from the detector data to successive degree of processing that refine this data (see table 4.1).

Event Format	Content	Purpose	Event Size	Events per year	Volume per year
RAW	Detector data, L1 + HLT info	Input to Tier-0, to be archived	1.5 MB	$3.3 \cdot 10^9$ (2 copies)	5.0 PB
RECO	Reconstructed physics objects (e, μ, jets, ...) + hits/cluster	Output of Tier-0 reconstruction/ re-reconstruction at Tier-1	250 kB – 500 kB	$8.3 \cdot 10^9$ (reprocessing)	2.1 PB
Analysis Object Data (AOD)	Reconstructed physics objects, some hit info	Physics Analysis	50 kB – 100 kB	$53 \cdot 10^9$ (copies at all Tier-1)	2.6 PB
SIM	Generator info, simulated detector data	Physics Analysis	2MB	1:1	5 PB

Table 4.1: *Overview of the CMS data formats and its sizes as well as its expected amount per year in terms of size and numbers [99]. In total the grid machinery has to deal with more than 15 PB per year once CMS is running at $\mathcal{L} = 2 \cdot 10^{33}$ cm^{-2} s^{-1}. RAW data are stored at the Tier-0/1, (re-)reconstructed to RECO at the Tier-1 and distributed to the Tier-2s in an AOD format.*

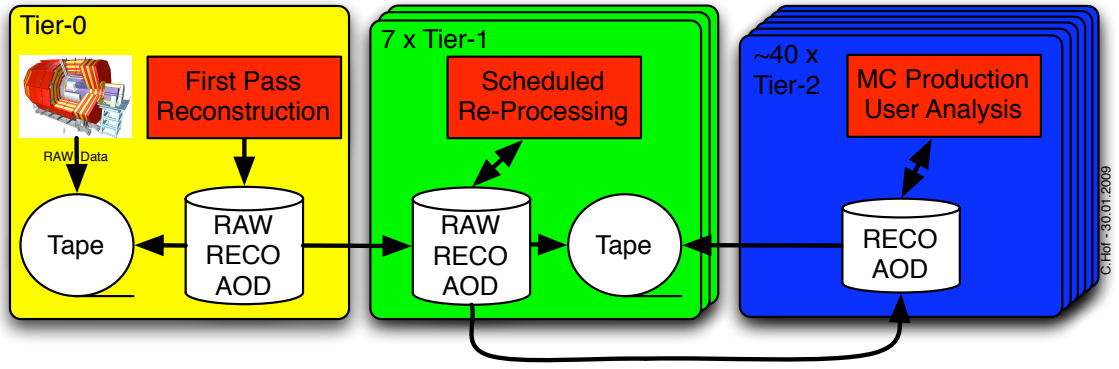

Figure 4.4: *CMS Data- and Workflow: The path of the data on its way from the detector (Tier-0) to the end-user at the Tier-2.*

A multi-Tier hierarchical distributed model (see figure 4.4) is adopted for serving and archiving of the raw and reconstructed data. The Tier-0 centre at CERN accepts data from the CMS online system, archives the data, performs prompt first-pass reconstruction, and distributes raw and processed data to Tier-1s. The Tier-1s are typically large regional

Requirement	Tier-0	Tier-1	Tier-2
Network [MBit/s]	6000	10000	2000
CPU [kSi2K]	4600	2500	900
Disk Storage [TByte]	400	1200	200
Tape Storage [PByte]	5	3	-

Table 4.2: *Nominal resource requirements for the different level of Tiers according to the Computing Technical Design Report [99]. The numbers assume a luminosity of* $\mathcal{L} = 2 \cdot 10^{33}\ \mathrm{cm}^{-2}\,\mathrm{s}^{-1}$.

computing centres like the German Tier-1 GridKa in Karlsruhe. Their tasks are to store a second copy of the data on tape and to provide services for scheduled data processing operations (reconstruction, calibration, skimming) and other data-intensive analysis tasks. Finally the Tier-2 centres, each consisting of one or several collaborating computing facilities, provide capacity for analysis, calibration activities, and Monte Carlo simulation. Individual scientists will access these facilities through Tier-2/3 computing resources, which can consist of local clusters in a university department or even individual PCs. Corresponding to their tasks the different Tiers have to meet certain resource requirements for CMS (see table 4.2).

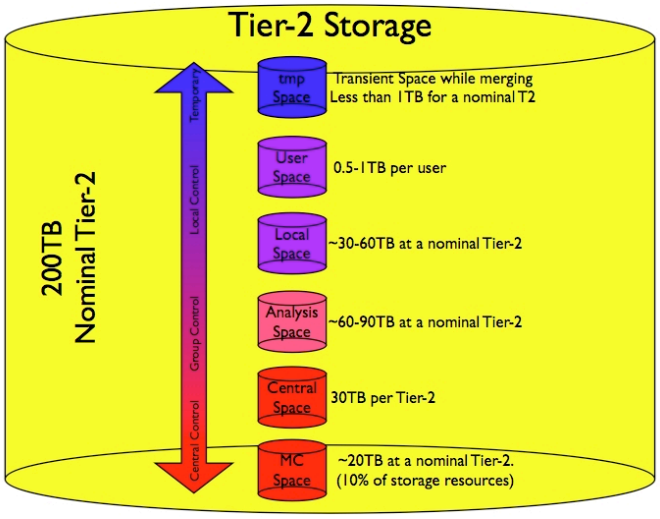

Figure 4.5: *Storage layout at a Tier-2. Resources have to be provided for each hosted user, for each hosted CMS group, but also for the central Monte Carlo production.*

4.4.1 The Data Management System

The CMS data management is based on a set of loosely coupled components which allow physicists to discover, access, and transfer event data. The typical workflow and the involved components are illustrated in figure 4.6 and discussed in the following sections in detail.

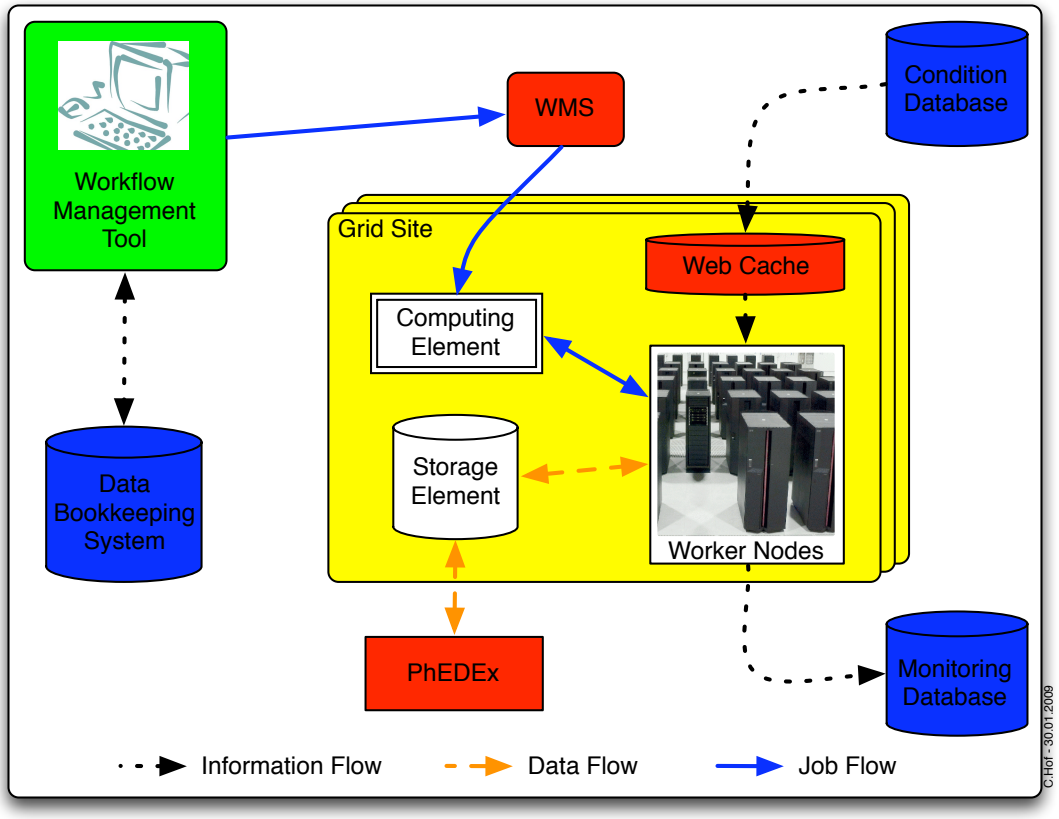

Figure 4.6: *Overview of the CMS grid workflow: The user interface provides access to the grid world wrapped by the CMS workflow management tools such as CRAB for user analysis and ProdAgent for official Monte Carlo production. The CMS tools explore the available datasets and its location within the Dataset Bookkeeping System (DBS) and send the jobs through the Grid Workload Management System (WMS) to the site holding the data. At the site the job is handed from the Computing Element (CE) to the next free worker node, which accesses data stored on the associated Storage Element (SE). Condition data are retrieved from a central condition database which is cached through a web cache at the site. To allow constant monitoring, the CMS jobs report their state on a regular basis to a central monitoring database. Once the job has finished, the output might be stored on a local or remote storage element or can be retrieved together with the log files at the user interface. In addition the processed files might be registered in DBS and CMS transfer tool (PhEDEx) for further processing or distribution.*

CMS Catalogues

The CMS catalogue used to define and discover the data and Monte Carlo simulated samples is the central Dataset Bookkeeping System (DBS) [100]. The DBS maintains the semantic information associated to the datasets such as which files belong to which dataset, their grouping into blocks, but also stores detailed meta-information about the files itself (type, size, checksums, content). It keeps track of the data parentage through their processing history and allows to discover which data exist. In addition it maps the file-blocks to sites holding a replica of them and allows to find the location of desired data. It is synchronized with the CMS data replacement and transfer tool PhEDEx [101, 102].

DBS supports the existence of local and global instances, for private or intermediate and public data, and allows the migration of data between them. DBS is implemented as a Tomcat-based Java server with an Oracle or MySQL back-end. A web interface called the DBS Discovery Page as a front-end to DBS allows to explore the datasets which are available within CMS. In a user-friendly manner it provides access to all the meta-information related to a dataset and gives handles to a simplified access of the data itself. It allows for example the download of predefined configuration files for the CMS framework or the CMS tool for data analysis within the grid (CRAB).

Local Data Access

For the simplified handling of files, the central databases store and deal only with logical file names. In order to access the files at the sites, e.g. through an analysis grid job, the logical file name has to be resolved into a physical file name such as a path to a local disk or a mass storage system like CASTOR or dCache. For this purpose each site maintains an XML[3]-based file containing simple, generalized rules to build physical paths from logical names and vice versa. The rules may depend on the desired access protocol and provide a fine-grained handle for the data organization to the site administrator. A common Tier-1 use case which is covered in that way, is the separation of data: files that should go to tape and data which should stay on disk only.

Data Placement and Transfer System

The Physics Experiment Data Export (PhEDEx) project [101, 102] manages the transfers of data among sites, dealing with grid File Transfer Services [103] and different storage systems. PhEDEx interacts with the CMS catalogues, cross-checks the file-level information in DBS for datasets mentioned in transfer requests, and updates the storage location when the data transfers are complete. Technically it is based on software agents that run autonomously at each site and exchange information via a central database. PhEDEx has been exercised in progressively increasing complexity and scale during several years of use in daily production and computing challenges. In the last year[4] 30 PB have been transferred with PhEDEx. In April 2008 the average global daily transfer rate was ∼180 TB/day or 2.0 GB/s with currently around 70 sites involved.

Handling of Calibration and Alignment Data

For the delivery of condition data to a world-wide community of distributed processing and analysis clients, CMS uses a multi-tiered web approach well-suited to the grid environment. Condition data include calibration, alignment, and configuration information used for online and offline event data processing. The conditions, which are stored in a central Oracle database, are keyed by time and have a limited validity. Since these data might be used by many thousand jobs in parallel all around the world, the caching of such

[3] XML (Extensible Markup Language) is a general-purpose specification for the creation of custom markup languages.

[4] March 2008 – 2009

information close to the processing activity results in a significant performance gain. CMS has adopted the solution of web proxies or caching servers (Squid) which are heavily used within the WWW since years and thus are readily deployable, highly reliable, and easily maintainable.
Each site deploys one or more squid caches which provide high performance access to the condition data requested by the jobs through the CMS framework and its interface FroNTier [104]. FroNTier is a simple web service approach providing client HTTP access to a central database service. The cache is loaded on demand and manages itself automatically. A lost or corrupted cache is simply repopulated with little or no intervention required. Several features have been developed to make the system meet the needs of CMS including careful attention to cache coherency with the central database, and low latency loading required for the operation of the online High Level Trigger.

4.4.2 The CMS Workload Management System

The CMS workflow management system manages large-scale data processing which is the principal focus of HEP computing. An example of distributed processing workflow that illustrates the interactions with data management components and the grid middleware is shown in figure 4.6. The basic steps are:

- Data discovery and location via DBS.
- Job submission to the site where the data are located.
- Handling of the output data stored on local storage or passed to the transfer system (PhEDEx).
- Publication of the produced data with the relevant provenance information in DBS.

Monte Carlo Production

CMS has a long-term need to perform large-scale Monte Carlo simulations. In addition it provides a way for testing the tools and infrastructure needed to process large amounts of events that will be available at detector startup. The MC production system consists of three components: ProdRequest, ProdManagers and ProdAgents. The request system (ProdRequest) acts as a front-end application for production request submissions into the production system. The production manager (ProdManager) manages these requests, performing accounting and allocating work to a collection of production agents (ProdAgents). The ProdAgent consists of a set of loosely coupled components executing production workflows in the grid environment. ProdAgents are responsible for job submission, job tracking, error handling, and automatic resubmissions, as well as data merging, and publication into the CMS cataloguing and data transfer system.
A production scale of more than 30.000 jobs per day and per ProdAgent has been achieved. By running several ProdAgents in parallel by 2–4 operations teams a production yield of a billion events per year, with a job efficiency of about 80%, is routinely reached [105]. More than 40 sites have been used for production with high job efficiency. The performance of

the production system is greatly affected by the unreliability and instability of global grid services and sites (local storage and batch systems).

Tier-0 Workflow

The CMS Tier-0 is responsible for all data handling and processing of real data events in the first period of their life. Data written by the data acquisition system in specific format (streamer files) are automatically transferred to the Tier-0 site. At the Tier-0 the repacking of the streamer files occurs, converting them into raw data and splitting into physics primary datasets. The output RAW data are archived on tape. Prompt reconstruction reads the RAW data and stages out the reconstructed data. These workflows are managed by ProdAgent instances and its evolution into a much wider system to support Tier-0 activities. Files are registered in DBS. The RAW and reconstructed data are transferred from the Tier-0 to dedicated Tier-1 sites via PhEDEx. Experience using the system is being gained with initial detector commissioning activities (monthly global data taking) as well as dedicated stress tests at nominal data rates.

4.4.3 User Analysis

A tool, CRAB (CMS Remote Analysis Builder) [106], has been developed to provide a user friendly interface for CMS physicists' interactions with data management and grid submissions. CRAB supports the direct submission to the grid, but also the submission with a CRAB server that aims at improving further automation and scalability of the whole system. CRAB has been used to analyse data during the past CMS challenges: studies of the CMS physics discovery potential based on MC simulation, analysis of Magnet Test & Cosmic Challenge data, and many other activities.

4.4.4 Monitoring

A key component of the grid is the monitoring. It allows the system to react on failures and enables site managers to check the health of the site and allows to detect the cause of a failure quickly. But it also provides valuable input for the users about the reliability of the resources to use.

The Experiment Dashboard

The CMS Dashboard project [107] aims to provide a single entry point to the monitoring data collected from the CMS grid environment and the jobs executed within this distributed system. By the inclusion of experiment-specific information (via MonALISA [108]) in addition to R-GMA data, Dashboard is able to display quantitative and qualitative characteristics of the experiment and is thus able to indicate problems of any nature. General monitored quantities are: how many jobs are running, pending, accomplished successfully or failed on a per user, per site, per input data collection basis. For an example see figure 4.7. Also the distributions evolving with time are available. Further resource usage (CPU, memory consumption, input/output rates) are aggregated. A detailed analysis of the job behaviour (success rate, reasons of failures as a function of time, execution centre,

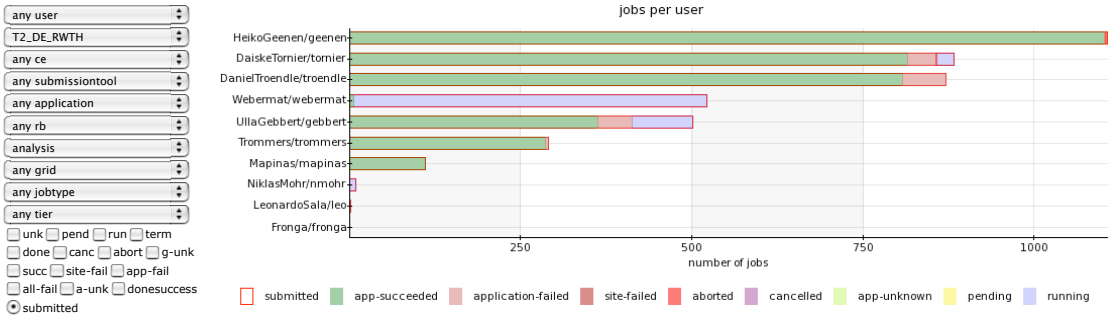

Figure 4.7: *The Dashboard Monitoring [107] provides a real-time monitoring for user and production jobs within CMS. With its detailed output it is even possible to debug causes of failures concerning CMS software or site problems.*

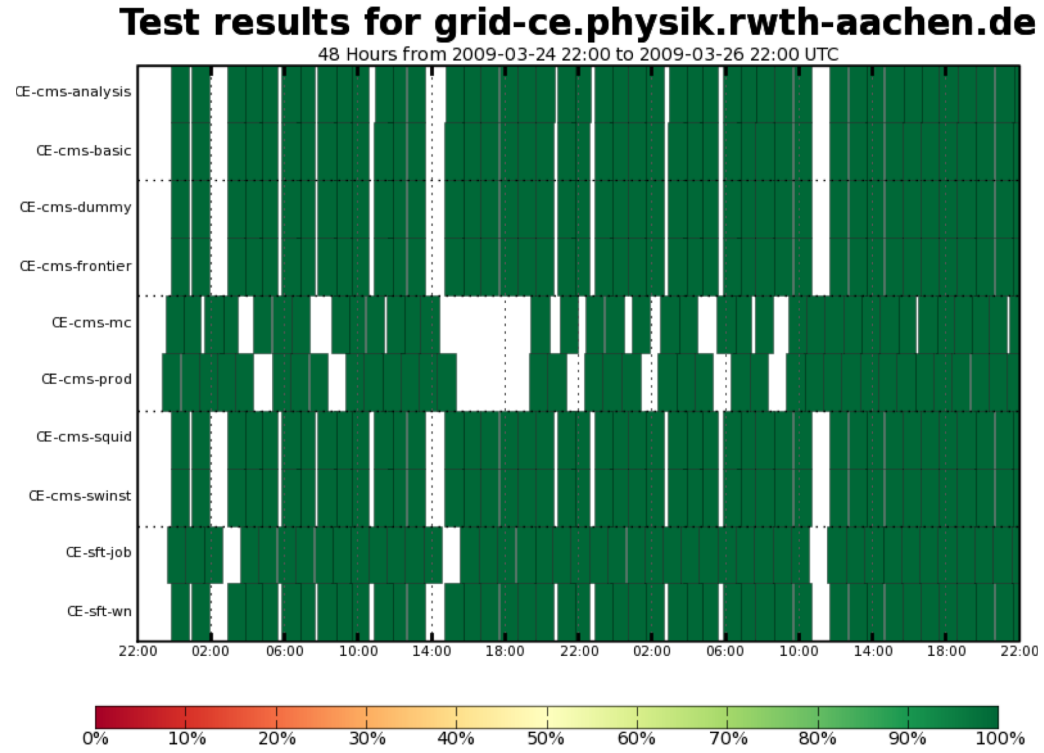

Figure 4.8: *The site availability monitoring enables the sites to follow their status. Only if all tests succeed a site is available for user jobs and MC production.*

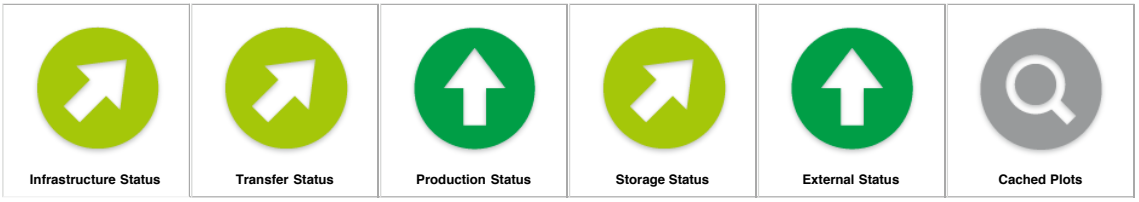

Figure 4.9: *HappyFace is a monitoring tool which aggregates the monitoring results of other tools. It allows a quick overview of the site status with respect to its productivity and helps to quickly spot the source of a failure.*

data collection) is possible and provides valuable feedback to the user to detect and identify the problem.

SAM – Site Availability Monitoring

The Dashboard includes the collection of Site Availability Monitoring (SAM) plots (see figure 4.8). SAM subsumes a collection of tests which check the basic functionality in terms of the CMS needs. These dedicated jobs, which run roughly every hour, imitate analysis, production, or software installation jobs accessing computing and storage resources as well as CMS specific services such as FroNTier or the local CMS catalogues. Only sites which pass these tests on a regular basis are available for the usage within CMS.

Other Monitoring Tools

There are various other monitoring tools on the market. Some, which are developed and used within the CMS and RWTH university context, are:

- The CMS JobRobot:
  The JobRobot [109] is an automated tool for the submission of fake analysis jobs using CRAB. It is used as a commissioning tool to test if a site is capable to run certain CMS workflows at the required scale. Currently around 300 of such fake jobs of the length of an hour are sent to each CMS site every day.

- CMS Site Status Board:
  The site status board [110] is a meta monitoring system which conflates the information from the various specific CMS monitoring tools. Within one view all relevant monitoring information are available including their evolvement with time.

- Happy Face:
  The Happy Face Project [111] was founded at the German Tier-1 and is currently part of the Helmholtz Alliance project. As a meta monitoring system it accesses existing monitoring sources and creates the simplified overview of a grid site and its services as displayed in figure 4.9.

4.4.5 Computing, Software, and Analysis Challenges

The CMS progress towards a full implementation for handling organized processing and analysis workflows in its distributed environment for the coming critical first year of LHC data taking is reviewed in a series of large-scale tests. Starting in 2004 (Data Challenge 04, DC04) with a 5% level of the expected requirements of the first year of data taking, the challenges gain in sophistication and scope with a 25%, 50%, and 100% test in 2006, 2007, and 2008, respectively, during the so-called Computing, Software, and Analysis Challenges.

The last CMS challenge in 2008 has tested the full scope of the offline data handling and analysis activities as expected for the CMS data taking during the first year of LHC operations. It was embedded in the Common Computing Readiness Challenge (CCRC08)

as a multi-VO computing stress test to emulate the simultaneous usage of the grid and its resources by all LHC experiments.

The challenge demonstrated the readiness of CMS for the LHC start-up and proved the functionality of the full reconstruction, calibration and analysis chain from the arrival of data at the Tier-0 up to the final analysis at the Tier-2s. In general the expectations concerning the stability and performance of the grid sites were excelled. Many lessons were learned and areas which need further optimization and development have been identified.

In 2009 another readiness challenge with several LHC experiments participating in parallel will be carried out to measure the progress in the system optimization and tuning. The so-called STEP'09 (Scale Testing for the Experimental Programme 2009) will focus on the long-term stability of the Tier operation and will concentrate on tape recall and event processing.

4.5 The RWTH Aachen Tier-2/3

In Aachen the first grid cluster prototype was already installed and operated within the LHC Computing Grid in the year 2004. With the gain of experience the cluster further developed and participated successfully in several of the CMS and WLCG challenges. In 2008 the cluster has been replaced by a brand new modern system and moved to the IT centre of the RWTH Aachen University. Currently[5] it holds in total 17 enclosures equipped with in total 261 blades with 8 cores and 16 GB RAM each. As an official CMS Tier-2[6] and Tier-3, it provides a storage and computing capacity of about 0.5 PetaByte and 2000 cores equal to 2200 kSPECINT2000[7]. Physically and technically there is no distinction between the Tier-2 and Tier-3 resources. However local users are granted higher priorities within the batch system and additional dCache storage space. For rapid installation and deployment all machines are configured with the Quattor toolkit [112]. The monitoring of the whole grid cluster and its operation status is done with Lemon [113].

Apart from the basic grid infrastructure as described in section 4.2 it holds specific CMS services: PhEDEx for data placement, a local DBS instance for the publication of private datasets, a CRAB Server for faster CMS job submission and also ProdAgents suitable for performing large-scale MC productions not covered by official CMS MC requests.

The successful interplay between computing and analysis in Aachen has been a long-standing tradition which is exemplarily reflected by the fact that the first full CMS grid-enabled analysis chain with cosmic data has been performed at the RWTH [114]. Cosmic data taken during the Magnet Test and Cosmic Challenge [115] in 2006 were transferred around the world using PhEDEx and placed into the CMS databases in order to be able to analyse them via grid jobs utilizing CRAB. In parallel a large statistic cosmic Monte Carlo dataset [116] has been produced using the full CMS detector simulation in order to compare the data with the expectation. Figure 4.10 shows the final result of this exercise: an agreement within the expected uncertainties of the cosmic muon p_{T} distribution measured by the CMS detector with the MC.

[5] Status as of March 2009.
[6] In Federation with DESY.
[7] A modern Intel Pentium IV processor with a 2.8 GHz CPU corresponds to about 1 kSI2k.

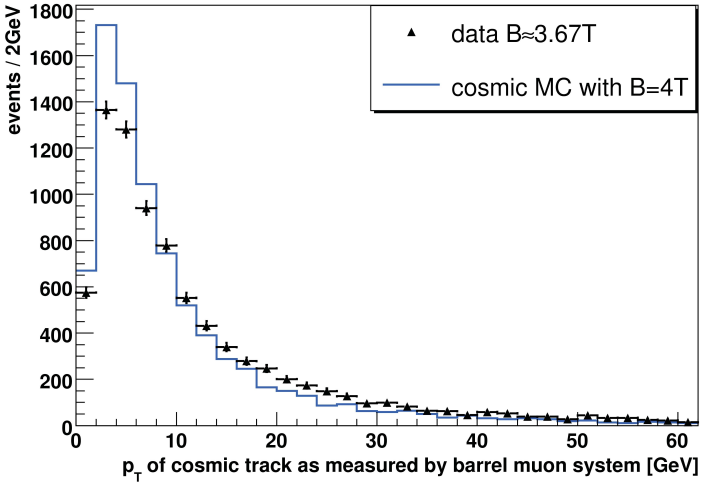

Figure 4.10: *Successful interplay between computing and analysis groups: For the first time the full offline analysis chain using cosmic muon data taken during the Magnet Test and Cosmic Challenge (MTCC) compared to a cosmic simulation has been demonstrated utilizing the LHC Computing Grid. The plot presents the comparison of the cosmic muon momentum with the MC expectation showing a good agreement.*

4.6 The German National Analysis Facility

The National Analysis Facility (NAF) [117] has been founded to enable German physicists to perform successful and internationally competitive analyses of the wealth of data expected from the LHC. The facility is designed complementary to the resources available at the German Tier-1 Centre GridKa and the federated Tier-2 operated by DESY and the RWTH Aachen University by providing an efficient infrastructure for end-user data analysis. The installation largely enhances the capability of the German groups from the ATLAS, CMS, and LHCb experiments as well as the ILC[8] group for their collaborative analysis efforts.

It tries to overcome the caveats of the tiered model:

- The usage of the grid technology adds an additional layer of complexity. While this is not a problem in large-scale productions, it might affect the user's productivity.
- Even the vast amount of grid computing resources might become scarce when the first LHC data are recorded.
- The event processing rate of many analysis jobs is limited by the data reading speed. However, typical grid sites are optimised for CPU-intensive tasks and overall bandwidth to the data. Thus individual analysis jobs might not get the highest possible bandwidth.

[8] International Linear Collider.

- The resources for the end-user analysis and their usage are not well-defined in the experiments' computing models. New classification methods and statistical tools may require rather large computing resources even for the last analysis step.

The NAF was initiated in 2006 as a sub-project of the Helmholtz-Terascale Alliance [118]. In 2007 a prototype installation has been setup at DESY with a close connection to the already existing infrastructure of the Tier-2. Since then it is operated successfully as a joint venture between the German particle physics groups and the DESY IT department. The facility forms the nucleus for the envisaged distributed facility and will provide valuable operational experience. It should trigger the development of collaborative tools supporting analysis by working groups and individual users.

The German groups of the ATLAS, CMS, and LHCb experiments have specified their requests for such a facility. The main requirements are:

- Additional storage resources to house data sets which are of interest for the German groups.
- Additional exclusive grid CPU resources.
- A batch farm for computing intensive end-user analyses.
- A storage system for input/output-intensive jobs.
- An interactive cluster for fast analysis of large data sets.

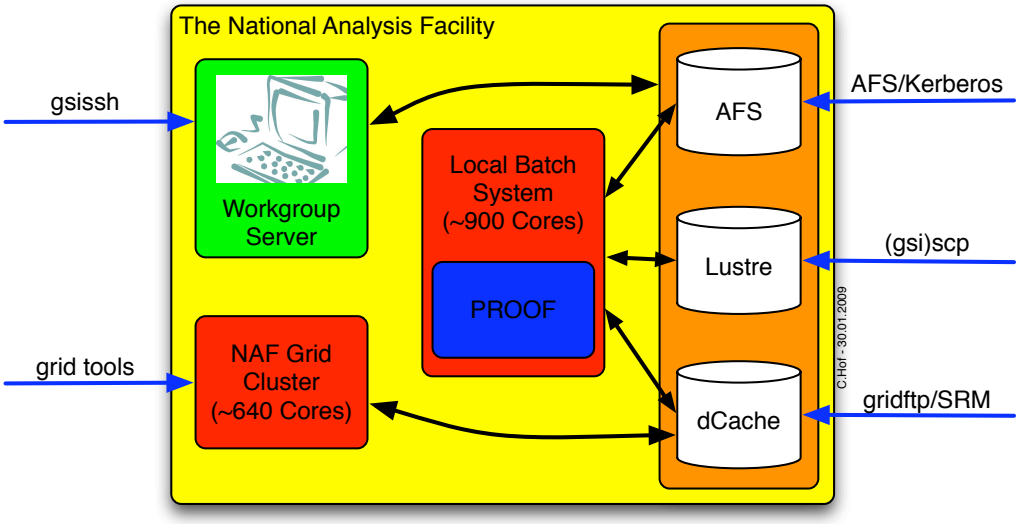

Figure 4.11: *Design of and access to the National Analysis Facility (NAF). Beside the usual grid approach jobs can also be submitted via an interactive local batch system. Data might be accessed/stored on different storage architectures: AFS for world-wide shared space, Lustre for input/output intensive tasks and dCache for large-scale storage.*

The current layout of the NAF is schematically visualized in figure 4.11. In addition to the resources at the Tiers which are mainly accessible via grid tools, a predominant

part of the NAF is designed for interactive usage. The total required resources have been estimated to be equivalent to 1.5 of an average Tier-2 centre, with a special focus on data storage. Registered users from German institutes can access the NAF authorized by their grid certificate through a workgroup server. From here one has the freedom to either submit jobs directly to the local batch system and thus have immediate control over the jobs or to send jobs via grid tools. A special VO subgroup for German users has been defined, allowing the prioritization of German users. For each user the workgroup server provides a home directory which is world-wide accessible via AFS[9]. In addition to the AFS space, which serves for the storage of small files, dedicated dCache storage resources are granted to the NAF users. The dCache storage space (∼500 TB) is aimed for holding additional datasets or dedicated skims useful for the German groups working at LHC experiments. Since a typical high energy physics analysis job tends to be input/output (I/O) dominated, i.e. the limiting factor of an analysis is not the CPU power, but the time needed to read (and write) the data from (and to) disk, a dedicated parallel cluster file system (Lustre [119]) of ∼100 TB has been set up. It suits for I/O-intensive and "burst-mode" like analysis in an interactive way like it is possible with the PROOF [120] toolkit.

Conclusion

Grid computing has already become a key technology for high energy physics: the LHC and its experiments heavily rely on the grid. The enormous improvements in the grid components and within the experiment specific tools proof already now, just in time for the start-up of the LHC, that the system is able to cope with the unprecedented amount of expected physics data.

The grid idea requires a new way of thinking. As the site administrators need to operate their system in a global context, the users must think globally when utilizing the grid. This globalization supports a single user with a huge amount of distributed computing, storage and network resources, but it also requires to learn some new tools, a different attitude towards computing, and finally also some discipline in the resource usage.

[9] The Andrew File System is a distributed file system.

Chapter 5

The Concept of the Model-Independent Search MUSiC

This chapter introduces the basic principles of model-independent analyses, especially its realization within CMS: MUSiC[1] – the Model Unspecific Search in CMS. The first part discusses the underlying ideas and the purpose of such a generic approach. It will motivate the benefits of such a search at the start-up of a collider experiment and point to its long-term goals. The general concept and the underlying algorithms are sketched in the following. The chapter will conclude with the scope of MUSiC, its timeline and an overview of its technical implementation.

5.1 Motivation

As outlined in the previous chapters the LHC will enter an unknown territory. There are multiple reasons why new physics is expected to appear. Unlike in experiments of the past there is an almost infinite number of predictions from theory of how exactly these new physics models will look like. Following the saying "expect the unexpected", a model independent search tries to cover a wide range of the phase space and is not limited to a specific topology. In this way it should be sensitive to surprises with spectacular final states as they might arise from mini black holes. MUSiC might reveal a consistent picture of the various channels where a possible supersymmetric signal contributes. Even without contributions from physics beyond the Standard Model, the tool will help to quickly discover discrepancies caused by detector effects or effects not properly accounted for in the Monte Carlo simulation. Especially at the start-up MUSiC will help to "commission" the physics objects and their reconstruction.

5.2 MUSiC – Principle and Guidelines

The concept of the model independent search stands in contrast to the traditional signal-driven analyses and searches for new physics. Inspired by a certain theoretical model,

[1] This analysis has been developed in close cooperation with [121].

traditional searches are highly tuned to find the optimal final variables which ensure the best separation of signal versus background. Instead, the MUSiC analysis has no optimization with respect to a certain signal. Requiring a solid object identification, all events full-filling well-understood basic triggers are classified and investigated for deviations of the data from the Standard Model. The optimization concept of the traditional searches with a clear focus on a certain signal, is substituted for MUSiC by the following guidelines:

- Model independence: no optimization of selection cuts with respect to a certain expected signal.
- Robustness: focus on well-understood physics objects, i.e. high p_{T}, central $|\eta|$ and solid object identification.
- Simplicity: the steps of the algorithm should be easy to follow, preferring standard statistical estimators and methods.
- Completeness: include any possible systematic differences between data and Monte Carlo prediction in the search algorithm.
- Allow for new physics that contributes predominantly to a single channel (resonances like W') and physics that produces deviations in numerous final states (SUSY).

During the development of the analysis these points lead to the decision which way to head.

5.3 The Analysis Methodology

The workflow of the model-independent approach is illustrated in Figure 5.1. Where a traditional analysis only investigates events of a dedicated topology, MUSiC classifies each event according to its topology. The usual optimization step to enrich the signal under investigation over the background within a final variable is completely missing. Instead a limited set of variables which are expected to be sensitive to physics beyond the Standard Model are inspected within each topology. The last step of both analysis strategies is to perform a significance test to actually quantify the degree of deviation between the data and the Standard Model expectation. Roughly speaking the model-independent approach is a multitude of traditional analyses in parallel without any optimization with respect to a signal. Obviously the challenges show up within the details. As the computing amount for a single signal-driven search is already quite significant, the requirements for MUSiC exceed this by two orders of magnitude. This explains why the approach is quite modern: such cheap and widely-accessible computing resources are only available within the last decade with large computing centres or the grid approach adopted by the LHC experiments (for details see chapter 4).

Also the implementation of systematic uncertainties in a precise and at the same time generic manner is difficult and sometimes even not possible due to limited resources. Therefore it is clear that such a model-independent search has to rely more on Monte Carlo predictions than other signal-driven searches which might rely on specialized background and uncertainty estimations from data.

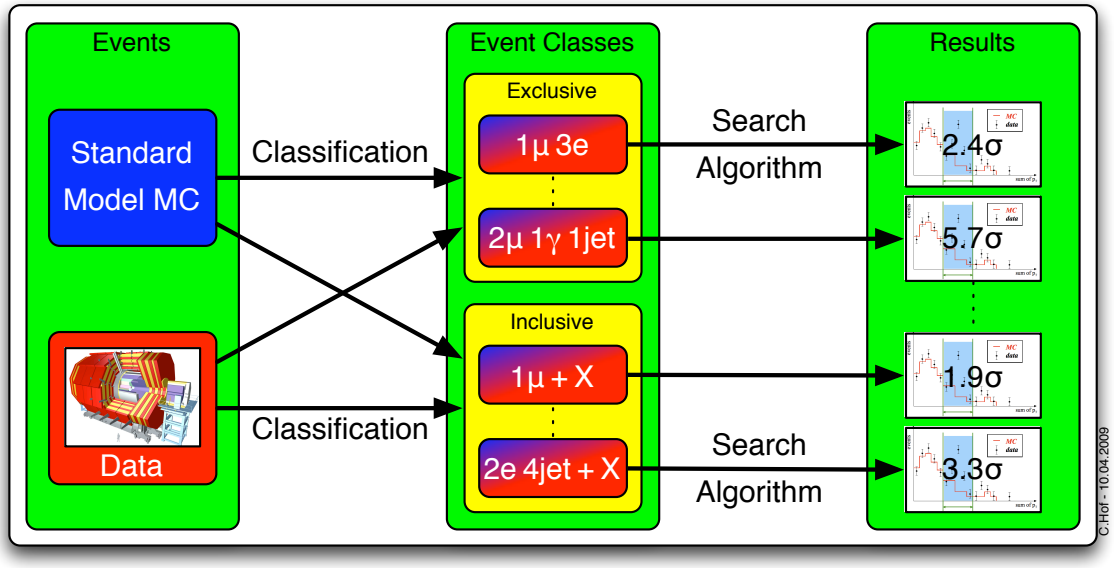

Figure 5.1: *The MUSiC analysis workflow: Events from the CMS detector and from the full detector simulation are classified by their topology into so called event classes. One distinguishes exclusive and inclusive event classes. While in the exclusive classes the exact number of particles have to be present, the events in the inclusive classes might contain further particles (denoted by $+X$). Within each class distributions of variables which might be sensitive to new physics are fed into a search algorithm identifying the region with the most prominent discrepancy between data and MC.*

5.3.1 Classification – The Event Class Concept

In order to have a well-defined trigger stream and in order to reduce the multi-jet background the analysed is restricted to events which contain at least a lepton (electron or muon) or a photon. This selection is in accordance with the guidelines of simplicity and robustness.

The selected events are sorted into so called event classes which group events according to their final state topology. These classes can be *exclusive* or *inclusive*. Each event class is defined by the amount of physics objects in the event, e.g. 1μ 3jet. In the exclusive case the exact number of particles is required (e.g. 1μ $0e$ 0γ 2jet $\not\!\!E_{\mathrm{T}}$), while the inclusive classes require only a minimal number of particles, e.g. 1μ 3jet $+ X$, that is at least one muon and 3 jets. Inclusive classes are denoted with the suffix $+X$. While each selected event is present in one and only one exclusive event class, it might populate several inclusive event classes. An event which contains for example two electrons and one jet is only in the exclusive class $2e$ 1jet, but in the following inclusive classes: $2e$ 1jet $+ X$, $2e + X$, 1jet $+ X$ and $1e + X$. The particles are ordered by their transverse momentum and only the quantities with the largest momenta are considered when calculating certain variables. This implies that only the photon with the largest transverse momentum enters the $1\gamma + X$ class although an event within this class might contain several photons or other particles.

Inclusive classes might be useful for complex decay chains which are present in supersymmetric events, but their statistical treatment is more complicated due to the overlaps.

Still, inclusive classes might be desirable in combination with exclusive classes. Final states with many jets might be investigated in exclusive classes up to a certain multiplicity from which on all event classes are grouped into a single inclusive event class (e.g. ≥ 5 jets). This would make sense since Monte Carlo generators are not expected to model the kinematics of the 8th or 9th jet correctly anyway.

In order to perform a classification by final state particles a clear definition of the physics objects is required. Currently the MUSiC analysis considers the following objects measured by the CMS detector:

- muons (μ)
- electrons[2] (e)
- photons (γ)
- hadronic jets (jet)
- missing transverse energy ($\not\!\!E_{\mathrm{T}}$).

The combination of all possible final states leads to approximately 200 event classes (100 exclusive and 100 inclusive) which contain at least one event within an integrated luminosity of 100 pb^{-1} [3].

Given the complexity of the analysis, the strategy will be to focus on well-measured and well-understood objects (high p_{T}, central η), even if this implies some loss in efficiency. The LHC is designed to probe the high-energy frontier, thus the analysis assumes that new physics will appear in events with high p_{T} objects. Selection cuts are desired to remain as simple as possible. Similar strategies are useful for any start-up physics study.

τ-leptons are not included so far. They only enter the selection via decays into an electron or muon. Once a τ-identification is well-studied and well-controlled with the first data one could imagine to include also τ-leptons as individual physics objects. A similar argumentation also holds for the inclusion of b-tagging, adding b-jets as separate physics objects.

5.3.2 The Search Algorithm

By comparing the data with the Monte Carlo expectation within each event class, one can quantify the degree of agreement of the data with the Standard Model. The following variables are analysed systematically within each MUSiC event class:

- The **total cross section**, i.e. number of events per class.
- **Kinematic distributions** of an event class such as the scalar sum of the transverse momenta $\sum p_{\mathrm{T}}$ of all its physics objects, the invariant mass M_{inv} or transverse invariant mass M_{T} in case of event classes containing $\not\!\!E_{\mathrm{T}}$. In addition the $\not\!\!E_{\mathrm{T}}$ distribution is investigated separately for all event classes which contain missing transverse energy.

[2] The word "electron" is used as a synonym for electrons and positrons within this work.

[3] Finally the number of event classes will be determined by the sum of event classes present in data and expected from the MC.

These variables are expected to be sensitive to new physics, but are, as shown in chapter 8, also practical to spot deviations caused by a limited understanding of the detector or imperfect tuning of the event simulation and generation.

The systematic analysis of these variables within all event classes is performed by a dedicated search algorithm. This algorithm represents a hypothesis test quantifying the agreement between data and MC expectation well-adopted to the model-independent specific case i.e. without a signal. Its details are explained in section 7. Some representative interpretations of the results/output of this algorithm are discussed in chapter 8.

5.3.3 Potential and Focus of MUSiC

With a first idea of the analysis concept in mind, one could think about its benefits and its application area. As any analysis technique MUSiC has advantages and disadvantages as well as areas where the approach works well and regions where other methods are superior. The perspective of this model-independent analysis can be summarized as follows:

- MUSiC is a *global physics monitor*, sending "alarms" in case of interesting deviations.
- The conclusion that a deviation is a discovery cannot be drawn by MUSiC alone but only in cooperation with more dedicated studies.
- It can help to improve the understanding of detector and SM backgrounds and contribute to the MC-tuning.
- MUSiC has a rather large coverage of new physics, but for some signals it is likely to be less sensitive compared to dedicated analyses in a specific channel.
- The generality of the approach allows to spot deviations in many regions not covered by a dedicated search. On the other hand it has to rely more on the background predictions made by Monte Carlo generators. As some physics cannot be modelled well with MC, like multi-jet background, an estimation from the data has to be implemented in a generic way for those cases.
- There is a clear trade-off between trying to cover a large amount of data and describing all of it properly.
- The key issue is to estimate and implement uncertainties with the correct order of magnitude such that problematic areas of the phase space are assigned with a reasonable uncertainty. In this way only indications of new physics, unexpected detector effects or insufficient knowledge of the Standard Model processes should remain as significant deviations.
- MUSiC has the potential of "unblinding" any analysis in CMS. Every information picked up from this broad data scan could bias a dedicated search investigating the same data.
- The objective search results need to be interpreted by a physicist to add some subjective knowledge or intuition. A deviation in a 10 jet channel for example is not really

a surprise since Monte Carlo generators are not expected to model such extreme topologies correctly.

- Even if one could conclude that the deviations are caused by physics beyond the Standard Model there still remains the question known as the inverse LHC problem: What is the underlying theory causing this signal?

Similar strategies have already been applied successfully at other accelerator experiments, see e.g. [122–128]. From the historical point of view the MUSiC concept follows the principal ideas of a similar strategy at L3 [128]. The search algorithm is inspired by the H1 approach [125, 126] and has already been exercised at Aachen with collision data taken by the DØ experiment [129]. In contrast to the Sleuth/VISTA approach [127] the MUSiC algorithm does not rely on a complex self-correction model trying to tune the simulation to the data. The idea is to benefit from the many detailed studies on particle identification and detector efficiencies and feed them in a transparent way into MUSiC. The current detector knowledge is used as an input and allows to learn from the results of a global data-Monte Carlo comparison. Therefore MUSiC is an excellent monitor to detect improvements or new discrepancies for high p_{T}-processes, allowing the cooperation with dedicated studies and contributions to the tuning of the simulation.

5.3.4 The MUSiC Timeline

The previous section already points to the fact that such a generic approach has its most suitable application in different areas through the life-time of an experiment. With the first pb^{-1} of data to arrive the focus will not be on the discovery of new physics but on re-establishing the Standard Model, understanding of the detector and validating the Monte Carlo predictions. One would concentrate on the high statistics parts of the distributions where the SM candles dominate. In this way it is also possible to ensure that one is not overwhelmed initially by deviations found by the algorithm, thus reducing the amount of distributions to be studied in detail. In this phase of data analysis MUSiC can contribute to the understanding of the detector and the tuning of the simulation and the event generators. In fact MUSiC is the first model-independent analysis which is already implemented before the arrival of the first data. For the first time it will be possible to exploit the benefits of such an approach in detector and physics commissioning.

In a next phase the focus will begin to shift also to the tails of the distributions where higher order effects like jet-multiplicities become important. Here the validation of the MC predictions will be crucial and comparisons of different event generators, e.g. MadGraph vs alpgen vs sherpa, will be important. Also here MUSiC can contribute, comparing data and MC predictions in a large part of the phase space. While one generator might describe one part of the data properly, it might fail in another part. Each time new generator parameter tunes are available MUSiC can compare them to data in a general way and thus be important for the overall generator validation.

After all initial problems have been solved and confidence in the understanding of the detector and the MC prediction is present, the full dataset available can be analysed with MUSiC and one can start looking for deviations from the Standard Model.

5.4 The Implementation

The analysis is implemented in several independent steps allowing for cross checks at each stage. An overview of the workflow is given in Figure 5.2. The analysis starts within the LHC computing grid with the skimming of the data samples which either stem from the full detector simulation or the detector itself. This step is performed to reduce the data from the order of several 100 TB to a more suitable and handy data format (~ 6 kB per event) containing only the information relevant for the MUSiC search algorithm and information for cross checks. This includes also a first loose application of object identification criteria (preselection). The skimming relies on the CMS C++ Physics Analysis Toolkit (PAT) [130] which suits as a user front-end for the CMS framework CMSSW [131]. PAT allows the simplified access of all physics objects at generator and reconstruction level including trigger information. In addition it offers a lot of other tools simplifying one's life e.g. code which matches generated and reconstructed particles with each other. It subsumes algorithms used in many analyses and allows adoptions for the individual analyses via the CMS configuration language. The objects and information extracted from PAT are stored in an object-oriented manner within the objects provided by the C++ Physics eXtension Library (PXL) [132]. PXL provides for example a handy event container which can be filled with particle objects, vertex objects and associations between them. It allows the storage and re-reading of these data in compact streamer files. The required information are extracted from the various datasets provided by the CMS data operation teams and stored in the grid dCache storage at Aachen in a PXL-specific format.

In the next step the events are read back with PXL and either fed into the so called control plot factory or the event class factory. Both are derived PXL classes and can run in parallel. While the former is developed to provide immediate feedback of the quality of the selected objects which enter the MUSiC algorithms, the latter performs the sorting of the events into event classes (first step in Figure 5.1). Following an object-oriented ansatz each event class is represented by an instance of a TEventClass class. TEventClass, as denoted by the leading "T", is derived from a basic ROOT [133] class. The class provides all the necessary containers for the storage and simple access to different distributions such as $\sum p_{\mathrm{T}}$, M_{inv} or $\not\!\!E_{\mathrm{T}}$. Upon arrival of an event the corresponding plots of the TEventClasses are filled automatically together with the bookkeeping information such as the number of analysed events, the physics processes or the cross sections. Also the information about the systematic uncertainties are kept. The ROOT Input/Output machinery is utilized for the persistent storage of the TEventClass objects and the information encapsulated within. For a rapid and efficient turnaround this step can be performed within the grid.

These TEventClass objects are picked up in the final step of the analysis which consists of the application of the MUSiC search algorithm. With the user's definition of which events should be considered as Standard Model reference and which as (pseudo-)data the algorithm evaluates the most significant deviation of both taking systematic and statistical uncertainties into account. This flexible design allows to compare MC versus MC distributions i.e. Standard Model expectation versus Standard Model plus a new physics signal, but will serve with the start-up of the LHC for Standard Model MC versus data compar-

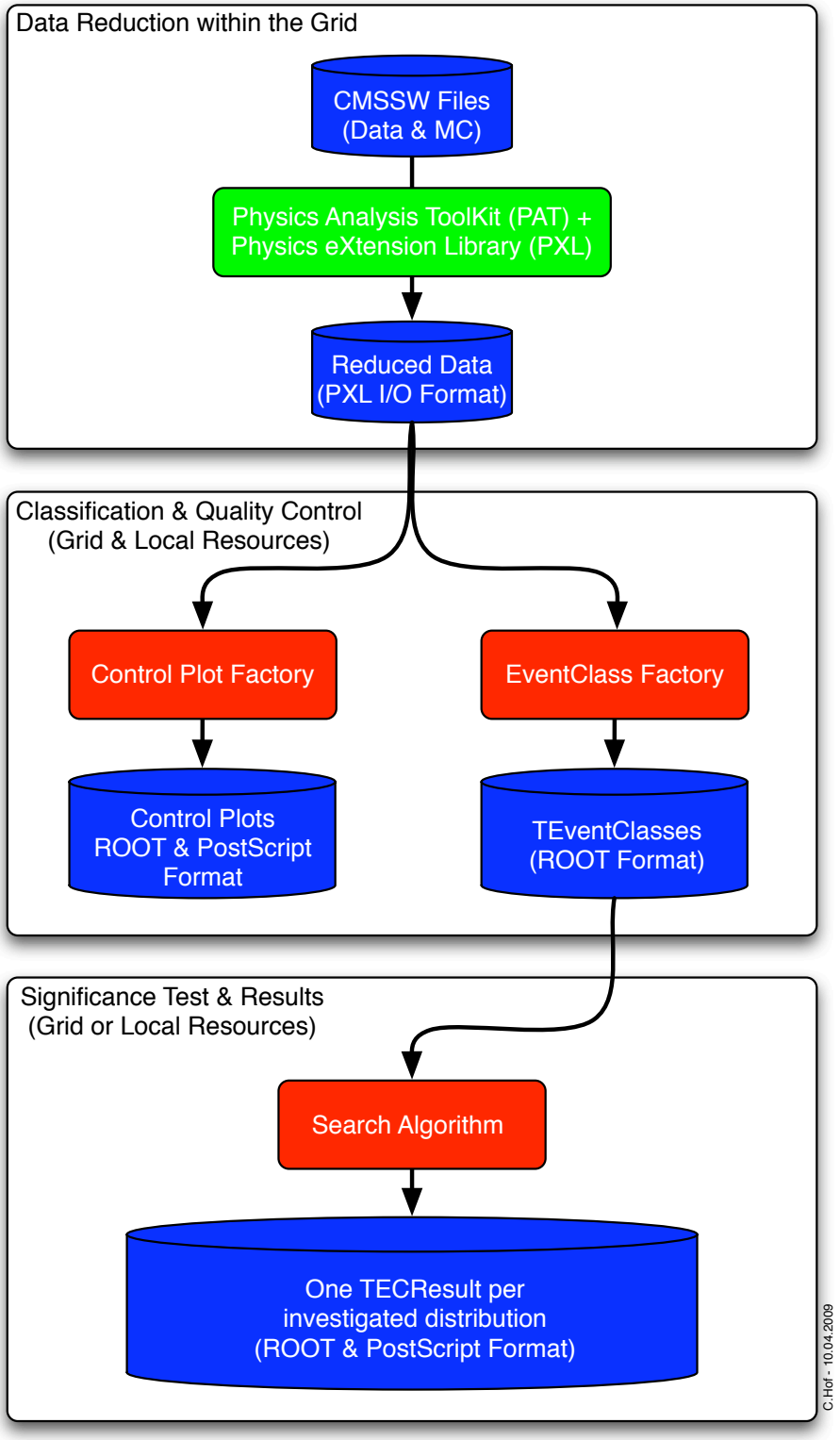

Figure 5.2: *Technical overview of the analysis steps and the intermediate data formats. The steps introduced in Figure 5.1 are reflected by the object-oriented program structure. Each step allows for cross checks and transparent fast feedback. Tools have been developed to be able to utilize the large computing resources of the LHC computing grid.*

isons. Finally the search results are stored both in a ROOT class and in an immediate visualization as a PostScript output. In addition tables with the significance ranking of the individual event classes are printed in the PDF format.

5.5 The CMS Monte Carlo Simulation

One input to the MUSiC analysis are Monte Carlo simulated events, which will be compared to data upon the start of the LHC. These MC events are generated centrally by the CMS production teams utilizing the full CMS detector simulation, which emulates the processes in the CMS detector as realistic as possible within a reasonable resource budget concerning timing, storage and CPU. As these events are the basis of our expectation to be confronted with data, the simulation and reconstruction framework is a cornerstone of the experiment.

The overall collection of software, referred to as CMSSW, consists of a framework, an event data model (EDM), and services required by the simulation, calibration and alignment. In addition reconstruction and analysis modules are implemented to process event data and to provide physicists the tools to perform analysis. The primary goal of the framework and the EDM is to facilitate the development and deployment of simulation, reconstruction and analysis software. The CMSSW event processing model [134] consists of one executable, called cmsRun, and many plug-in modules which are managed by the framework. All the code needed in the event processing (calibration, reconstruction algorithms, etc.) is contained in the modules. A module is a piece (or component) of CMSSW code that can be plugged into the CMSSW executable cmsRun. Each module encapsulates a unit of clearly defined event-processing functionality. Modules are implemented as plug-ins (core libraries and services). They are compiled in fully-bound shared libraries and must be declared to the plug-in manager in order to be registered to the framework. The framework takes care to load the plug-in and instantiates the module when it is requested by the job configuration. This configuration file (implemented as Python code) instructs cmsRun which data to use, which modules to execute in which order with which parameter settings. In addition filters can be declared within each executed sequence. Finally, the configuration of the output module defines which data are stored persistently within the output file.

Unlike the previous event processing frameworks, cmsRun is extremely lightweight: only the required modules are dynamically loaded at the beginning of the job. This concept makes the compilation of the binary executables superfluous. The CMS event data model is centred around the concept of an Event. The Event is a C++ object container for all simulated and reconstructed data related to a particular collision. During processing, data are passed from one module to the next via the Event, and are accessed only through the Event (see figure 5.3). All objects in the Event may be individually or collectively stored in ROOT files, and are thus directly browsable in ROOT. This allows tests to be run on individual modules in isolation. Auxiliary information needed to process an Event are stored and accessed via the EventSetup.

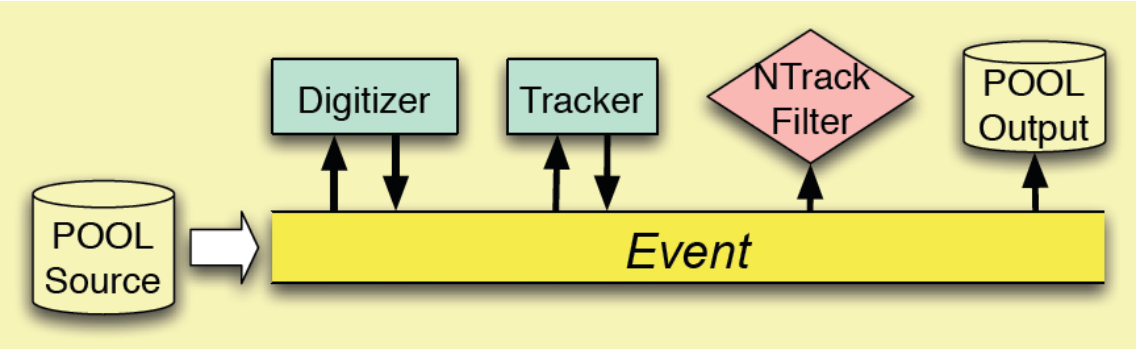

Figure 5.3: *Illustration of the EDM processing model centred around the C++ class of an event [134]. The event object can be read from disk via a pool source or can be created from scratch. According to the user defined schedule (configuration file) the event is passed from one module to another (here: "Digitizer", "Tracker") which interact with the event i.e. modify information. Filter (here: "N Track Filter") allow the selection of certain objects. Finally, the event object can be stored on disk via the output module.*

A typical simulation chain starts with the generation of a single proton-proton collision of interest. This is the field of the so-called "event generators". These programs like PYTHIA, MADGRAPH, SHERPA, or MC@NLO simulate the hard interaction (including the underlying event i.e. the proton remnants) at a certain precision (tree level, LO or even NLO). While certain generators only provide a description of the hard scattering, programs like PYTHIA also perform the hadronization and fragmentation of the produced partons up to stable particles.

In the following step the CMS framework CMSSW picks up this list of particles including their momentum and timing information. Based on the GEANT 4 toolkit [135] the simulation traces the particles through the detector, calculates their interactions, energy losses and deposits (detector hits). Also decays and secondary particles which might result in interactions with the detector are handled. In addition the mixing of pile-up events (multiple *pp*-collisions which occur within the same bunch crossing (in-time pile-up) as well as overlying events from different *pp*-collisions (out-of-time pile-up)) is supported.

The next step emulates the electronic response of the particle hits within the detector. The hits are digitized mimicking the readout electronics as closely as possible, including the simulation of the trigger. In the last step the reconstruction of the physics objects is performed using the hits and energy deposits within all detector parts. This involves also the access of the latest calibration and alignment data. The framework provides a persistency mechanism which allows a modular storage of the data at every stage in the chain including different data tier definitions (see also chapter 4.4).

At this stage all physics objects required for analysis are reconstructed and can be accessed via the full CMSSW framework (e.g. by utilizing the Physics Analysis Toolkit PAT), a light-weighted set of libraries (framework light) for the fast analysis on a laptop, or even in bare ROOT.

Chapter 6

Object Identification and Selection

The concept of a model independent search implies that selection cuts are not chosen or optimized according to a specific signal beyond the Standard Model since this would introduce a strong bias. For MUSiC the aim of the selection cuts is to analyse standard physics objects which are robust and well-understood within the experiment, even if this implies some loss of statistics. As an example leptons with a relatively high p_{T} threshold of 30 GeV are used since these are expected to be well under control very early and not affected by any trigger threshold effect (turn-on). Following this concept the analysis at this stage does not distinguish between light quark jets and b quark jets in particular since b-tagging requires detailed studies and a reasonably well-commissioned detector. Following the guidelines described in section 5.2 the strategy is to keep it simple and to focus on objects which are well-studied and recommended by the CMS physics object groups. Relying on standard physics objects MUSiC can benefit from dedicated studies which e.g. determine efficiencies from data or develop selection cuts well-suited for rejecting misreconstructed objects in real data. In this context MUSiC serves as an additional cross check of these numbers or variables in a more general frame. This ability has already proven to be very useful at this stage of the analysis as many coding mistakes in the CMS simulation and reconstruction framework or misconfigurations of produced datasets have been spotted by the MUSiC analysis first (for further details see section 8.1) .

This chapter introduces the physics objects which enter the MUSiC analysis and thus the object's definition. Their reconstruction principle is outlined briefly and the applied selection and quality criteria are discussed and documented with representative distributions. Finally the trigger paths which have to be full-filled for this study are presented. The trigger menu used is designed for a centre of mass energy of 10 TeV with a focus on a luminosity of the order of $\mathcal{L} = 10^{30}\ \mathrm{cm}^{-2}\,\mathrm{s}^{-1}$.

The variable *(identification) efficiency* will appear several times. It is defined utilizing the Monte Carlo truth information of the simulated events. After the application of the identification, selection and acceptance cuts like $|\eta|$, p_{T} or isolation, the generated objects are matched to the reconstructed objects using a $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$ criterion (cone size depending on the physics object). The efficiency is then given as:

$$\varepsilon_{\mathrm{ID}} = \frac{N(\text{generated and matched to reco})}{N(\text{generated})}. \tag{6.1}$$

This efficiency is not a pure reconstruction efficiency, but includes also the identification and selection. It might be given as a function of a kinematic variable such as η or p_{T}. In this case the reference is defined by the generator truth e.g. the η or p_{T} at generator level. Studies to derive these numbers from data are currently under investigation within numerous CMS analyses, e.g. using the tag–and–probe method [136, 137].

In a similar fashion as the efficiency, so-called *fakes* are defined. All reconstructed particles which cannot be matched to a corresponding object type at generator level with a ΔR criterion are considered as "fake". This definition serves the needs for such a generic approach. However, one should be aware that it might lead to cases where a certain object which arose in a subsequent interaction with the detector material is considered as fake. For example a photon, which is emitted in the interaction of a charged particle within the tracker and reconstructed as photon within the electromagnetic calorimeter would be a fake since it has no matching photon at generator level. As the inclusion of fake-rates is an important item, MUSiC will rely on external studies focusing on the determination of fake rates from data as explained e.g. in [138]. A detailed discussion of the precise handling of fake rates and its uncertainties within MUSiC is given in section 7.7.

6.1 Muon Selection

Global muons [139] are reconstructed utilizing the muon system and the inner tracking detectors. Starting from the muon spectrometer, hits within each drift tube and cathode strip chamber are connected to segments compatible with the beam spot. A combination of matching segments is used as seeds for the actual track building and fitting within the DT, CSC and RPC subdetectors via a Kalman filter [140]. The final fit through the whole muon spectrometer results in a track known as "standalone muon". Via geometrical constraints and momentum comparison, standalone muons are matched with a track inside the tracker. Finally a "global muon" is obtained by a global refit using the hits from the tracker track and the stand-alone muon track. In case of multiple matches inside the tracker the fit with the best χ^2 is chosen.

The addition of calorimeter information allows the calculation of isolation quantities and a cross check of the compatibility with a track of a weakly interacting particle i.e. a muon with a momentum up to a TeV.

For the selection of muons the MUSiC analysis follows the recommendations of the muon physics object group [141]. The applied selection criteria are:

- Global Muons
- $p_{\mathrm{T}}(\mu) > 30$ GeV
- $|\eta(\mu)| < 2.1$
- $R_{\text{Track Isolation}} = \frac{\sum p_{\mathrm{T}} \text{ of tracks in 0.3 cone excluding the muon itself}}{p_{\mathrm{T}}(\mu)} < 0.1$

- $N_{\text{Tracker Hits}} \geq 11$
- $\chi^2/DoF \leq 10$
- Vertex compatibility: $|d_0| < 2$ mm (in the xy-plane w.r.t. the primary vertex)
- Calorimeter and segment compatibility.

The chosen η-acceptance is induced by the coverage of the muon detectors which provide the input to the L1 single muon trigger. The p_{T} requirement of at least 30 GeV ensures that the muons are well above the trigger threshold (the HLT requirement for a single muon is approximately 15 GeV, see section 6.7).

Muons with such momenta should easily cross the iron of the return yoke, leading to track-segments within the muon system which can be combined with a tracker track. The cuts on the number of hits and on the normalized global χ^2 of the muon track fit are designed to suppress mismeasured muon candidates which tend to have unphysically high p_{T}.

The isolation variable helps to suppress muons originating from multi-jet events. These non-prompt muons tend to be within or close to hadronic jets and are therefore more difficult to reconstruct, given the higher silicon track multiplicities. In addition high energetic particles might leak out of the calorimeter into the muon system, so-called "punch-through", and cause higher muon segment multiplicities within the first muon stations. Still it has to be stressed that since the isolation criterion is quite loose there is no strict focus on only prompt muons. The loose cut on the vertex compatibility suppresses out of time signals arising from cosmic or beam-halo muons, but keeps isolated muons from secondary vertices.

The calorimeter compatibility cut represents a likelihood which checks if the muon candidate has a calorimeter deposit consistent with a muon hypothesis. In addition a segment compatibility likelihood evaluates if the muon has caused segments in the muon system where it is expected to traverse. These variables further clean the muon selection and allow a separation of prompt muons from muons which arise in decays of kaons or pions [141].

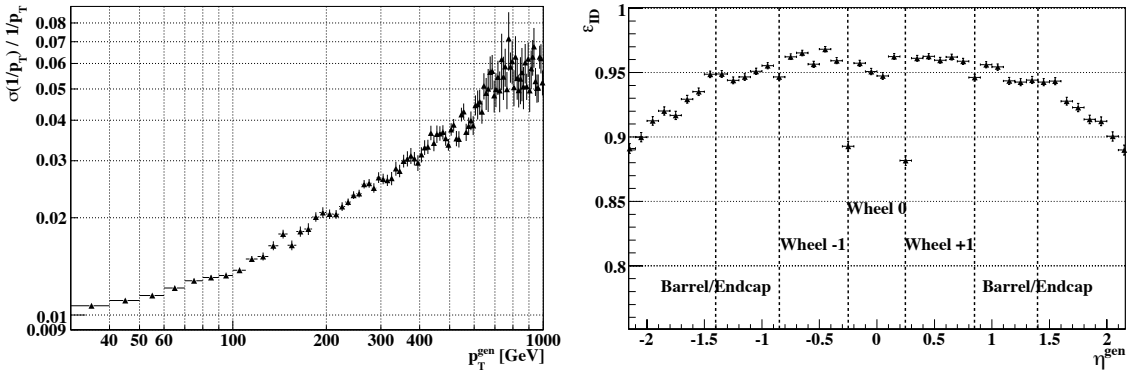

Figure 6.1: ***Left:*** *p_{T}-resolution of global muons within the barrel ($|\eta| < 1.1$) as a function of the muon momentum p_{T} at generator level using Drell-Yan events.* ***Right:*** *Muon reconstruction efficiency as a function of the pseudorapidity η at generator level using Drell-Yan events around the Z-pole.*

Figure 6.1 shows representative quality plots for the selected global muons. The left plot displays the transverse momentum resolution as a function of the momentum at generator level for barrel muons from Drell-Yan events (40 GeV $\leq m_{\mu\mu} \leq$ 2 TeV). The resolution is determined by fitting a Gaussian function to the variable $1/p_{\mathrm{T}}^{\mathrm{gen}} - 1/p_{\mathrm{T}}^{\mathrm{rec}}$. This difference is proportional to the spatial resolution (sagitta measurement) of the detector which is expected to be Gaussian. The distribution comprises two regimes with different slopes: for energies up to 100 GeV the resolution is dominated by the tracker due to the distortion of the measurement due to multiple scattering within the return yoke. For higher momenta the larger lever arm of the muon system is needed to obtain a reasonable measurement.

The right plot of figure 6.1 reflects the high muon reconstruction efficiency as a function of the pseudorapidity of more than 95% (90%) in the barrel (endcap). It has been created using Drell-Yan events around the Z-pole, matching generated muons to reconstructed, selected global muons. The dips at $\eta \approx \pm 0.25$ and ± 0.85 are caused by the not instrumented gaps between the wheels of the return yoke. The effect is most prominent in the transition region of the central wheel since muons originating at the vertex more likely travel along the gap. The identification efficiency is almost flat in p_{T} with a slight decrease at the TeV-range where muons start to emit significant amounts of bremsstrahlung disturbing the global muon reconstruction algorithm. Dedicated reconstruction procedures for TeV-muons are currently being under investigation in CMS [142].

6.2 Electrons

Electrons [143] are reconstructed by combining energy measurements within the calorimeters and momentum measurements in the central tracking devices. The electron as well as the photon reconstruction starts with the combination of clusters of energy deposits (so-called superclusters) inside the electromagnetic calorimeter. These superclusters represent the electron plus the collected bremsstrahlung emitted along the electron trajectory in the tracker volume.

Using the energy estimate from the supercluster, the reconstruction algorithm searches for geometrically matching hits within the pixel detector assuming electrical charges of ± 1. The matching pixel hits are used as seeds for the trajectory building via a Gaussian Sum Filter (GSF) which is a modified Kalman filter that takes the electron specific energy losses into account [144]. The electrons are classified according to variables sensitive to the amount of emitted bremsstrahlung. The classification is used to apply energy scale corrections to the superclusters and to estimate the associated uncertainties. Finally the electron energy is derived from a weighted combination of the corrected supercluster energy and the momentum measurements within the tracker. Measurements within the hadronic calorimeter are used to determine the hadronic shower fraction of the electron candidates and serve for isolation purposes.

A robust and simple identification is demanded at the LHC start-up period until data are available to verify and tune the selection criteria. Therefore MUSiC focuses on the most predictable and stable electron variables possible. The applied selection criteria use

Variable	Detector	Cut Value			Description
		Type 1	Type 2	Type 3	
$E_{\mathrm{had}}/E_{\mathrm{elm}}$	Barrel	< 0.042	< 0.050	< 0.045	energy ratio of the deposits within HCAL and ECAL
	Endcap	< 0.037	< 0.055	< 0.050	
$\sigma_{\eta\eta}$	Barrel	< 0.011	< 0.0125	< 0.010	shower shape variable, lateral width in η
	Endcap	< 0.0252	< 0.0265	< 0.026	
$\lvert\Delta\phi_{\mathrm{in}}\rvert$	Barrel	< 0.016	< 0.032	< 0.0525	difference of supercluster ϕ & track ϕ at ECAL
	Endcap	< 0.035	< 0.025	< 0.065	
$\lvert\Delta\eta_{\mathrm{in}}\rvert$	Barrel	< 0.0030	< 0.0055	< 0.0065	difference of supercluster η & track η at ECAL
	Endcap	< 0.0055	< 0.0060	< 0.0075	
$E_{\mathrm{seed}}/P_{\mathrm{in}}$	Barrel	> 0.94	> 0.24	> 0.11	ratio of seed cluster energy and track momentum
	Endcap	> 0.83	> 0.32	> 0.00	

Table 6.1: *Variables and cuts used for the selection of "tight" electrons. Type 1 electrons have a f_{brem} (relative momentum change of the track between vertex and calorimeter entrance, which is proportional to the amount of bremsstrahlung emitted by the electron) of less than 6% / 10% in the barrel/endcap. If electrons exceed the previous cut but have an E/p between 0.8 and 1.2 they are classified as Type 2 otherwise as Type 3. Electron candidates with $f_{\mathrm{brem}} < 0.2$ and $E/p < 0.8$ are discarded as well as candidates which fulfill $E/p < 0.9 \cdot (1 - f_{\mathrm{brem}})$.*

a standard cut-based electron definition of the electron physics object group [145] which relies on the classification of the electrons into three categories[1]:

- Type 1: Electrons with only a few bremsstrahlung deposits (high population from both real and fake electrons).
- Type 2: Electrons with reasonable bremsstrahlung deposits (electron-like region with little contamination from fakes).
- Type 3: Electrons with bad matching of energy and momentum measurement (region with not many real electrons).

The cut-based electron identification uses the variables E/p, the hadronic and electromagnetic energy ratio $E_{\mathrm{had}}/E_{\mathrm{elm}}$, the cluster shape $\sigma_{\eta\eta}$, and the matching between the track and the supercluster in η and ϕ. Different cuts are applied to different electron classes, also distinguishing electrons measured in the endcap and the barrel sub-detectors (see Table 6.1).

The complete list of selection criteria for electrons reads as:

- Pixel Matched Gaussian Sum Filter Electrons
- Electron identification: *tight (category based)*
- $p_{\mathrm{T}}(e) > 30$ GeV

[1] An even more robust electron identification forbears the categorization and just uses the most stringent cut of Table 6.1. This is likely to be used within the analysis of the very first data within MUSiC, but has a quite low efficiency. Since this study is focusing on the feasibility of the first years of data taking, it relies on the more efficient and almost similar robust category-based identification.

- $|\eta(e)| < 2.5$
- $R_{\text{track isolation}} = \frac{\sum p_{\text{T}} \text{ of tracks in 0.3 cone}}{p_{\text{T}}(e)} < 0.1$
- Vertex compatibility: $|d_0| < 2$ mm (in the xy-plane w.r.t. the primary vertex)

Electrons within $|\eta| < 2.5$ are expected to have the complete electromagnetic shower contained within the electromagnetic calorimeter. In addition the preshower detector[2] placed in font of the endcap ECAL can be used to suppress neutral pions. Since electron candidates are required to have a matching inner (pixel) track, the electron reconstruction is geometrically limited by the η-coverage of the tracker ($|\eta| \approx 2.5$). The cut $p_{\text{T}} > 30$ GeV matches to the muon requirement and also ensures that the electrons are well above the trigger threshold (see section 6.7). The isolation criterion, as in the muon case, is based on tracker tracks within a cone of $\Delta R < 0.3$ around the electron. The electron momentum is subtracted by excluding an inner cone of $\Delta R < 0.015$ from the sum. This isolation ensures a clean electron measurement and rejects contamination from multi-jet events, e.g. jets with many $\pi^0 \to \gamma\gamma$ decays. A loose vertex compatibility cut matching the muon case is applied. Note that this cut still allows for non-prompt electrons a priori.

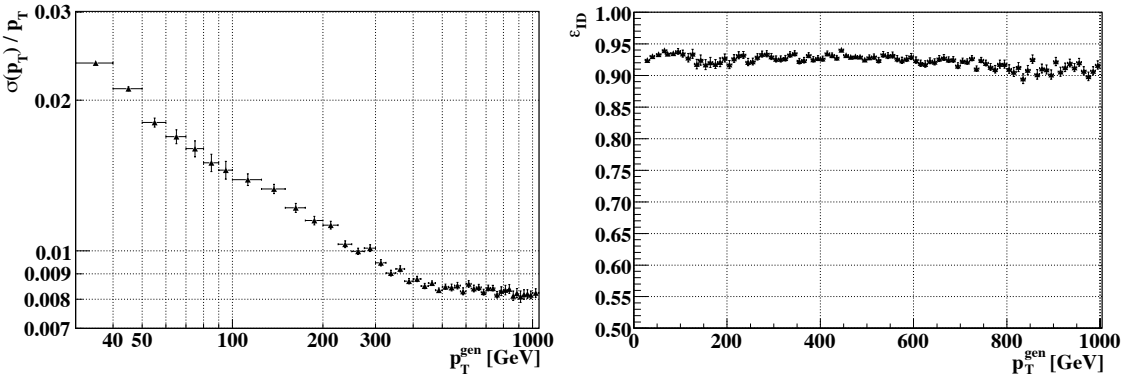

Figure 6.2: ***Left:*** *Electron p_{T}-resolution as a function of the electron momentum p_{T} at generator level.* ***Right:*** *Electron reconstruction efficiency within the barrel ($|\eta| < 1.4$) as a function of the electron momentum p_{T} at generator level. For both distributions W+Jet events over a wide invariant W mass spectrum up to 2* TeV *have served as input.*

Representative performance plots are shown in figure 6.2 and 6.3 for the selected electrons from a W+Jet dataset with W masses around the W peak up to 2 TeV. The p_{T}-resolution of barrel electrons (figure 6.2, left) has been determined by fitting a Gaussian function to the variable $p_{\text{T}}^{\text{gen}} - p_{\text{T}}^{\text{rec}}$. It obeys the expected behaviour of increasing relative precision with energy given by the statistical nature of the electromagnetic shower. The resolution ranges from $\sim 2\%$ at 30 GeV to $\sim 0.7\%$ at 1 TeV .

The overall identification efficiency is above 90% over the large p_{T}-range from 30 – 1000 GeV (see figure 6.2, right). This can be mapped to the excellent geometrical coverage of the electromagnetic calorimeter as seen in the efficiency plot as a function of the pseudorapidity η (see figure 6.3, left). The efficiency is almost flat in η with only moderate dips

[2]Note that the preshower detector has been disabled in the Monte Carlo events used in this study since it was not expected to be finished at the LHC start-up.

Variable	Cut Value		Description
	Barrel	Endcap	
Cluster Shape R_9	> 0.8	> 0.8	energy ratio of the deposits in 3 x 3 cluster & total supercluster energy
ECAL Isolation	< 10.0 GeV	< 10.0 GeV	Hollow cone $0.06 < \Delta R < 0.4$ excluding an η bar of size 0.08
HCAL Isolation	< 5.0 GeV	< 10.0 GeV	Hollow cone $0.1 < \Delta R < 0.4$
Track Isolation	< 30.0 GeV	< 30.0 GeV	Hollow cone $0.04 < \Delta R < 0.4$

Table 6.2: *Variables and cuts used for the selection of "tight" photons.*

at the ECAL module borders ($|\eta| \approx 0.0, 0.4, 0.8, 1.1$), at the insensitive transition region between barrel and endcap ($|\eta| \approx 1.5$), and at the acceptance edges of the inner tracker ($|\eta| \approx 0.0$ and > 2.4).

6.3 Photons

The photon reconstruction [146] relies on the same clustering algorithm as the electron reconstruction. Thus, a priori every electron candidate is also a photon candidate. The distinction between both needs to be done at the identification level e.g. by a track veto. The clustering algorithm allows to recover clusters which are spread due to bremsstrahlung and photon conversions within the relative large amount of material budget in front of the electromagnetic calorimeter. However, because of the strong 4 T magnetic field the energy reaching the calorimeter is spread mainly in ϕ. Similar to the electron case photon-specific energy scale corrections are applied to the supercluster. Using the superclusters obtained by the clustering algorithms one determines the position of the photon at the ECAL impact point by an energy-weighted mean position of the crystals in the cluster.

The MUSiC photon selection uses the robust cut-based standard photon identification of the e/γ particle object group. To ensure a high reconstruction efficiency and a low misidentification rate "tight" identification criteria are applied. The variables considered in this cut-based identification are the isolation within the electromagnetic and hadronic calorimeter as well as a coarse track isolation and the shower shape variable R_9 (ratio of the energy deposit in a 3 x 3 cluster and the total supercluster energy). The detailed "tight" cuts are given in table 6.2. The complete set of cuts which photons have to fulfill in order to enter the MUSiC analysis are:

- Photon identification: *tight*
- $p_{\mathrm{T}}(\gamma) > 30$ GeV
- $|\eta(\gamma)| < 2.5$
- veto on a matched pixel seed
- $E_{\mathrm{had}}/E_{\mathrm{elm}} < 0.2$

- $R_{\text{track isolation}} = \frac{\sum p_{\text{T}} \text{ of tracks in 0.3 cone}}{p_{\text{T}}(\gamma)} < 0.1$

Due to the ambiguity of the electron and photon candidates a clear separation of both has to happen with cuts at the identification level. While prompt electrons should place at least two hits in the pixel detector, prompt isolated photons can only produce a track inside the pixel detector if they already convert that early. Since the pixel material budget (see figure 3.12) is 0.1 X_0 in the barrel and 0.4 X_0 in the endcap, the conversion probability is relatively low (Barrel: $\approx$ 3%, Endcap: $\approx$ 10%). Thus, to obtain disjoint photon and electron candidates the photons are required to have no hits inside the pixel detector.

The acceptance cuts on $|\eta|$ and p_{T} follow the electron case. Measurements within the tracker are used to apply a track veto. The cut on the hadronic over electromagnetic energy deposits is chosen to minimize the probability of a jet faking a photon. No explicit veto is applied to photons which converted to a e^+e^- pair within the tracker. However one should notice that the shower shape variable R_9 used within the identification algorithm is sensitive to conversions within the tracker and removes a fraction of them (see performance plots below). Since the conversion probability for photons with a momentum of the order of 100 GeV is about 30 – 70% (strongly depending on η), the inclusion of converted photons increases the identification efficiency significantly while keeping the fake rate at a reasonable level. A dedicated reconstruction algorithm is currently developed [147] which will further enhance the reconstruction quality of converted photons.

Isolation is required in order to reject contaminations from π^0 decays within hadronic jets. In this way photons with a considerable p_{T} originating from initial or final state radiation within the hard interaction as well as isolated photons coming from the decay of new particles (e.g. excited leptons) are selected.

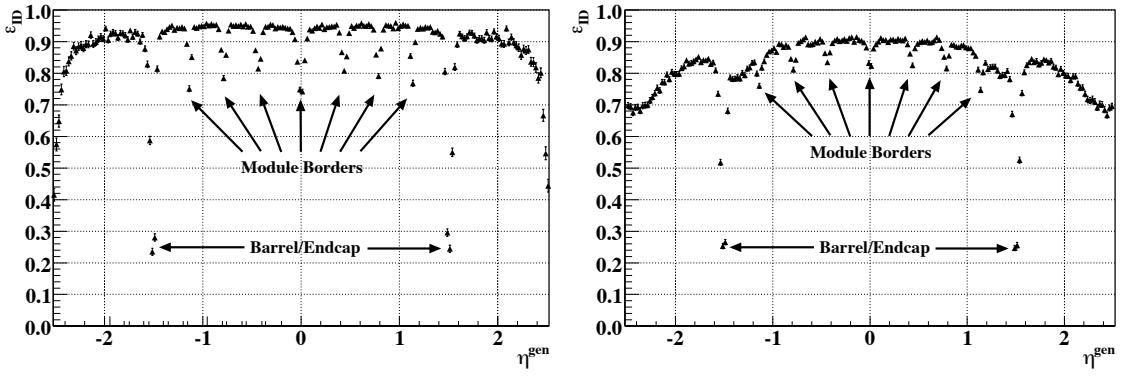

Figure 6.3: *Comparison of the electron* ***(left)*** *and photon identification efficiency* ***(right)*** *as a function of the pseudorapidity η. The electron efficiency is derived from a W+Jet sample while the photon efficiency has been calculated from a Photon+Jets sample with transverse photon momenta around 100* GeV.

The photon resolution is very similar to the electron resolution given in figure 6.2 (left). Figures 6.3 and 6.4 show some representative performance plots for the selected photon candidates. The efficiencies have been determined by matching reconstructed to generated photons from Photon+Jets samples. The identification efficiencies as a function of the

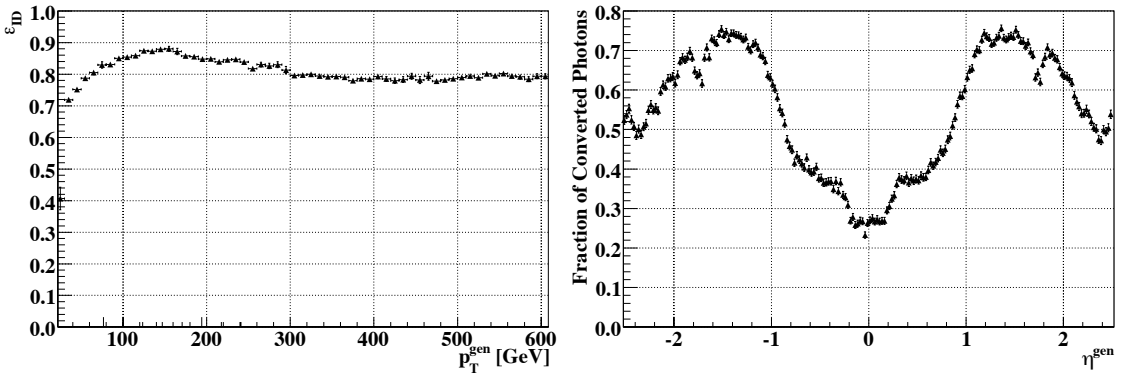

Figure 6.4: ***Left:*** *Photon identification efficiency as a function of the transverse photon momentum at generator level.* ***Right:*** *Probability for a photon to convert into an electron-positron pair before reaching the electromagnetic calorimeter as a function of the pseudorapidity. For both plots Photon+Jets samples have been utilized. While in the left plot photons with transverse momenta from 15 – 800* GeV *are taken into account, the right plot is restricted to photons with transverse momenta from 80 – 120* GeV*.*

transverse momentum at generator level is around 80 – 90% almost flat in p_T for Photon+Jets samples generated with PYTHIA (15 GeV $\leq \hat{p}_T$[3] $\leq$ 800 GeV).

The identification efficiency versus the pseudorapidity η shows the same characteristics as the electron distribution (see figure 6.3). While the efficiency is almost 90% within the barrel ($|\eta| < 1.0$), the efficiency decreases in the region beyond $|\eta| > 1$. This is given by the material budget and the related increase of photon conversions which have a lower identification efficiency (see figure 6.4, right).

6.4 Hadronic Jets

Jets are groups of particles or energy deposits collected by dedicated algorithms. These algorithms allow to deduce parton level information from the measurements within the calorimeter. Various jet algorithms have been implemented within CMS in such a modular way that any set of four-vectors can serve as an input. This allows to study jets of partons (partonic jets), jets of particles remaining after the hadronization (particle jets), and jets of energy deposited in the detector (calorimeter jets). A powerful algorithm results in similar jet collections at all levels. Jet algorithms should fulfill the criteria of infrared and collinear safety: the algorithm should be invariant under the addition of very soft particles (e.g. soft gluons) and should be insensitive to small substructures (e.g. energy distributed to two very close particles instead of one). In practice collinear unsafety is introduced by the limited granularity of the calorimeter and any energy/momentum threshold on the objects which enter the jet algorithm.

The MUSiC analysis relies on the recently developed cone-based algorithm: the Seedless Infrared Safe Cone (SISCone) algorithm [148] with a radius of $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2} < 0.5$. This cone algorithm, with reasonable execution time, is infrared as well as collinear safe

[3] Transverse momentum of a PYTHIA $2 \rightarrow 2$ process in the rest frame of the interaction before the parton shower.

at all orders. It is thus superior to the other implemented cone jet clustering algorithms and has properties comparable to other algorithms such as the kt-jet finder [149]. It uses the energy deposits from the electromagnetic and hadronic calorimeters above a certain threshold as input.

The minimal required standard L1–L3 jet energy scale corrections [150] are applied in order to have a proper estimate of the jet at parton/particle level. These factorized corrections are associated with different detector and physics effects: The first level offset correction accounts for pile-up and noise and tries to subtract, on average, the unwanted energy from the jet. Since the response of the CMS detector for a jet with fixed transverse momentum varies with the pseudo-rapidity an angular dependent correction is applied at level two. Finally, at level three, the jets are corrected for the absolute response as a function of transverse momentum. Thus the total jet energy can be symbolically written as:

$$\text{Corrected CaloJet Energy} = (\text{CaloJet Energy - offset}) \times C(rel., \eta) \times D(abs., p_{\mathrm{T}}) \quad (6.2)$$

Further corrections for the electromagnetic fraction within the jets or the jet flavour are not taken into account within MUSiC due to the loss of robustness and generality. The discussed corrections are currently determined from MC simulations and test beams, but will be derived in a data-driven manner from di-jet, photon-jet or Z+Jet events [150].

Jets must fulfill the following criteria to enter the MUSiC analysis:

- SISCone jets with $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2} < 0.5$
- $p_{\mathrm{T}}(\text{jet}) > 60$ GeV
- $|\eta(\text{jet})| < 2.5$
- $E_{\mathrm{had}}/E_{\mathrm{tot}} > 0.05$

The acceptance cut in $|\eta|$ ensures that the whole hadronic shower is contained within the barrel and endcap of the hadronic calorimeter. The p_{T} threshold ensures that the energy resolution of the hadronic jets is reasonable ($< 20\%$). Thus jet energy scale corrections should be well under control. A certain amount of hadronic energy is required in order to separate jets from electromagnetic objects in the calorimeter such as electrons or photons. For jets with a considerably large electromagnetic energy fraction standard jet energy scale corrections should not be applied. Thus the selection is restricted to hadronic jets which are easier to handle.

Two representative performance plots for jets are given in figure 6.5 using multi-jet events over a broad $\hat{p}_{\mathrm{T}}$-range from 15 GeV up to 3 TeV. The distributions have been obtained by matching jets at generator level, using the four-vectors of all particles in the detector acceptance except neutrinos as input, to reconstructed jets, using calorimeter deposits as input. The relative resolution plot (left) shows a characteristic curve with increasing precision at higher transverse momenta. The resolution is limited by the resolution of the hadronic calorimeter and ranges from 15% at 60 GeV to smaller than 5% above 1 TeV of transverse jet momentum.
The identification efficiency is almost 100% over the full p_{T}-range of the applied selection

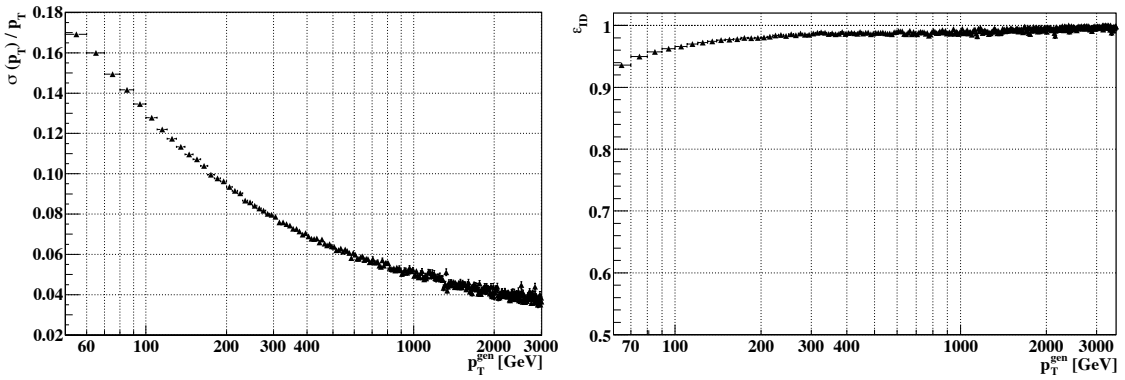

Figure 6.5: *Relative jet transverse momentum resolution* ***(left)*** *and jet identification efficiency* ***(right)*** *as a function of the jet momentum at generator level. The jets have been reconstructed with the SISCone algorithm with* $\Delta R < 0.5$. *For both plots multi-jet events with transverse momenta from* $15 - 3000$ GeV *have been utilized.*

(figure 6.5, right). Also in η the efficiency is flat at almost 100% with only a slight dip in the less sensitive transition region between barrel and endcap.

6.5 Missing Transverse Energy

The almost hermetical coverage of the CMS calorimeter over a wide pseudorapidity range allows a rather precise measurement of the momentum conservation in the transverse plane i.e. perpendicular to the beam direction. Any measured significant transverse momentum imbalance $\not\!\!E_{\mathrm{T}}$ can be considered as signature of weakly interacting particles such as neutrinos or hypothetical dark matter candidates which typically escape undetected.

The missing transverse energy is the magnitude of the vector which balances the vector sum of the uncorrected transverse energy deposits inside the calorimeter towers [151]. Similar to the jets, the missing transverse energy needs to be corrected for various effects. Since muons escape the calorimeters almost undetected their contribution to the $\not\!\!E_{\mathrm{T}}$ needs to be accounted for. Also, any correction which is applied to the jet collection needs to be fed back into the missing transverse energy.

Obviously, the missing energy relies on all detector components and is therefore extremely sensitive to detector malfunctions and small regions which are not instrumented within CMS. Its detailed understanding is a great challenge at the LHC start-up and an absolute pre-requisite for the discovery of $\not\!\!E_{\mathrm{T}}$-based signatures of and beyond the Standard Model.

To have a robust $\not\!\!E_{\mathrm{T}}$ object MUSiC considers only missing transverse energies above a relatively high threshold of

- $\not\!\!E_{\mathrm{T}} > 100$ GeV.

The left plot of figure 6.6 presents the relative $\not\!\!E_{\mathrm{T}}$-resolution as a function of the generated missing transverse energy, using a W+jets sample and events from a supersymmetric benchmark point without requiring a selected lepton or photon. The $\not\!\!E_{\mathrm{T}}$ at generator level is defined by adding up all stable particles within the calorimeter acceptance, excluding

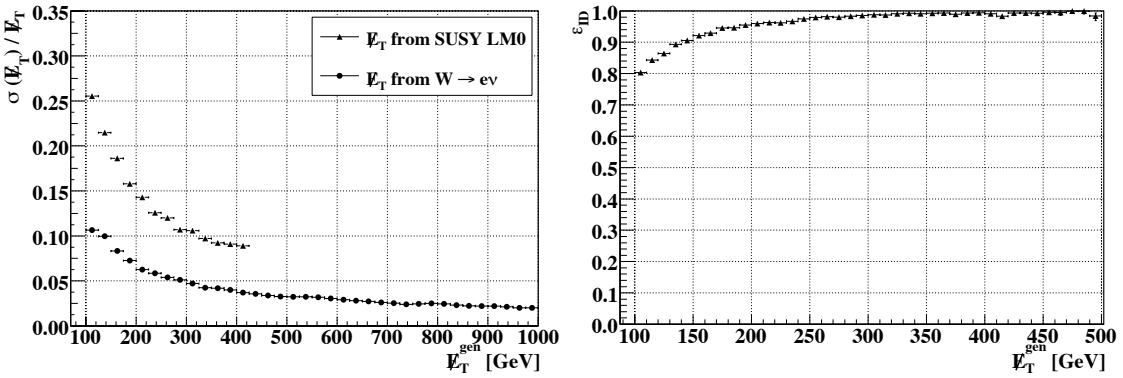

Figure 6.6: ***Left:*** *$\not\!\!E_T$-resolution as a function of the generated missing transverse energy. One curve is obtained using $\not\!\!E_T$ from the supersymmetric benchmark point LM0 (resolution limited by the HCAL), the other is $\not\!\!E_T$ given by W+Jet events ($\not\!\!E_T$ dominated by the electron measurement within the ECAL).* ***Right:*** *$\not\!\!E_T$ identification efficiency as a function of the $\not\!\!E_T$ at generator level utilizing the SUSY LM0 sample.*

neutrinos and weakly interacting particles beyond the SM like supersymmetric neutralinos. In the case of the W+Jet sample the missing transverse energy is dominated by the measurement of the electron, smeared by further jets leaving the hard interaction, the underlying event, the detector acceptance, and noise. The expected $\not\!\!E_T$-resolution ranges from 10% at a $\not\!\!E_T$ of 100 GeV to below 2% for $\not\!\!E_T$ larger than 1 TeV. The $\not\!\!E_T$ resolution is much worse in SUSY events where large jet multiplicities deteriorate the resolution further due to the limited HCAL resolution.

The $\not\!\!E_T$-identification efficiency as a function of the generated $\not\!\!E_T$ is close to 100% with a slight "turn-on" from 80% at 100 GeV to 100% at 300 GeV (see figure 6.6, right).

6.6 Suppression of Instrumental Background

With the arrival of first data various "cleaning" steps will be needed to select "good" runs without detector problems. Also at the level of physics object reconstruction additional criteria are needed in order to minimize instrumental background from "fakes". This cleaning mainly refers to the removal of duplicate objects and the ambiguous interpretation of objects in the detector. For example an ECAL supercluster can be interpreted as an electron as well as a photon. The listed cleaning steps are carried out in the following sequence:

- Muon candidates which are closer than $\Delta R < 0.2$ to each other are cleaned, keeping only the one measured best (smaller normalized χ^2). The cut is designed to remove ghost muons and other sources of duplicate muons.
- Electron candidates which are closer than $\Delta R < 0.2$ to each other, and which share either the inner track or the supercluster seed are cleaned, keeping only the more energetic one.
- Photon candidates which are closer than $\Delta R < 0.2$ to each other and which share the supercluster seed are cleaned, keeping only the more energetic one. Also photon

candidates closer than $\Delta R < 0.2$ to an already selected electron are removed if the photon has the same supercluster seed as the electron. This removes the ambiguity imposed by the fact that all superclusters can be interpreted as electrons as well as photons. Thus well-measured electrons receive a higher priority than photons.

- Jet candidates closer than $\Delta R < 0.2$ to an already selected electron or photon are removed to avoid an overlap of those collections.

So far no e/μ separation cut is included but could be added in the future.

6.7 High Level Trigger

The choice of the trigger menu used in this analysis is driven mainly by the requirement to combine triggers with a prescale factor of unity (or at least triggers which use the same L1-definition) and high level triggers which are expected to be "standard" at the LHC start-up and which are therefore commonly used and well-understood.
A logical "OR" of various high level triggers is used:

- single muon or di-muon HLT (both with and without isolation)
- single electron or di-electron HLT (both with and without isolation)
- single photon or di-photon HLT (both with and without isolation)
- single high and very high energy e/γ trigger.

Table 6.3 shows the full list of triggers used in this analysis including a detailed description and their expected rates for two luminosity scenarios. Note that data taken with the CMS detector will be delivered in trigger streams which subsume triggers of identical objects similar

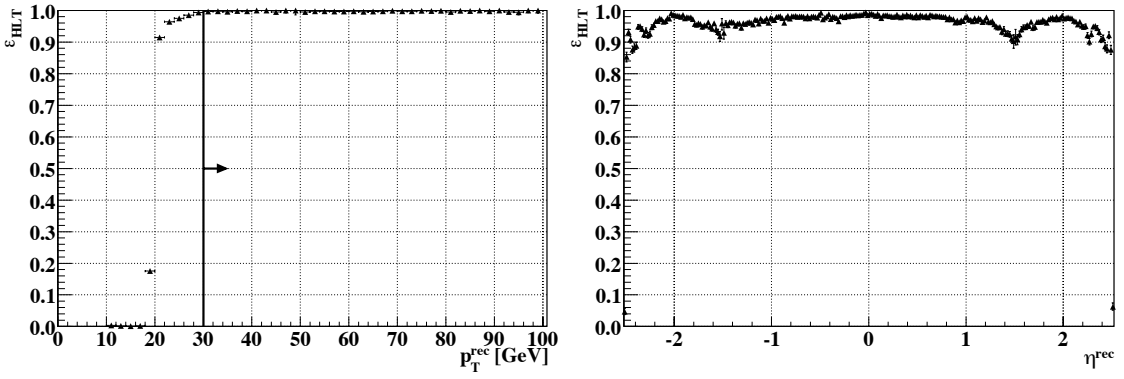

Figure 6.7: ***Left:*** *Trigger turn-on for the photon triggers as a function of the reconstructed photon momentum with respect to the selected photon events. The trigger efficiency of the "OR" of all photon triggers has been calculated using a Photon+Jet sample matching the reconstructed photon to the fired trigger object.* ***Right:*** *Electron trigger efficiency as a function of the reconstructed pseudorapidity with respect to the selected electron events. Electrons from a W+Jet sample with typical momenta from resonantly produced Ws have been matched to the trigger candidate.*

HLT Path	Requirements	Rate (Hz)	
		"8e29"	"3e30"
Chosen Muon HLT Paths			
Single Isolated μ *IsoMu15*	Input: L1 μ with $p_T > 10$ GeV Threshold: $p_T > 15$ GeV	0.01	0.03
Single μ *Mu15*	Input: L1 μ with $p_T > 10$ GeV Threshold: $p_T > 15$ GeV	0.06	0.23
Double Isolated μ *DoubleIsoMu3*	Input: 2 L1 μ's with $p_T > 3$ GeV Threshold: $p_T > 3$ GeV	0.04	0.16
Double μ *DoubleMu3*	Input: 2 L1 μ's with $p_T > 3$ GeV Threshold: $p_T > 3$ GeV	0.06	0.22
Chosen Electron HLT Paths			
Single Isolated e *IsoEle18_L1R*	Input: L1 e/γ with $p_T > 15$ GeV Threshold: $p_T > 18$ GeV Track Isolation	0.10	0.35
Single e *Ele15_LW_L1R*	Input: L1 e/γ with $p_T > 10$ GeV Threshold: $p_T > 15$ GeV	4.17	15.13
Double Isolated e *DoubleIsoEle12_L1R*	Input: L1 e/γ with $p_T > 10$ GeV Threshold: $p_T > 12$ GeV Track-based isolation	0.00	0.00
Double e *DoubleEle10_LW* *OnlyPixelM_L1R*	Input: 2 L1 e/γ with $p_T > 5$ GeV Threshold: $p_T > 10$ GeV Matching Pixel requirement	1.57	5.71
Chosen Photon HLT Paths			
Single Isolated γ *IsoPhoton20_L1R*	Input: L1 e/γ with $p_T > 15$ GeV Threshold: $p_T > 20$ GeV # tracks isolation cut	0.21	0.76
Single γ *Photon25_L1R*	Input: L1 e/γ with $p_T > 15$ GeV Threshold: $p_T > 25$ GeV	1.38	5.01
Double Isolated γ *DoubleIsoPhoton20_L1R*	Input: 2 isolated L1 e/γ, $p_T > 8$ GeV Thr: $p_T > 20$ GeV, # tracks isolation	0.02	0.06
Double γ *DoubleIsoPhoton20_L1I*	Input: 2 L1 e/γ with $p_T > 10$ GeV Thr: $p_T > 15$ GeV, # tracks isolation	0.00	0.00
Chosen High Energy e/γ HLT Paths			
Single High E_T e/γ *EM80*	Input: L1 e/γ with $p_T > 15$ GeV Thr: $p_T > 80$ GeV, # tracks isolation	0.00	0.00
Single Very High E_T e/γ *EM200*	Input: L1 e/γ with $p_T > 15$ GeV Threshold: $p_T > 15$ GeV Various loose cuts	0.00	0.00

Table 6.3: *Details on the High Level Triggers used within this analysis. The rates are estimated from the detector simulation within the "HLT exercise" [78] for two different luminosity scenarios ($\mathcal{L} = 8 \cdot 10^{29}$ cm^{-2} s^{-1} and $\mathcal{L} = 3 \cdot 10^{30}$ cm^{-2} s^{-1}).*

to the list given above. After the application of all selection criteria, MUSiC will merge all streams into a single dataset, avoiding double counting in events where triggers from more than one stream have fired.

For each of the three different trigger objects a representative distribution is given in figures 6.7 and 6.8. For each of the plots the efficiency is defined as the percentage of events which pass the selection including the topology cut of at least one electron, muon or photon that also are accepted by an "OR" of the considered high level trigger bits.

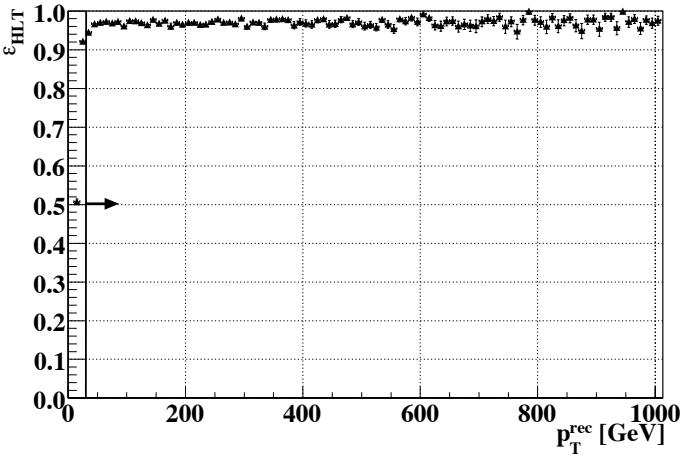

Figure 6.8: *Muon trigger efficiency with respect to the selected muon events as a function of the reconstructed muon momentum utilizing the muon with a largest momentum from a broad Drell-Yan spectrum for invariant masses from* 40 GeV $< m_{\mu\mu} <$ 2.5 TeV.

The left plot of figure 6.7 represents the so-called turn-on curve for the considered photon triggers as a function of the leading photon momentum. The efficiency has been determined utilizing a Photon+Jets sample relaxing the photon momentum cut from 30 GeV to 10 GeV. Although the p_{T} threshold of the single (relaxed) photon trigger is at 20/25 GeV the trigger is already fully efficient at the lowest momenta considered within this analysis. In this case the efficiency of more than 95% is completely dominated by the single relaxed photon trigger.

An archetypical plot for the electron trigger efficiency is given in the right plot of figure 6.7. Here the efficiency as a function of the reconstructed electron pseudorapidity is drawn for the leading p_{T} electron from a W+Jets sample with boson masses at the resonance peak. The efficiency is around 95% – 100%, slightly depending on η. Especially in the transition region between barrel and endcap ($|\eta| \sim 1.5$) and at the tracker acceptance boundaries ($|\eta| \sim 2.4$) the offline reconstruction is superior to the high level trigger reconstruction leading to minor efficiency losses.

The efficiency of the muon trigger as a function of the leading muon momentum is given in figure 6.8 for momenta up to 1 TeV. The trigger efficiency for Drell-Yan events with invariant di-muon masses from 40 GeV to 2.5 TeV is around 95% over the whole p_{T}-range.

Figure 6.9 shows the rate of fired triggers for events from the supersymmetric benchmark point LM1. The rate has been defined after applying all selection cuts with respect to events which contain at least one muon, electron or photon. The detailed trigger efficiencies are

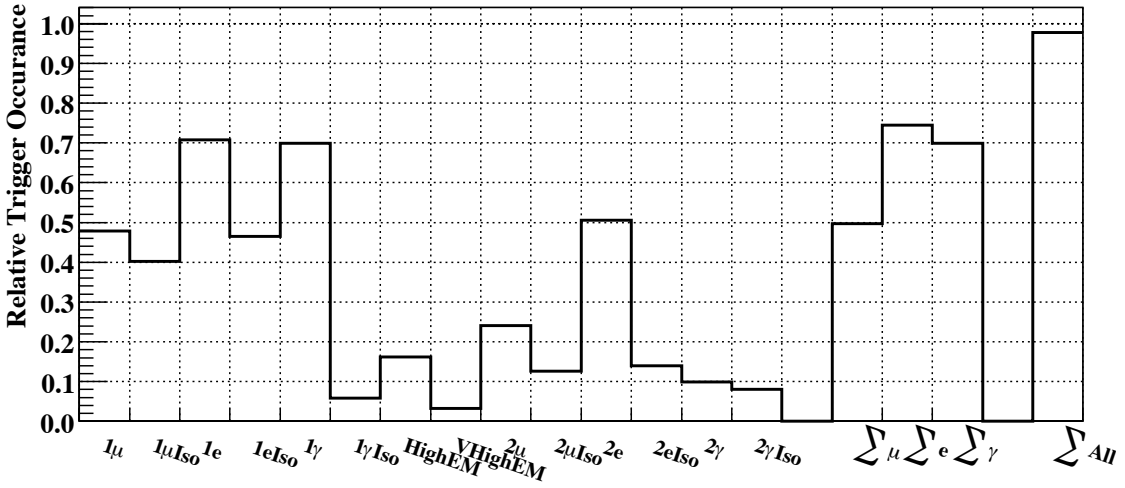

Figure 6.9: *Relative trigger rates for the different triggers used within the MUSiC analysis (see table 6.3) for events which fulfill the topology criterion of at least one electron, muon or photon utilizing the supersymmetry benchmark point LM1. Bins labelled with Σ represent an "OR" of the corresponding triggers.*

stated below. They reflect which of the LM1 events that contain e.g. a reconstructed muon are triggered by an muon trigger. The list contains statistical uncertainties only:

- Muon: $\varepsilon_{\mathrm{HLT}} = 94 \pm 1.6\%$
- Electron: $\varepsilon_{\mathrm{HLT}} = 98 \pm 1.6\%$
- Photon: $\varepsilon_{\mathrm{HLT}} = 93 \pm 3.8\%$
- Muon || Electron || Photon: $\varepsilon_{\mathrm{HLT}} = 98 \pm 1.1\%$.

Chapter 7

The Implementation of MUSiC

This chapter describes the details of the search algorithm which is used to perform a broad data – Monte Carlo comparison. It discusses the differences between a common signal-driven analysis and a model-independent approach and the issues one has to solve. A dedicated part will be devoted to trial factors which play an important role when testing many regions in many distributions for the same hypothesis. Also the importance of systematic uncertainties and their incorporation in the search algorithm is discussed in detail.

7.1 Input Variables to the Search Algorithm

As outlined in the previous chapters, the events have been processed and physics objects satisfying the selection criteria have been identified. The composition of the event, i.e. the number of muons, jets, and other objects, determines to which event classes it is assigned. At the present three distributions are investigated for each event class, thus limiting the number of distributions looked at and focusing on distributions which seem to be promising for spotting new physics, but also detector or MC related deviations:

- **Scalar sum of the transverse momentum** $\sum p_{\mathrm{T}}$ of all physics objects.
  For example for the class $2e$ 1jet $\not{E}_{\mathrm{T}} + X$ one calculates:

$$\sum p_{\mathrm{T}} = p_{\mathrm{T}}(\mathrm{e}_1) + p_{\mathrm{T}}(\mathrm{e}_2) + p_{\mathrm{T}}(\mathrm{jet}_1) + \not{E}_{\mathrm{T}}.$$

- **Invariant mass** M_{inv} of all physics objects. For classes with missing transverse energy the transverse invariant mass M_{T} is calculated. Utilizing the four-vectors p of the objects the mass is calculated e.g. for the class 2μ 1jet $+ X$ via

$$M_{\mathrm{inv}} = \sqrt{p^2(\mu_1) + p^2(\mu_2) + p^2(\mathrm{jet}_1)}.$$

  The M_{T} calculation is performed with the four-vectors which contain the energy projected to the transverse plane ($E \cdot \cos\theta$) and the z-component set to zero.

- For classes with **missing transverse energy** $\not{E}_{\mathrm{T}}$ this variable is investigated separately.

The $\sum p_{\mathrm{T}}$ distribution is the most general quantity to be checked. The invariant mass has an obvious advantage for new particles produced as resonances like new heavy gauge bosons. Since many models beyond the SM aim to give a candidate particle explaining dark matter in the universe, this particle would lead to a considerable amount of missing energy in the event. An obvious example would be the lightest supersymmetric particle in supersymmetric extensions of the Standard Model: here $\not{E}_{\mathrm{T}}$ is known to be prominent for separating supersymmetric events from the Standard Model background. Still, this quantity will be hard to control and understand at the beginning of data taking and thus the use of this quantity might be challenging.

While in $\sum p_{\mathrm{T}}$ a model independent scan of all classes will be performed with the first data, this is not clear at the moment for M_{inv} and $\not{E}_{\mathrm{T}}$. Also, in principle the implementation of even further variables can be done easily, if desired. However, one should always keep in mind that the increase of the number of distributions weakens the global sensitivity due to the higher trial factor (see below).

All distributions are input to the MUSiC algorithm (similar to the H1 analysis [125]) which scans them systematically for deviations, comparing the Standard Model expectation (Monte Carlo prediction) with the measured data.

7.2 Prelude: Statistical Interpretation of Search Results

The aim of each search for new physics is the quantification of the deviation (or lack thereof) from the Standard Model expectation. Due to the nature of the measurement process and the probabilistic foundation of quantum field theory, results can only be drawn on a statistical basis. In a typical high-energy new physics search one or more so-called final variables are chosen which are expected to be most sensitive to deviations from the Standard Model. These variables or distributions thereof are taken as input to a statistical test.

The final variables are designed to have a high separation power for two distinct hypotheses which will be confronted to data: One is the *null hypothesis*, representing the model which is in full agreement with the Standard Model, i.e. without a new physics signal. The other is the *alternative* or *test hypothesis* considering a model where in addition to the SM prediction a distinguishable new physics effect is present. The hypotheses are also referred to as *background-only* and *signal + background hypothesis*, respectively.

A test statistic has to be constructed that is used to quantify the degree to which the data are consistent with the two hypotheses. In general the final variable depends on a variety of parameters such as the luminosity, reconstruction and detection efficiencies, theoretical cross section predictions. Those parameters quoted as nuisance parameters are not of immediate interest, but are important ingredients to the hypothesis test. The uncertainty of the nuisance parameters known as systematic uncertainties in high energy physics need to be taken into account and generally degrade the power of the test to distinguish the null and the alternative hypothesis. Deviations are classified by a significance estimator usually quoted in Gaussian standard deviations which quantifies the significance level of the deviation from a certain hypothesis. Typically a deviation of 5σ from the background-

only hypothesis is interpreted as a discovery, while an 95% exclusion level with respect to the signal+background hypothesis is quoted if no signal is spotted.

The model-independent search presented here deviates from the traditional model-driven searches due to the lack of a signal and thus an alternative hypothesis. Nonetheless a hypothesis test can be performed, which quantifies the degree of agreement of the data with the null hypothesis. Since the present analysis is a feasibility study without data, pseudo-data are generated involving SM expectations, but also new physics benchmark channels mimicking data which agree or deviate from the null hypothesis.

Despite the long history of statistics, hypothesis tests and their interpretation remain a delicate topic: the Bayesian foundation of statistics as degree of belief depends on a prior probability introducing a subjective element in a somewhat arbitrary manner [152]. This definition of probability which provides the basis of our daily life decisions is opposed to the Frequentist's interpretation of probability as a relative frequency familiar to us in role dicing. The interpretation of Frequentist confidence intervals, however, is unintuitive and often misinterpreted as a Bayesian statement about the theory given the data.

Both interpretations of probability are mathematically perfectly well-defined, but their appliance and interpretation in the area of hypothesis tests lacks a clear and unique answer although in practice Bayesian credible intervals and Frequentist confidence intervals tend to converge in the limit of the central limit theorem.

Nowadays several methods exist to construct confidence intervals motivated by either Frequentist or Bayesian statistics. Even mixtures of both statistics are pragmatically taken into account and are accepted by both communities as long as the properties of the hypothesis tests are in agreement with the Bayesian and Frequentist foundation. E.g. a Frequentist expects for the background only hypothesis in not more than a fraction of $2 \cdot 10^{-7}$ of the repeated pseudo-experiments without signal, a deviation of more than 5σ. A method which fulfills this criterion is said to have coverage in contrast to overcoverage (the method claims 5σ but its actually more, conservative) or undercoverage (the method claims 5σ but its actually less, liberal).

MUSiC tries to cope with these mentioned issues in multiple ways. For each hypothesis test all values for the input variables are stated so that everyone can in principle redo the significance calculation with another hypothesis test. In addition several different methods for the confidence level calculation have been implemented to check and validate the approximate correctness of the significance. Still one should notice that the exact significance is not of great interest for this analysis representing an alarm system for deviations. In case of an interesting deviation a dedicated analysis would be performed anyway.

Having all that in mind the reader will find the detailed steps of the hypothesis test in the following sections.

7.3 The Search Algorithm

The aim of the search algorithm is to spot the region with the largest statistical significant deviation. In order to do so one needs to carefully incorporate systematic uncertainties

and quantify the discrepancy by taking the trial factor into account. Logically the search algorithm can be separated into two parts, which are discussed separately in the following.

- Step 1: Find the region with the most signification deviation per event class and distribution (Region of Interest).
- Step 2: Take the trial factor or look-elsewhere-effect for looking at many regions into account.

7.3.1 Spotting the Region of Interest

Each connected bin region (see figure 7.1 for illustration) is considered within the distributions like $\sum p_{\mathrm{T}}$ within each event class. This can be single bins (e.g. bin 10 or bin 200), broad regions (e.g. bins $3-100$ or bins $300-305$) or even the whole distribution. The combination of unconnected bins, e.g. bin 20, bin 100, and bin 314, as one region is not considered meaningful.

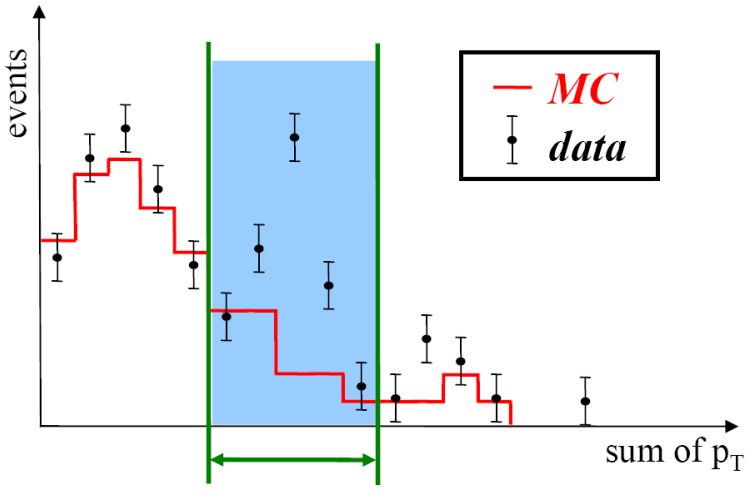

Figure 7.1: *Illustration of a connected bin region within a kinematic distribution.*

For each connected region, a counting experiment is performed, adding up the various expected Monte Carlo contributions (N_{SM}) and comparing this sum to the amount of measured data (N_{data}). In addition to these two numbers also the uncertainty of the prediction $\delta(N_{\mathrm{SM}})$ is used, i.e. the combined systematic and statistical uncertainties of the simulated events contributing to this specific region. Then a Poisson probability is computed, determining how likely the prediction fluctuates up (down) to or above (below) the number of events seen in the data in the case of an excess (deficit). The systematic errors, taking correlations into account, are included using a convolution with a Gaussian:

$$p = p_N = \begin{cases} \sum_{i=N_{\mathrm{data}}}^{\infty} A \cdot \int_0^{\infty} \mathrm{d}b \, \exp\left(\frac{-(b-N_{\mathrm{SM}})^2}{2(\delta N_{\mathrm{SM}})^2}\right) \cdot \frac{e^{-b}\, b^i}{i!} & \text{if} \quad N_{\mathrm{data}} \geq N_{\mathrm{SM}} \\ \sum_{i=0}^{N_{\mathrm{data}}} A \cdot \int_0^{\infty} \mathrm{d}b \, \exp\left(\frac{-(b-N_{\mathrm{SM}})^2}{2(\delta N_{\mathrm{SM}})^2}\right) \cdot \frac{e^{-b}\, b^i}{i!} & \text{if} \quad N_{\mathrm{data}} < N_{\mathrm{SM}} \end{cases} , \quad (7.1)$$

where A ensures the normalization. From all the possible combinations of connected bins, the region with the smallest p-value ($p_{\mathrm{min}}^{\mathrm{data}}$) is chosen. This is the place in the distribution where the biggest discrepancy between data and Monte Carlo prediction is found. It is called the **Region of Interest**.

This effective approach is sensitive to an excess of data as well as a deficit. It can detect large single bin fluctuations as well as possible signals spread over a large part of the distribution. The bin width is variable and chosen dynamically in multiples of 10 GeV taking the expected detector resolution for the different objects into account. The resolution is assumed to be dominated by the object which can be measured worst. For example the bin width of the $1e + 1\mathrm{jet}$ $\sum p_{\mathrm{T}}$ distribution is given by the resolution of the jets. This binning ensures that the algorithm does not pick up effects which cannot be resolved by the detector.

One should stress the importance of including the uncertainty on the estimate of the Monte Carlo simulation into the probability definition. In this way the p-value gives the probability for the background to fluctuate up to the data and further, given the intrinsic uncertainties of the MC estimate. One can easily assign large errors to the value N_{SM} if the Monte Carlo events are expected to not describe the data well in a specific part of the phase space. Still, this does not necessarily spoil the potential to reveal deviations: If one expects a 100% uncertainty in some exotic final state where one cannot trust the Monte Carlo prediction, some new physics signals such as spectacular mini black hole signatures might well lead to discrepancies far exceeding this large uncertainty.

A detailed discussion of the significance estimator and alternative implementations within MUSiC are given in section 7.5. The systematic uncertainties and their determination are explained further in chapter 7.7. The different uncertainty contributions are assumed to be Gaussian and uncorrelated (e.g. luminosity uncertainty and jet energy scale uncertainty) and are thus added in quadrature. Correlations within a single uncertainty like the luminosity uncertainty between simulated samples, are carefully included in the uncertainty estimate. The individual contributions will be discussed in detail in section 7.7. Ultimately the total systematic uncertainty can be expressed as:

$$\delta N_{\mathrm{SM}} = \sqrt{\sigma_{\mathrm{stat}}^2 + \sum_i \sigma_{\mathrm{i,syst}}^2}\,, \tag{7.2}$$

where σ_{stat} represents the statistical uncertainty given the limited MC-statistics of the various samples used. The sum runs over all systematic uncertainties discussed in chapter 7.7.

7.3.2 Taking the Trial Factor into Account

It is important to understand that the statistical estimator p alone is not sufficient to claim any evidence for a signal. A statistical penalty factor has to be applied to account for the large number of investigated regions (connected bin combinations). This is done in the second step of the algorithm, determining the **event class significance** (per distribution)

of the deviation found in the first step:
Toy Monte Carlo experiments are performed, assuming the background-only hypothesis. Therefore hypothetical data histograms are created numerous times by varying the Monte Carlo prediction for each bin according to its statistical and systematic uncertainty. Again correlations within single uncertainty contributions have to be accounted for when creating the pseudo data. These hypothetical data are then fed again into the first step of the algorithm and compared to the Monte Carlo mean (results in $p_{\text{min}}^{\text{SM}}$). Again all possible connected regions are examined, not only the Region of Interest from the initial step 1. The event class significance of the deviation is defined as:

$$\widetilde{P} = \frac{\text{Number of SM-only toy experiments with } p_{\text{min}}^{\text{SM}} \leq p_{\text{min}}^{\text{data}}}{\text{Total number of toy experiments}} \,. \tag{7.3}$$

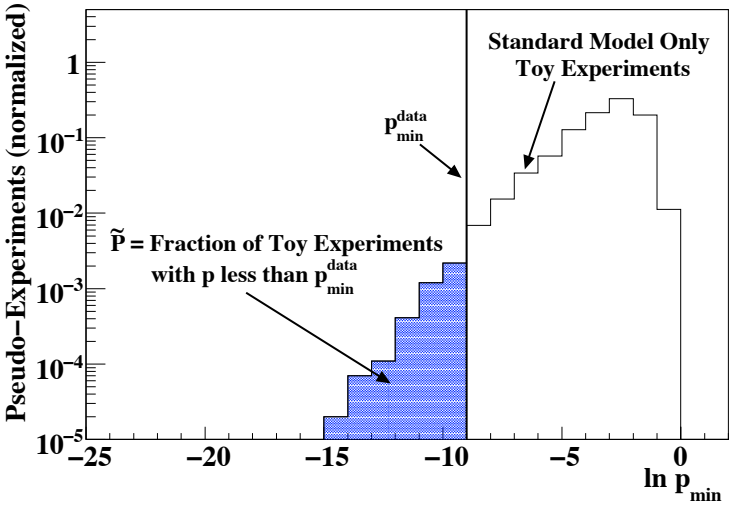

Figure 7.2: *Illustration of the $\tilde{P}$-calculation. $\tilde{P}$ marks the fraction of background-only events which have a more prominent deviation than the data $p_{\text{min}}^{\text{data}}$.*

The value of $\widetilde{P}$ as illustrated in figure 7.2 is the fraction of background-only toy experiments where a deviation even bigger than the one observed in the data is found. Performing these pseudo-experiments one jitters the Standard Model expectations and tests for signal-like fluctuations of the Standard Model. These fluctuation may appear in all regions considered within the algorithm, not only in the Region of Interest. The $\widetilde{P}$ can directly be translated into standard deviations Z (see figure 7.3) and is comparable to the widely used CL_b. Since MUSiC is sensitive to an excess of data as well as a deficit, a two-sided Gaussian is used for this translation. In principle the trial factor of looking at many regions within one distribution can be calculated analytically using binomial statistics (see section 7.6). However, this would neglect all correlations between the different regions which are automatically taken into account when determining the effect of the trial factor via toy experiments. The disadvantage of this approach is the huge amount of computing power required to determine e.g. a 5σ effect, which needs at least $5 \cdot 10^6$ toy experiments. Therefore it might be desirable for a fast analysis turn-around with the first LHC data to estimate the effect of the trial factor analytically and only switch to the more precise Monte Carlo method upon an improved understanding of the data.

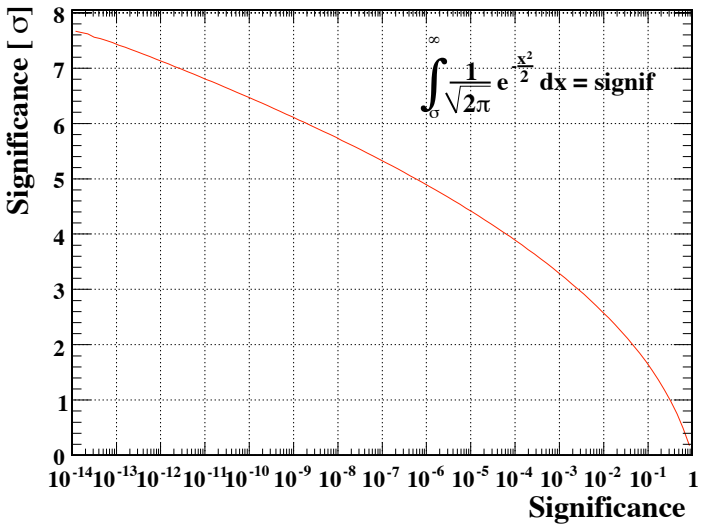

Figure 7.3: *Translation of significance $\widetilde{P}$ into number of standard deviations σ.*

7.4 Sensitivity Study with Simulated Events

Since the LHC has not started it is clear that there are no pp-data (N_{data}) yet to compare with the Monte Carlo prediction. Still one can pick representative models beyond the Standard Model and test the sensitivity of MUSiC with them. Instead of only producing pseudo-data for the background-only hypothesis one can also create toy data as input to step 1 assuming signal + background, i.e. by adding a signal distribution on top of the SM ones. In this way one can repeat several pseudo-CMS experiments and determine the expected event class significance of a possible signal present in the data. Figure 7.4 illustrates this procedure, using the event class $1e\ 5\text{jet} + X$ as an example: The green curve represents the pseudo-experiments where signal (LM4) plus background are assumed. With data this would correspond to a single line. The red curve on the other hand displays the multiple repetition of the SM expectation including its uncertainties, representing step 2 of the algorithm. The p and $\widetilde{P}$ values stated in the plots refer to the median of the left curve, integrating the background-only curve beyond this median p_{min}. The interpretation of the two curves is clear: In the case that they are well-separated, $\widetilde{P}$ will be quite low and a discovery is easy, as shown in the left plot where no systematic uncertainties are assumed. By the inclusion of systematic uncertainties in the algorithm, the two hypotheses move closer to each other and only a deviation less than 3σ ($\approx 10^{-3}$) remains. This also underlines the importance of the implementation of systematic uncertainties into MUSiC which will be discussed in section 7.7.

Producing pseudo-data

Testing the significance of the deviation is done by dicing hypothetical data histograms. This means that one changes the true value N_{SM} slightly to reflect the inherent statistical and systematic uncertainties. Of course, the assumption that these uncertainties

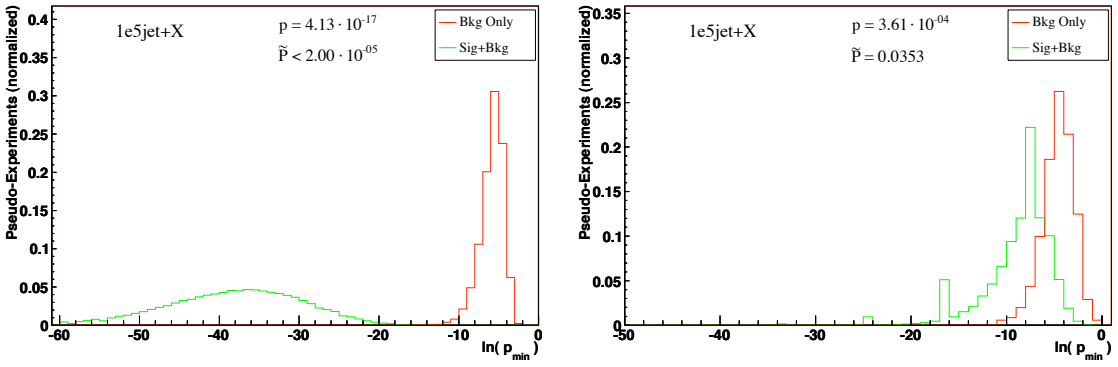

Figure 7.4: *Signal plus background and background-only hypotheses for an LM4 event class, on the left without systematic uncertainties and on the right with all uncertainties included. The striking difference between both plots shows the importance of systematic uncertainties.*

are both well-understood and realistic is crucial for this procedure. On the other hand huge deviations found in numerous event classes of the first LHC data could indicate that some uncertainties have been underestimated and/or additional uncertainties have to be included.

In order to be able to separate deviations caused by new physics from background-only fluctuations the uncertainties have to be included in a way similar to a real measurement of the CMS detector. Thus correlations between bins and simulated samples are important. These will be discussed in detail in section 7.7, but some general comments on the implementation are done here as well.

The basis for the dicing is the significance estimator p in equation (7.1), even though the actual dicing process is divided into several parts with respect to all uncertainty contributions. It is essential that contributions which are *statistically independent* can be *decoupled* and diced separately. There are three main dicing-contributions for each hypothetical data histogram:

- the assumed **systematic uncertainties** as part of the Gaussian convolution
- the **statistical uncertainties** of the MC datasets as part of the Gaussian convolution
- the **Poisson probability** to account for the actual measurement.

An example for a systematic uncertainty could be the uncertainty on the luminosity estimate. Assuming a 10% uncertainty all bins and all simulated samples are correlated with respect to this uncertainty. Thus for each set i of pseudo data a single number $\sigma_i(\text{lumi})$ is generated assuming a Gaussian with a mean of $\mu = 0$ and a width of $\sigma = 0.1$. This variable number could be -3% in one pseudo-experiment and $+10\%$ in another one, and all bins of all samples are scaled with the corresponding factor. Thus magnitude and direction of the uncertainty is preserved for all bins and all MC-contributions. Since the individual systematic uncertainties are assumed to be uncorrelated, similar considerations can be made also for them.

Figure 7.5 illustrates the correctness of the dicing procedure. Here the sum of the SM background is shown together with the total systematic uncertainty (shaded area). The data points correspond to the mean after many repetitions of the background-only hypothesis. The error bars correspond to the width of the variation for the many pseudo-data sets. As expected, one can see nicely that the data points match the mean expectation of the Monte Carlo and that the error bars reflect the total uncertainty estimate.

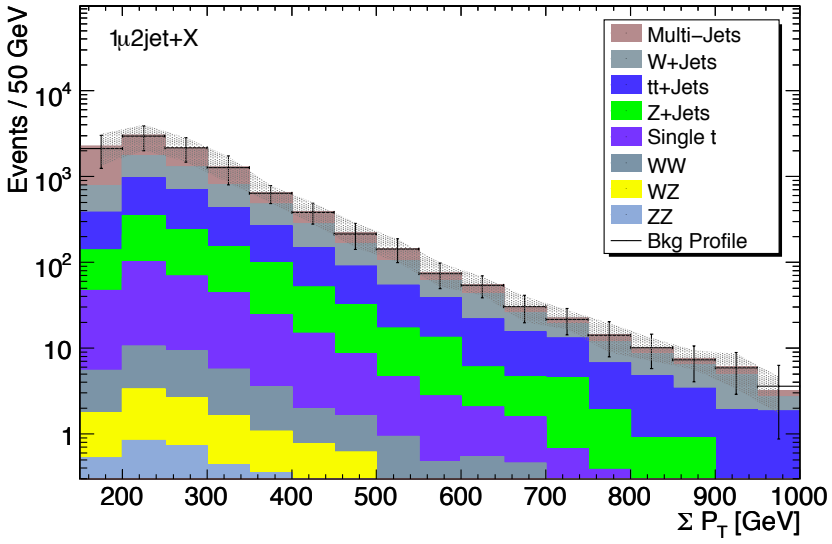

Figure 7.5: *Sum of SM backgrounds and assumed systematic uncertainties (shaded area) in comparison with the distribution of the numerous pseudo-data sets. Data points correspond to the mean of these sets, the error bars to the width of the variation. This closure test shows that the distribution is diced correctly according to the assumed uncertainties.*

7.5 Discussions Concerning the Hypothesis Test

The definition of the significance estimator p given in equation 7.1 represents a Bayesian-Frequentist hybrid. The Poisson distribution describes the statistical fluctuation (Frequentist), while the Gaussian reflects the Bayesian prior integrating out the nuisance parameters (systematic uncertainties). As the inclusion of systematic uncertainties within significance estimators has not been solved under general circumstances within the professional statistics community, this hybrid method represents a reasonable and practical ansatz. It has good (Frequentist) properties over a broad range when applied to a problem with a mean background expectation and a Gaussian systematic uncertainty. Of course approximating the uncertainties by a Gaussian is a strong assumption which may not be true in all cases. Thus this should be understood as a pragmatic solution and the reader should be aware of possible deficits due to non-Gaussian tails. Still, one should keep in mind that correct Frequentist coverage cannot be guaranteed even for this Gaussian-mean background problem for all possible parameters. Especially when the Gaussian tails have significant contributions in the unphysical region of event numbers smaller than zero the method becomes unreliable and it might be more adequate to e.g. use a lognormal prior as discussed below.

A detailed discussion of the properties of such a Bayesian-Frequentist hybrid as well as comparisons to other methods can be found in [153].

One should emphasize that in the context of MUSiC the focus is not to give very precise significances but to act as an alarm system detecting interesting deviations. Since the number of data events, expected Monte Carlo events and its corresponding uncertainty is always known and stated, cross-checks using alternative statistical methods are possible and desired. Therefore further significance estimators have been studied within MUSiC, but the amount of precise estimators suitable for a generic model-independent search are rare due to the stringent requirements:

- No-Signal: The estimator must work without the assumption of a signal. A prominent estimator which does not fulfill this criterion is the CL_S method [154]. Here even the background-only test statistic depends on an assumed signal.
- Generality: Since MUSiC is interested in excesses as well as deficits, the estimator has to support both scenarios.
- Speed: The code to evaluate the significance must be sufficiently fast to allow for the calculation of the significance for the several hundreds of distributions and up to several thousand regions within these distributions in a reasonable time.
- Simplicity: The method must be generally applicable and simple e.g. it cannot involve fitting of a complex probability density function.
- Coverage: Approximate coverage for all possible scenarios from regions with many events and small uncertainties to rarely populated regions with large uncertainties should be guaranteed.

Apart from the estimator given in equation 7.1, two other methods have been tested within the MUSiC framework. These are introduced briefly in the following.

7.5.1 Alternative Significance Estimator I

In [153] a method called Z_{Bi} is promoted since it is based purely on Frequentist assumptions using products of Poisson probabilities and shows good performance in many cases. In case of an on/off-problem like given in gamma ray astronomy where one observes n_{on} events while looking at the source (signal + background) and n_{off} events off source (background only), the estimator can be written as:

$$p_{Bi} = B\left(n_{\text{on}}/(n_{\text{on}} + n_{\text{off}}), n_{\text{on}}, n_{\text{off}} + 1\right), \tag{7.4}$$

where B denotes the incomplete beta function. The problem can be translated to the MUSiC case which is closer to the gaussian-mean background problem, by a rough estimate using $n_{\text{on}} = N_{\text{data}}$ and $n_{\text{off}} = (N_{\text{SM}}/\delta N_{\text{SM}})^2$. This estimate leads to overcoverage when applied to a Gaussian-mean background problem especially when the background has a large uncertainty. This can also be seen in figure 7.6 where the p-values from both methods are compared. As an input for the comparison typical numbers from a scan of the

exclusive event classes where the pseudo-data have been supplemented with SUSY LM4 are used. One can clearly see a correlation between the results of both statistical methods. However, the purely Frequentist estimator Z_{Bi} is always more conservative than the Bayesian-Frequentist hybrid Z_{N} defined by equation 7.1, confirming the results in [153].

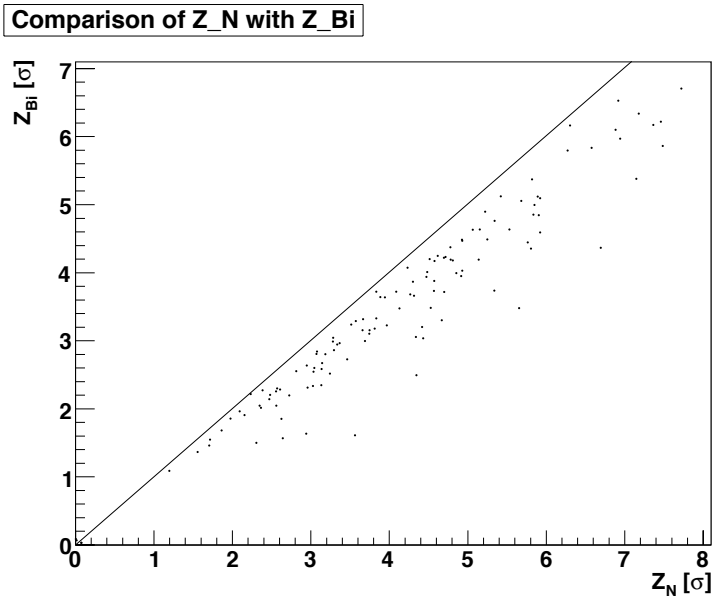

Figure 7.6: *Comparison of p-values computed by two different statistical methods.*

7.5.2 Alternative Significance Estimator II

The significance estimator in equation 7.1 assumes the systematic uncertainties to be Gaussian. Following the usual HEP assumption of Gaussian uncertainties for the different uncertainty contributions (e.g. on cross sections), their combination is again Gaussian as long as the contributions are combined in a sum (sum of Gaussians are Gaussian) or one of the Gaussian uncertainties dominates. However, if several uncertainties with similar sizes have to be combined multiplicatively this is inappropriate. The product of Gaussian probability density functions results in a log-normal distribution. When looking at the average expected number of events for a certain process with cross section σ, recorded within a given integrated luminosity $\mathcal{L}_{\mathrm{int}}$ and efficiency ϵ

$$N_{\mathrm{events}} = \mathcal{L}_{\mathrm{int}} \cdot \sigma \cdot \epsilon, \tag{7.5}$$

the replacement of the Gaussian prior with a log-normal one seems a reasonable ansatz for the propagation of the uncertainties . Not being the focus of this work the approach is only discussed briefly here, but details for the evaluation of the method within MUSiC are given in [155].

Figure 7.7 shows for some selected parameter values log-normal distributions based on the parametrization

$$f_{\mathrm{LN}}(x; b_0, k) = \frac{1}{\sqrt{2\pi}\ln^2 k} \cdot \frac{1}{x} \cdot \exp\left(\frac{-\ln^2(x/b_0)}{2\ln^2 k}\right) \tag{7.6}$$

The parameter b_0 defines the median of the log-normal distribution, while k is related to the width. For k close to one the shape of the log-normal probability density function (pdf) is similar to a Gaussian with mean $\mu = b_0$ and variation $\sigma = b_0 \cdot (k-1)$.

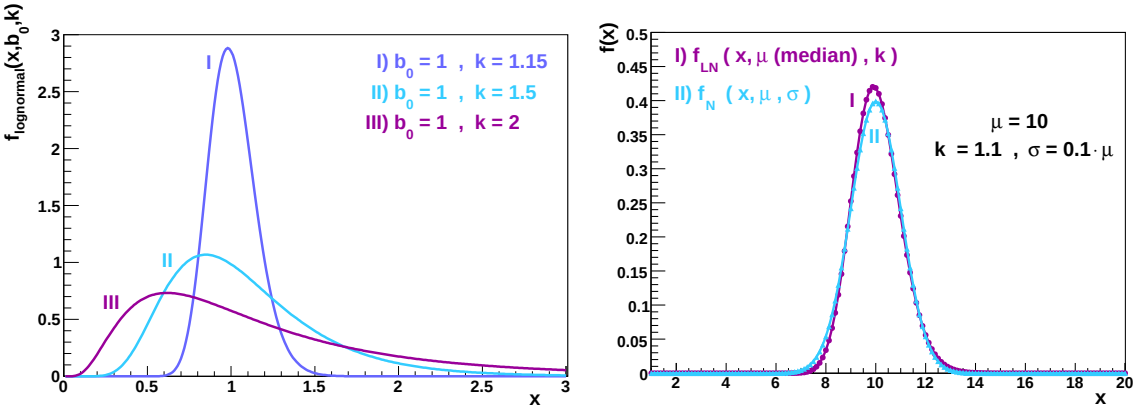

Figure 7.7: ***Left:*** *Log-normal distributions with median μ and different widths k.* ***Right:*** *Comparison of a log-normal distribution with a Gaussian.*

Utilizing the fact that products of log-normal pdfs are again log-normal distributions, the single uncertainty contributions included within the MUSiC search algorithm are approximated by a log-normal pdf and combined into a total log-normal distributed uncertainty. The combination of uncertainties as given in 7.2 is therefore replaced by a product.

In general the log-normal has longer "tails" and thus the corresponding significance estimator is more conservative than the Gaussian prior based estimator (see table 7.1 for an example). The log-normal distribution has another advantage: While the Gaussian prior is truncated at zero which leads to unphysical pdfs for small background values and large uncertainties, the log-normal prior converges smoothly to zero for small background expectations and arbitrary uncertainties. The estimator has good Frequentist coverage as shown in [155] and is especially superior to the Gaussian-based estimator when investigating log-normal distributed backgrounds.

$N_{\mathrm{SM}} = 50$, $\delta N_{\mathrm{SM}} = 20\%$			
N_{data}	Z (no uncertainty)	Z_N	Z_{LN}
60	1.32	0.77	0.78
75	3.24	1.84	1.92
90	5.04	2.75	2.99
110	7.28	3.79	4.34
130	9.38	4.68	5.61
150	11.36	5.46	6.83

Table 7.1: *Comparison of the significances of the estimator using a Gaussian prior (Z_N) and the estimator using a log-normal prior (Z_{LN}) assuming $N_{\mathrm{SM}} = 50$ and a relative uncertainty of 20%. In addition the Poisson probability is stated, which ignores the uncertainty.*

A problem arises when combining the log-normal distributed uncertainties of the different bins into one region. As the additive combination of log-normal distributed variables is no more a log-normal distributed variable, slight inconsistencies between the dicing of the pseudo-experiments (bin-wise) and the p-value calculation (region-wise) might appear.

7.6 Global Interpretation of Search Results

So far the individual event classes have been interpreted apart from the complete set of events. For each event class a significance $\widetilde{P}$ has been computed which easily can be translated into standard deviations, see figure 7.3.

When combining these numerous event classes a final trial factor can be estimated to account for the multiple number of final state topologies looked at. A similar punishment factor could also be used when considering the large number of independent analyses conducted by the whole CMS collaboration.

Conservatively neglecting correlations between the event classes (which is not true for the inclusive ones for sure), the final statistical estimator for the overall degree of agreement with the Standard Model can be quantified using the formula

$$P_{\mathrm{CMS}} = 1 - (1 - \widetilde{P})^n, \tag{7.7}$$

where $\widetilde{P}$ is the significance of a certain event class and n refers to the total number of distributions analysed. Figure 7.8 displays this translation for various number of event classes considered. As an example, if 1000 classes are used, a *local* 5 σ deviation in a certain topology leads to roughly 3.5 σ for *global CMS*.

This global significance P_{CMS} corresponds to a *single* significant deviation found in the context of the many other classes analysed. It gives an answer to the question "if there is a single class with 5 σ, how probable does one get such a single 5 σ or more deviation in any of the event classes when repeating the whole CMS experiment".

As it is expected that deviations show up in several distributions one could also compute a similar global significance using Binomial statistics for other cases, e.g. four classes with a 3 σ deviation or two classes with a 4 σ effect.

Another approach to quantify the global CMS accordance of data and Standard Model expectation is to plot the frequency distribution of the $\widetilde{P}$ values using all event classes analysed. In a dataset where no signal beyond the SM is present these $\widetilde{P}$ values are distributed uniformly as all values are equally probable. If there is a signal leading to significant deviations in several event classes the tails of this global distribution are expected to differ from the SM-only case. More entries than expected with small $\widetilde{P}$ should be observed, thus a discrepancy in the tails of this distribution between a SM-only CMS experiment and a CMS dataset including some signal should be seen.

Figure 7.9 gives an example for such a distribution, using exclusive $\sum p_{\mathrm{T}}$ event classes. Here the $\widetilde{P}$ values ($-log_{10}\widetilde{P}$, thus $3 \mathrel{\widehat{=}} 3.3\sigma$) of all event classes with pseudo-data entries are charted in a histogram[1]. The black curve refers to the expectation of a SM-only dataset.

[1] The distribution is shown as a function of $-\log_{10}(\tilde{P})$ in order to better visualize the deviations in the tails ($> 2\ \sigma$). In this representation the background only curve corresponds to a straight line with a slope of -1.

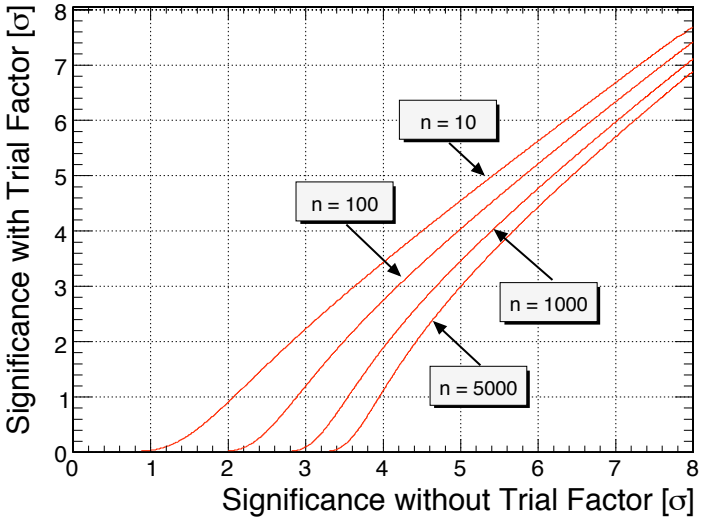

Figure 7.8: *Effect of the global trial factor when investigation n distributions in parallel.*

Here the distributions of several single CMS experiments without any signal are averaged in order to give a reliable prediction. The points on the other hand correspond to a single CMS experiment assuming SUSY LM4 is realized in nature (14 TeV centre of mass energy and 1 $\text{fb}^{-1}$of integrated luminosity). One can clearly see that the SUSY contribution leads to significant deviations in numerous classes. Thus one gets many entries at small $\widetilde{P}$ which are not expected from the SM prediction. Note that classes where only an upper limit can be set ($\widetilde{P} < X$, indicated by the arrow) all contribute to the very last bin.

Integrating the SM-only curve one can determine a similar estimator as the P_{CMS} discussed above. The tail corresponds to the global trial factor, but again only for a deviation in a single event class.

Hypothesis Ranking

The discussion of global trial factors above indicates that it might be desirable to constrain the number of distributions looked at to a minimum. In the context of MUSiC it is clear that $\sum p_{\text{T}}$ of all event classes will be scanned for deviations in a generic way minimizing any bias towards a certain model beyond the SM. Including transverse mass, $\not{E}_{\text{T}}$ or additional distributions looks promising for certain models, e.g. M_{inv} for Z' or $\not{E}_{\text{T}}$ for SUSY.

An interesting approach to lower the penalty of the trial factor is to use a so-called hypothesis ranking [156]. Its feasibility within the MUSiC algorithm is currently under investigation. In general this technique could be a good solution to include additional promising kinematic distributions for certain event classes without blowing up the global trial factor. An example is given in section 8.6.1. The hypothesis ranking always requires to add additional information which rate the variables or classes under investigation. In the simplest way this could be the physicist's experience to classify some classes as more promising than others. This subgroup of classes or distributions would then be analysed first and would benefit from a much lower global trial factor. Therefore the chances for a

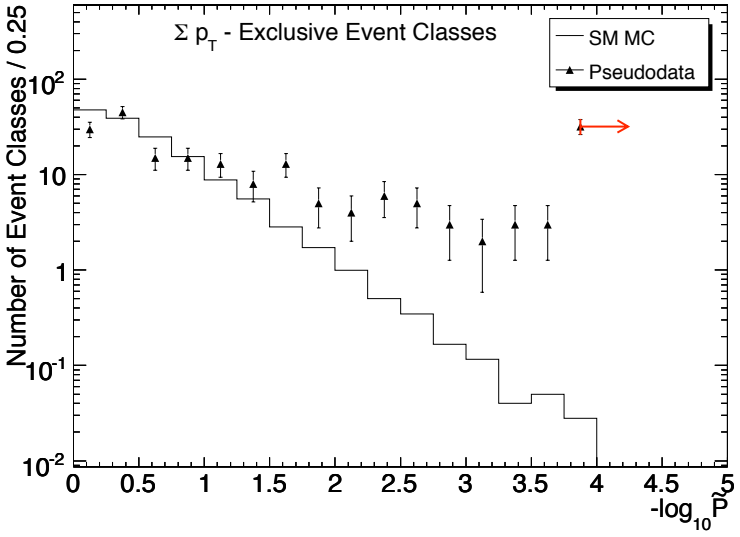

Figure 7.9: *Frequency distribution of the $\widetilde{P}$ values using all exclusive event classes which have pseudo-data entries, using the $\sum p_{\mathrm{T}}$ distribution and assuming* 1 fb^{-1}*. The black curve refers to an averaged CMS experiment with SM-only, the points correspond to a single CMS dataset with SUSY LM4 present (here a centre of mass energy of 14* TeV *is assumed).*

significant deviation are enhanced. In any case – especially when the analysed classes did not show any discrepancy – one can still decide to investigate the other classes with the burden of full trial factor penalty.

7.7 Systematic Uncertainties

As mentioned in the previous section, it is crucial to implement correct systematic uncertainty estimates in the algorithm in order to distinguish a true signal from a "fake" deviation caused by an unanticipated detector effect or an incorrect theoretical estimation of the Standard Model expectation. The following relative systematic uncertainties are assumed and included in MUSiC. Their magnitude is estimated in the context of 1 fb^{-1} of data, but the values can be adapted very easily:

- σ(integrated luminosity) $= 5 - 10\%$
- σ(parton distribution function uncertainty): dynamically, typically $2 - 5\%$
- σ(cross sections) $= 10\%$ (pragmatically)
- σ(jet energy scale) $= 5\%$, change in jets propagated also into $\not\!\!E_{\mathrm{T}}$ estimate
- σ(efficiency correction factor) $= 2\%$ for e, μ, γ and 1% for jets
- σ(fake probability) $= 100\%$ for e, μ, γ
- statistical uncertainty of the Monte Carlo prediction, based on the amount of originally produced events per sample

It is important to stress that the algorithm also accounts for correlations within one error in the context of systematics. For global factors like cross sections all bins in a distribution and the different sub-samples (jet multiplicity bins or p_{T} bins) are correlated. For the integrated luminosity even all physics processes are correlated. In addition to this, variations are not always just "up or down", the JES uncertainty actually redistributes the bins and is again correlated for all generated samples. These correlations have to be taken into account when computing p-values for a certain region and when generating pseudo-data for the whole distribution.

Luminosity

A luminosity uncertainty of 10% should be realistic at the start-up phase up to an integrated luminosity of 1 fb^{-1} of data. At this stage the LHC machine parameters should be well-known and the luminosity monitors should operate smoothly. In addition to this the W- and Z-peak or even $t\bar{t}$ can be used as "standard candles" to determine the luminosity assuming a fixed and precisely known cross section.

Cross Section

The limited theoretical knowledge of cross sections represents a prominent uncertainty to be taken into account within the MUSiC search algorithm. Several uncertainties feed back into the cross section uncertainty. The main limitation is given by the fact that partonic cross sections are only known up to a limited order within the perturbative expansion. Since most of the event generators used at the LHC are only leading-order (LO) implementations so-called k-factors are used within MUSiC to transform the LO cross sections to higher orders, wherever these are known.

In order to obtain an inclusive cross section at a hadron collider these partonic cross sections need to be folded with the parton distribution functions $\mathrm{PDF}(x, f, Q)$ which represent the probability to find a parton of flavour f with a momentum fraction x at a given scale Q (factorization scale) within the hadron:

$$\sigma(pp \to X) = \sum_{i,j} \mathrm{PDF}_{i,p}(x_1, f_1, Q) \otimes \mathrm{PDF}_{j,p}(x_2, f_2, Q') \otimes \sigma_{ij \to X}(Q). \tag{7.8}$$

Therefore, any uncertainty on the PDFs directly propagates into an uncertainty on the cross section. In addition the partonic cross section usually depends on an unphysical cut-off parameter (renormalization scale Q') induced by the limited knowledge of the perturbative expansion and potential divergencies, which require the renormalization of coupling constants at a certain scale Q'. Further uncertainties might be given by the parton shower evolutions within the event generators.

The uncertainty induced by the parton distribution functions are estimated with a reweighting technique which has been pioneered by the MUSiC group within CMS [157]. The method will be sketched briefly here while the details are given in appendix C.

The method relies on the fact that the groups which evaluate the parton distribution functions do not only provide the *best-fit parton distribution functions*, but also hand out $2n$ variations. These "up" and "down" variations of the n variables used within the

PDF fit, transformed into an orthogonal basis, can be used to propagate the experimental uncertainties in the global PDF determination to any variable one is interested in. The exact approach (brute force method) would involve the calculation of the variable X one is interested in $2n+1$ times each time using a different PDF variation (1 best-fit + $2n$ error variations). The uncertainty on the variable X induced by the PDF uncertainty is then given by the master formula

$$\begin{aligned} \Delta X^{+}_{max} &= \sqrt{\sum_{i=1}^{N} \left[max\left(X_i^{+} - X_0, X_i^{-} - X_0, 0\right)\right]^2} \\ \Delta X^{-}_{max} &= \sqrt{\sum_{i=1}^{N} \left[max\left(X_0 - X_i^{+}, X_0 - X_i^{-}, 0\right)\right]^2}. \end{aligned} \tag{7.9}$$

Here, X_i^{+} and X_i^{-} represent the ith up and down variation while X_0 is the value obtained by the best-fit PDF. This approach requires to generate the Monte Carlo $2n+1$ times where n is typically 10 – 20 depending on the considered PDF set. At generator level this might be still feasible for some processes, but when involving the full GEANT based detector simulation to e.g. take acceptances and selection cuts into account this approach becomes impractical or even impossible due to the enormous computing resources needed.

The reweighting technique assumes that for small variations the parton distribution convolution within formula 7.8 can be factorized out. Following this idea one can define for each event a set of $2n+1$ weights, defined by the ratio of the PDF values evaluated for the different error PDFs with respect to the best fit PDF:

$$w^j := \frac{\mathrm{PDF}^j(x_1, f_1, Q) \cdot \mathrm{PDF}^j(x_2, f_2, Q)}{\mathrm{PDF}^0(x_1, f_1, Q) \cdot \mathrm{PDF}^0(x_2, f_2, Q)} \qquad \text{for} \qquad 0 \leq j \leq 2n \tag{7.10}$$

This approach has the advantage that the Monte Carlo has to be produced only once. With the knowledge of the flavours f_1, f_2, the momentum fractions x_1, x_2 and the factorization scale Q provided by the event generators, these weights can be obtained by evaluating the PDFs. Using these weights one can calculates the variable of interest $2n+1$ times. Finally theses values are fed into the master formula (7.9) as in the case of the brute force method.

Both methods are usually in good agreement and predict PDF uncertainties typically in the range of 2% to 8% (see tables C.1 – C.3 in the appendix). The method has been integrated within the MUSiC analysis by storing the weights during the event processing within the grid utilizing the CTEQ 6.1 parton distribution[2] provided by LHAPDF [158]. During the analysis step the distributions of interest ($\sum p_{\mathrm{T}}$, M_{inv}, and $\not{E}_{\mathrm{T}}$) are drawn $2n+1$ times. Applying the master formula bin by bin the uncertainty can be estimated as a function of the variable. This is especially important when looking at distributions ranging over a very broad area as for example in case of a W' search where one needs to know the uncertainty on the W background far off the W peak (see figure 7.10). While the PDF uncertainty at the W peak is only $\sim 5\%$ the uncertainty grows to more than 10% for transverse invariant W masses larger than 2 TeV.

[2] A more recent set of PDFs will only be available within the next major CMSSW release.

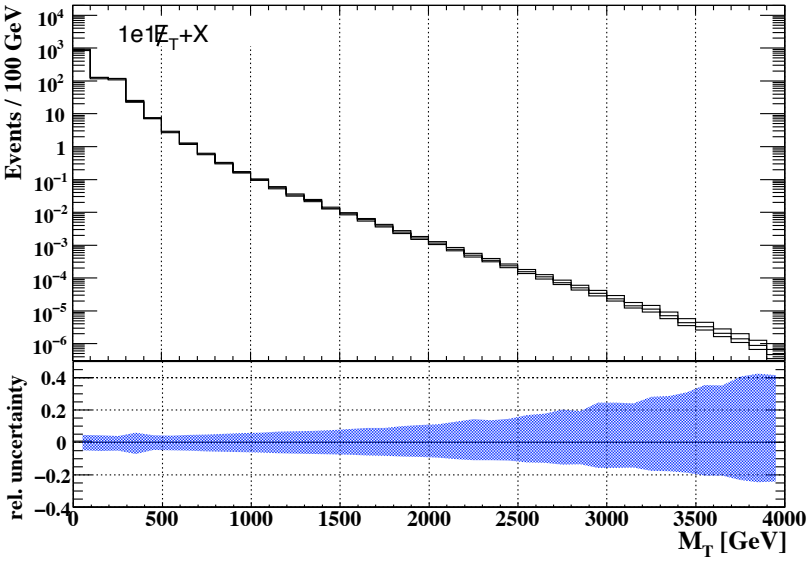

Figure 7.10: *Application of the reweighting technique to a distribution. The top plot shows the transverse invariant mass distribution of a W sample. The relative uncertainty induced by the limited knowledge of the parton distributions (lower plot) increases significantly with larger invariant masses.*

All further cross section uncertainties are assumed to be absorbed by a 10% uncertainty applied to all Standard Model background processes, not distinguishing between different jet multiplicity bins or p_T bins of the generated samples.

The determination of the theoretical uncertainty on the cross section is a quite delicate issue for a model-independent analysis. Many theorists have calculated next-to-leading order or even higher order cross sections and state an uncertainty estimate. However, these calculations concern mostly inclusive processes. Exclusive cross sections, which are the focus of MUSiC with the classification according to final states, might have completely different uncertainties and might vary strongly even for a single process from final state to final state.

The discussion of these theoretical uncertainties within MUSiC is still in flux, and the understanding of these numbers is likely to change in the future. Therefore, it has been decided to use a single "conservative" number (10%). Note, however, that for some processes the uncertainties might even be higher. Nonetheless, the infrastructure is in place so that it is possible within MUSiC to specify the cross section uncertainty for each process individually. Thus the 10% only reflects our current understanding and might well be refined in the future.

Jet energy scale

The 5% uncertainty on the jet energy scale is taken from evaluations done by calorimeter and jet-reconstruction experts [159]. Within 1 fb^{-1} of data both MC truth based calibration techniques and data-driven methods can be used and compared to each other. In this way the simulation can be tuned to match the data, resulting in reliable jet energy scale corrections.

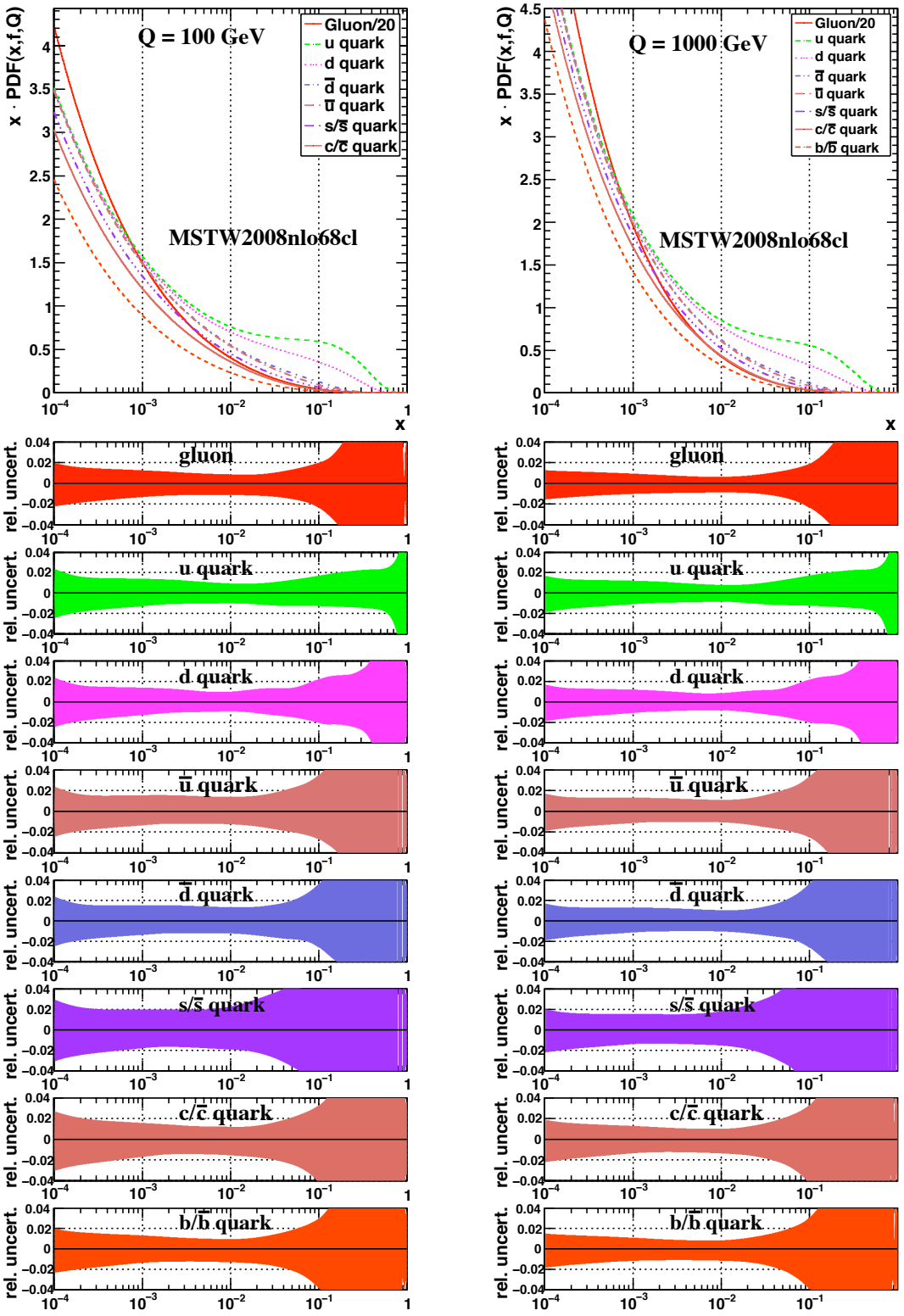

Figure 7.11: *Recent parton distribution functions and their relative uncertainty at the factorization scale $Q = 100\,\mathrm{GeV}$ (left) and $1\,\mathrm{TeV}$ (right) obtained using the distributions provided by the MSTW group[57]. The values from the error varied PDFs have been fed into the master formula to calculate the relative uncertainties.*

As mentioned above this error actually redistributes the bins and cannot be determined directly from the MC mean. All distributions analysed with MUSiC are created also in an up-down variation, scaling all selected jets according to the assumed 5% error. The changes in the 4-vectors of the jets are summed up and the residual variation is vectorially subtracted from the $\not{E}_{\mathrm{T}}$ estimate. In order to compute a p-value according to equation (7.1) the error has to be symmetrized. The direction of the error is taken from the up-variation and thus preserved for all bins, the value is symmetrized by averaging the up- and down-variation for each bin.
In principle also the muon energy scale and the electron energy scale have a certain uncertainty. Still these should be small compared to the JES, but the implementation within MUSiC should be easy following the JES example.

Efficiency correction factor

For electrons, muons and photons efficiency correction factors are included into the MUSiC algorithm. These account for possible efficiency differences between data and Monte Carlo. Using data-driven methods (e.g. tag-and-probe technique on $Z \rightarrow \mu\mu$ events) reconstruction efficiencies can be measured and compared to the Monte Carlo estimate. This will result in correction factors to be applied to the simulated events. These correction factors are the result of careful and complex studies done by the various physics object groups (POG), see also recent studies for muons [136] and for electrons [137]. It should be stressed that MUSiC depends on the input of these numbers from the different groups. That is the reason why standard objects such as global muons or pixel-matched-electrons with standard identification cuts are used. In this way a duplication of work is avoided, synergy effects can be exploited and the scope of MUSiC remains feasible. On the other hand MUSiC can give feedback and spot possible limitations utilizing these numbers in a broader context.
So far the correction factors are implemented as a function of p_{T} and η, with dummy values of unity until first data arrive. We get for the original bin entry N_i

$$N_i^{'} = N_i \cdot f_e^2 \cdot f_{\mathrm{jet}} \qquad \text{for the } 2e \text{ 1jet } (+X) \text{ event class.} \tag{7.11}$$

Of course these Monte Carlo correction factors f_e and f_{jet} are only known up to a certain precision. For muons, electrons and photons we assume a constant relative error of 2% for the correction factor. Since the jet reconstruction efficiency is close to 100% anyway and since QCD events will be available with almost unlimited statistics only a 1% error is assumed for jets. The error can be computed using simple error propagation on equation (7.11), respecting that all bins and all physics processes are correlated. For $\not{E}_{\mathrm{T}}$ no efficiency correction is planned. Here differences between data and MC are likely to be caused by resolution effects and thus an offline-smearing of the MC objects could be performed.
One should note that in the context of reconstruction efficiencies at first approximation MUSiC assumes them to be independent of the number of particles in the event. Thus the efficiency for one muon is the same as for two other muons which are in the event. Of course for very complex particle topologies this might not be true, still it is hard to solve special issues and problems like this in a generic way for all event classes. This relates to

the introductory remarks of this note: A deviation found by MUSiC has to be investigated in the following with more dedicated checks. MUSiC acts more like a warning system which cannot account for all details. Questions like the efficiencies within complex final states have to be addressed when investigating the deviation(s) found by MUSiC.

Fake probabilities

The estimation of fake probabilities for the reconstructed objects using first data will probably be a challenging task. Also huge differences with respect to the misreconstruction probabilities predicted by the detector simulation should not be surprising. In principle one could perform similar MC corrections as in the case of reconstruction efficiencies. Still it is not clear which level of detector understanding and MC tuning is needed before this seems realistic. First studies using data-driven techniques can be found in [138]. In the scope of this note and as a first attempt to implement this uncertainty we have decided to rely on the MC-truth-knowledge for the moment. A reconstructed object not matching within a $\Delta R < 0.2$ criterion to a generated isolated particle is labelled as "fake". For jets and $\not\!\!E_{\mathrm{T}}$ the dominant uncertainty is already covered with the jet energy scale, thus fake errors are only assumed for muons, electrons and photons. As a conservative "guess" of the error on fake probabilities 100% uncertainty is assumed. Thus for each event processed the number of fake objects is counted and an event weight for the "up"-variation is calculated:

$$weight^{\mathrm{up}}_{\mathrm{fake}} = 1 + \sqrt{(N_{\mathrm{fake}}(e)\cdot\sigma_{\mathrm{fake}}(e))^2 + (N_{\mathrm{fake}}(\mu)\cdot\sigma_{\mathrm{fake}}(\mu))^2 + (N_{\mathrm{fake}}(\gamma)\cdot\sigma_{\mathrm{fake}}(\gamma))^2} \quad , \tag{7.12}$$

where $N_{\mathrm{fake}}(e)$ denotes the number of fake electrons in this specific event and $\sigma_{\mathrm{fake}}(e)$ the relative error of the fake probability. This results in an additional distribution where the fake probabilities are varied by one sigma. Again the differences between this distribution and the mean MC values can be computed and used for the algorithm. Since misreconstruction is an overall detector effect all bins and all physics samples are correlated.

Smearing corrections

Once first data have arrived the tuning of the detector simulation will start. In this context resolution differences between data and MC are likely to appear which would demand to further smear reconstructed objects in the simulation. Also the widths of these smearing functions are known only up to a certain precision. Thus one might consider varying this width by one sigma and further smear the MC. This would give an error estimate on the effect of these smearing steps performed in the simulation.
The implementation of these errors is very similar to jet energy scale variations and parts of the infrastructure could be re-used. Since without data these smearing corrections are not needed this systematic uncertainty is not included for the time being. Still it should be straightforward to include them in MUSiC in future iterations.

Non-collision backgrounds

Since MUSiC is analysing the event contents of pp-collisions the contributions of other sources of particles can be regarded as another systematic uncertainty. While the effect of pile-up is expected to be small at initial luminosities contributions from beam halo and cosmic muons are always existent. Both sources could cause deviations between data and SM simulation, especially in "exotic" channels with very high particle multiplicities. Still both sources are also expected to be relatively rare: The cosmic muons as well as the beam halo particles are asynchronous with respect to the hard interaction. Also their passage through CMS is quite different from the particles originating from the vertex. Additionally, CMS is 90 m underground, such that the rate of cosmic muons arriving at CMS is only O(100 Hz). Thus, these events are unlikely to fire a trigger. When overlaid to a pp-triggered event the differences in timing and direction can hardly lead to high-quality reconstructed objects with central η and high-p_{T}.

Still they are an irreducible background which can affect data-MC comparisons. Luckily for both sources of particles dedicated Monte Carlo generators exist [116, 160]. For future MC productions events from both beam halo and cosmics are planned to be mixed under the hard collision, just like for pile-up events. Ultimately of course real cosmic/beam halo/pile-up background events could also be overlaid. In addition to this, loose cuts on the extrapolations of the tracks to the vertex (Δz) can help to further reduce these backgrounds.

Chapter 8

Probing MUSiC with Benchmark Channels

A difficulty of a model-independent search is the quantification of the search results, especially without data. While a feasibility study can state expected discovery or exclusion limits within the parameter space of the theoretical model under investigation, this is not possible for the model-independent search due to a lack of signal. Nonetheless it is possible to show its performance with benchmark scenarios generating pseudo-data which include in addition to the Standard Model further signatures of new physics. Other deviations which might be profiled are pseudo-data which include unexpected detector or Monte Carlo effects. Still one should keep in mind that these representative use cases or toy examples reflect only a small part of the enormous phase space that can be covered by such a generic approach.

Serving as an alarm system for deviations, the threshold for an interesting deviation within the MUSiC analysis is defined to be three standard deviations corresponding to a $\widetilde{P}$ of at most 10^{-3}. This threshold is still far away from the region where conventionally a discovery is stated (5σ), but already in a regime where a statistical signal-like fluctuation is relatively rare. Such a deviation would be worth to study in greater detail, possibly with a new dedicated analysis.

The investigated benchmark channels which demonstrate the feasibility of the MUSiC analysis in the order of the expected time-line from the detector start-up to a mature stable-running and well-understood experiment, are:

- Physics object and software commissioning
- Detector commissioning: spotting a detector effect
- Event generator and simulation tuning
- First Day Physics: detection of a prominent deviation
- Signatures of new physics with deviations in many distributions

In each section the possible benefits of the approach are explained and whenever possible compared to a dedicated analysis. Finally, some possible extensions of the MUSiC analysis are discussed.

8.1 MUSiC as Physics Debugging Tool

Already without the presence of collision data, MUSiC has proven its benefit for the CMS collaboration as a tool to spot coding bugs and configuration mistakes mostly affecting physics silently. Due to the flexible nature of the CMS reconstruction framework the very same code which is currently applied to simulated data will be used upon the arrival of the first proton-proton data. Therefore any coding mistake which is fixed now will help to speed up the detector and physics commissioning. This debugging mainly consists of understanding the basic quantities of all studied physics objects from simple momentum distributions, over isolation quantities to variables used within the object identification. For this purpose the control plot factory part of MUSiC (see section 5.4) currently contains more than 500 distributions which are created in parallel to the classification of the events into final states. This complementary class allows to gain a decent understanding of the physics objects which is an absolute prerequisite for the understanding of the search results within the different event classes.

Out of the numerous bug discoveries MUSiC was involved, only a few should be mentioned here representing the potential of the model-indepentent search in a few concrete examples.

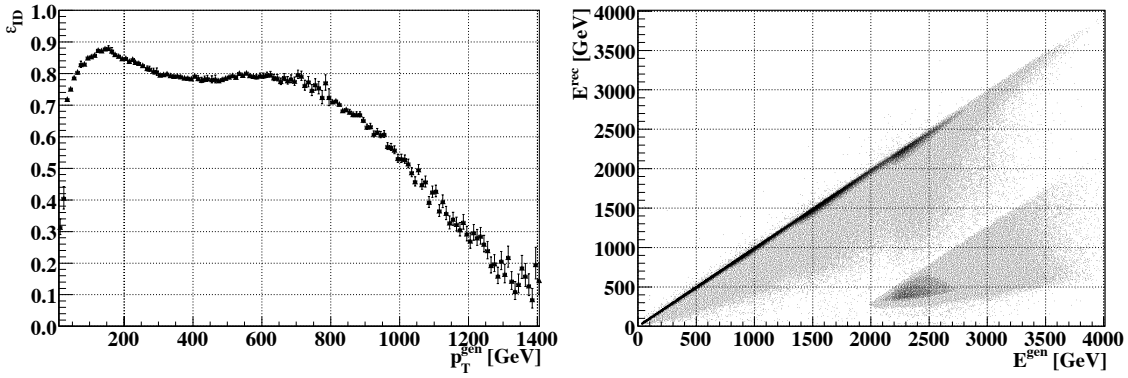

Figure 8.1: *Flaws in the photon identification and reconstruction. The* ***left*** *distribution shows the photon efficiency which drops at high photon momenta due to a cut on the calorimeter isolation not suitable for such photons.* ***Right:*** *Reconstructed versus generated photon energy revealing an incorrect handling of ECAL cells with saturated readout electronics (≈ 1.7 TeV in the barrel).*

- **The MadGraph High Level Trigger Bug**
  Due to a faulty implementation of the MadGraph generator interface within CMSSW, all its generated events claimed to be real data. As a consequence the high level trigger simulation was skipped since real data are already expected to have gone through the whole trigger chain. This resulted in more than 30 million events which are of limited use for physics analyses and a tremendous waste of computing power

before MUSiC detected the defect. The bug has been fixed quickly and the events have been produced again.

- **High Energy Photons**
  Since the photon identification has only recently been established, there is still room for improvements, especially for high energy photons. Figure 8.1 (left) shows the identification efficiency as a function of the reconstructed transverse momentum. For momenta above $\approx$ 700 GeV the efficiency drops quickly to 0. This is related to the fact that the official tight identification (see section 6.3) uses an absolute isolation cut on the ECAL (5 GeV) and HCAL (10 GeV) deposit excluding the photon by a cone with a fixed size in ΔR. For photons with momenta above 700 GeV the photons start to leak significantly into the isolation area and are thus removed.

  Another issue can be seen in figure 8.1 (right). It shows the energy of the reconstructed photons versus the energy of the generated photons within the barrel matched by an ΔR criterion. The energy of most photons is reconstructed properly leading to a clear one to one correlation with some tails towards smaller reconstructed energies. However, photons with energies above $\approx$ 2 TeV are often reconstructed with a deficit of about 1.7 TeV. This is related to the fact, that the ECAL readout electronics saturates at energy deposits of about 1.7 TeV in the barrel, and 3.0 TeV in the endcaps. Although this is correctly simulated at GEANT level, a bug within the unpacking/reconstruction of the information leads to zero energy deposits within crystals with saturated read-out electronics. Therefore, very high energy electromagnetic showers which deposit such energies in a single crystal are reconstructed with an energy deficit of $\approx$ 1.7 TeV.

  Both issues have been fixed within the latest CMS software releases.

The examples demonstrate that especially at the start-up of an experiment MUSiC might help to improve the understanding of the detector and the reconstructed physics objects. It might serve for physics validation purposes in a way complementary to the data quality monitoring due to its focus towards a physics analysis.

8.2 MUSiC and First Data

Especially during early data taking ($\ll 1\ \mathrm{fb}^{-1}$) the physics focus will not be to discover some signal beyond the Standard Model, but rather to re-establish the SM with the CMS detector. In order to measure the various SM candles properly a lot of work will be needed to understand the CMS detector. After years of construction and simulation studies for the first time data will be recorded and can then be compared to the "ideal" Monte Carlo world. Differences between data and simulated events can arise from Monte Carlo generators not properly describing nature at 10 TeV center of mass energy, or from a detector not working exactly as predicted by the detector simulation. Both aspects can be addressed using a generic search approach since measurements differing from the expectation can be revealed by the algorithm. Since a large part of the data are divided into event classes, MUSiC can perform a general scan of the different detector properties, possibly revealing unexpected

discrepancies. Of course many detailed studies will examine efficiencies, resolutions and other detector properties to a great extent. Still MUSiC can serve as a cross check. Thus it is interesting to see how well the data agree in general in the various event classes without extensive tuning and optimizations. On the other hand MUSiC can assist in the process of Monte Carlo tuning, monitoring the improvements of the SM Monte Carlos and giving feedback where additional changes might be needed.

8.2.1 Noise in the Calorimeter

While in previous MUSiC studies the effect of a non-accurate jet energy scale calibration or an efficiency difference between data and Monte Carlo have been discussed [161], another common issue is outlined here: noise within the calorimeters. As a variety of physics objects such as photons and jets heavily rely on calorimetric measurements the understanding and commissioning of these detector parts is extremely important. Experiences at past colliders teach us that great care has to be taken to identify and remove objects not related to a "physics signal", but to an unforeseen detector effect. Dedicated algorithms to identify noisy cells have already been studied and implemented on the basis of the operational experience gained during the frequent cosmic data takings in the last years [162]. While "hot" cells i.e. detector parts with very frequent "fake" signals can be identified relatively easily, cells which only rarely emit a signal related to an unforeseen detector effect are much more complicated to spot [163]. Several layers of quality control are therefore put in place to monitor the data recorded by the subdetectors. Unphysical energy deposits within the calorimeters induced by occasional electronics malfunctions, noise and hot cells are supposed to be flagged by the online and offline shift crew. Still, such malfunctions can appear at any given time and might need a significant amount of data and time to be identified and treated appropriately. Therefore any analysis which cross checks the quality of the data might improve the understanding and trust in the CMS data. As MUSiC investigates many distributions it might serve as an additional check complementary to the official data quality monitoring with the focus of a physics analysis.

Here a scenario is presented which shows how the MUSiC algorithm might be sensitive to such a detector malfunction. In order to do so random "noise" within three fixed detector regions (at $\eta = -0.1, 0.8, -1.6$ and $\phi = 0.2$) is generated within on average one per mille of the selected events. An energy deposit of 600 GeV roughly corresponding to half of the electronics saturation energy of a cell [162] is considered with a Gaussian spread of 30 GeV. As the treatment within the full detector simulation would be too complicated and beyond the scope of this example, these deposits are added as an additional four-vector jet possibly merged with a nearby jet at the final stage of the analysis. Figure 8.2 represents two of the many classes which would see a significant deviation. The left plot corresponding to the inclusive 1γ+jet class, reflects the effect on the $\sum p_{\mathrm{T}}$ distribution. In total three deviations can be spotted at $\approx$ 300 GeV, $\approx$ 500 GeV, and $\approx$ 650 GeV. While the former is hardly visible, the latter is considered as most prominent due to the smaller Standard Model contribution in this region. The threefold structure is related to the fact that the jets are added with fixed energy at three different pseudorapidities and thus different transverse

momenta. The simulated noise contribution is further "smeared" by the momentum of the photon.

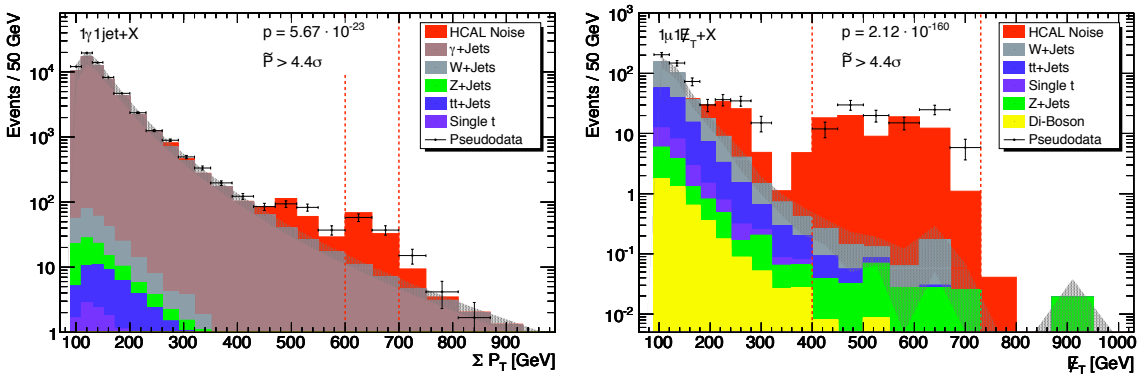

Figure 8.2: *Effect of noisy cells on a jet* ***(left)*** *and missing transverse energy* ***(right)*** *distribution. The regions of interest contain* $N_{\text{data}} = 95$, $N_{\text{MC}} = 17.9 \pm 3.3$ *and* $N_{\text{data}} = 117$, $N_{\text{MC}} = 0.8 \pm 0.2$ *events, respectively.*

Additional energy deposits caused by detector noise or a malfunction also have a severe influence on the transverse missing energy distribution. To demonstrate this, the missing transverse energy within this scenario is also changed when adding an additional "noise jet". The effect can be clearly seen in figure 8.2 (right) where the missing transverse energy distribution for the $1e\not{E}_{\text{T}} + X$ class is displayed. Here the first peak is also visible while the two others are almost merged representing the region with the largest deviation of the pseudo-data and the Standard Model expectation. The significance of the deviation is very high and underlines the well-known fact that the understanding of the missing transverse energy is challenging at a detector start-up. All object mis-reconstructions, malfunctions and detector effects not under control are propagated into the measurement of the missing energy.

8.2.2 Monte Carlo Tuning

Apart from the detector commissioning and understanding of physics objects the first LHC data will be extensively used to tune the Monte Carlo simulations to the observations at 10 TeV or later 14 TeV centre of mass energy. Extrapolations from the measurements of previous collider experiments like the parton structure functions from HERA or the underlying event tunes from the Tevatron will be probed, validated and improved by various dedicated analyses. Other differences might originate from theoretical uncertainties like the missing inclusion of higher order contributions (k-factors).

Such a scenario is constructed here within a toy example where two different Monte Carlo predictions are used, basically comparing an advanced tree level prescription (PYTHIA) with a matrix element event generator (MADGRAPH). The example assumes that the data follow the MADGRAPH prediction while in the Standard Model MC the Drell-Yan sample is replaced by an equivalent PYTHIA sample. Figure 8.3 shows two representative distributions. While the inclusive two electron class agrees very well, the distributions differ with an increasing number of jets. The differences are expected as the parton shower is not

able to model the second or further jets as accurate as the matrix element implementation. The excess of the pseudo-data reflect that MadGraph more often produces events with a second, harder jet. The regions of interest are found accordingly: while in the distribution without jets a random non-significant deviation within the tails is picked, the inclusive two jet class shows a broad region of interest with a 3σ deviation.

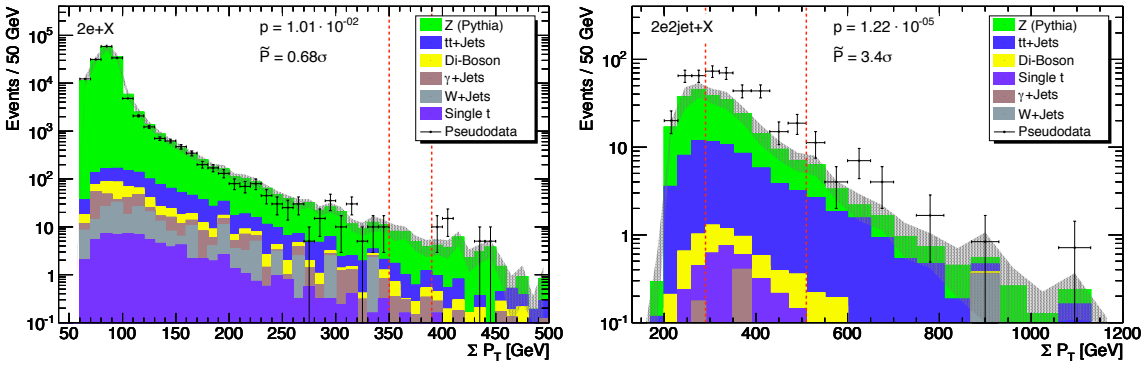

Figure 8.3: *MC tuning example. The pseudo-data are assumed to follow the* MadGraph *description, while the SM MC reference used a Z sample generated with* pythia. *While the inclusive two electron class* ***(left)*** *shows a good agreement, the distributions differ with increasing jet multiplicity* ***(right)****. The regions of interest contain* $N_{\text{data}} = 0$, $N_{\text{MC}} = 4.9 \pm 0.8$ *and* $N_{\text{data}} = 183$, $N_{\text{MC}} = 87.7 \pm 18.6$ *events, respectively.*

8.3 Interlude: Multi-Jet Background Estimation from Data

While the modern Monte Carlo generator tools produce fairly reliable predictions of shapes for the various distributions of Standard Model processes like W+jets or $t\bar{t}$ +jets with high statistics, it is clear that for QCD multi-jet production the enormous cross sections exceed the computational resources available. In addition the theoretical uncertainties for multi-jet events are orders of magnitude larger than in the case of electro-weak processes. This analysis investigates events with at least a single isolated lepton or photon. Within multi-jet events these objects are only produced via non-prompt mechanisms or via misidentification, e.g. muons from b-jets or electrons/photons from misidentified jets with a large pion fraction. Compared to the inclusive di-jet cross section these "fake" leptons are very rare, and thus difficult to model using inclusive multi-jet Monte Carlo samples.

The MUSiC approach aims to estimate the multi-jet contribution from the data in order not to rely on the simulated prediction only. Since a generic search is looking at many different distributions and a diversity of final states, it is not practical to define control regions for each specific event class. One has to use a more general estimate of the multi-jet background applicable to all classes. The uncertainties of such cross-class extrapolations have to be absorbed by an appropriate global uncertainty of the multi-jet estimate, which can be easily incorporated into the search algorithm.

The strategy used to estimate the multi-jet contribution from data is similar to the methods commonly applied at the Tevatron [164], also known as "ABCD"-method within CMS. The basic idea is to cut the phase space into four regions utilizing two variables each

being prominent for separating the contribution to be determined from data from the rest. Using the shape from one of the regions rescaled by the ratio of the event numbers in two other regions the contribution within the fourth region can be determined (see figure 8.4 for an illustration).

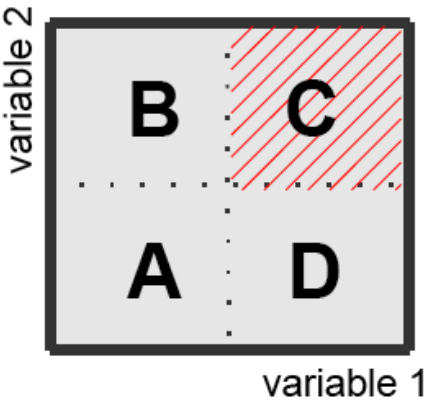

Figure 8.4: *Illustration of the background estimation via the "ABCD"-method. Two uncorrelated variables have to be found, which are prominent for separating the contribution to be determined from data (e.g. a certain background) from the rest (e.g. a signal enriched region). The certain background contribution in region C can then be calculated by taking the distribution within region D (or B) weighted by the ratio of the events in region B and A (or D and B). A prerequisite for this technique is that regions A, B, and D are "signal-free".*

For MUSiC the approach is a bit more challenging as in general no variable can be found which separates multi-jet events from the rest of the Standard Model within each class i.e. one has to correct for a "signal" contamination.

Here a single selection cut, which is prominent for distinguishing "fake" leptons from well-measured isolated ones, is inverted or relaxed. The sample with the inverted cut is then used to model the shape of the QCD background, and a control region (p_{T} cut) is defined where the sample is scaled to fill up the gap between the remaining SM Monte Carlo samples and the data. Great care has to be taken that the shape of the relaxed distribution is still equal to the (independent) shape to be estimated from data.

This method is exercised here using final states with electrons. Previous studies show that it works similarly in the muon case [161]. Two variables are suitable for distinguishing "fake" electrons from well-measured isolated leptons: the (track) isolation, which also has been used in the multi-jet estimation in the muon classes, and the identification variable. Both variables work equally well. Here the method is only exercised using the identification variable. The second variable required by the "ABCD"-method is chosen as the $\sum p_{\mathrm{T}}$.

In order to aggregate a distribution where the multi-jet contribution is enhanced compared to the other Standard Model processes, the electron identification requirement is turned from *tight* into *loose* and at the same time *not tight*. As the Standard Model processes like W- and Z-production contain clean, isolated, and thus *tight* electrons, these contributions are suppressed. At the same time the shape of the distribution is kept unchanged. Figure 8.5 shows the multi-jet enhancement induced by the relaxed ID criterion for two event classes.

By only relaxing but not inverting the electron identification cut a sample is obtained with similar kinematics compared to the multi-jet events entering the final selection. By inverting the cut one would risk to introduce larger differences in the distributions.

Two control regions are defined which are used to determine the scale factor,

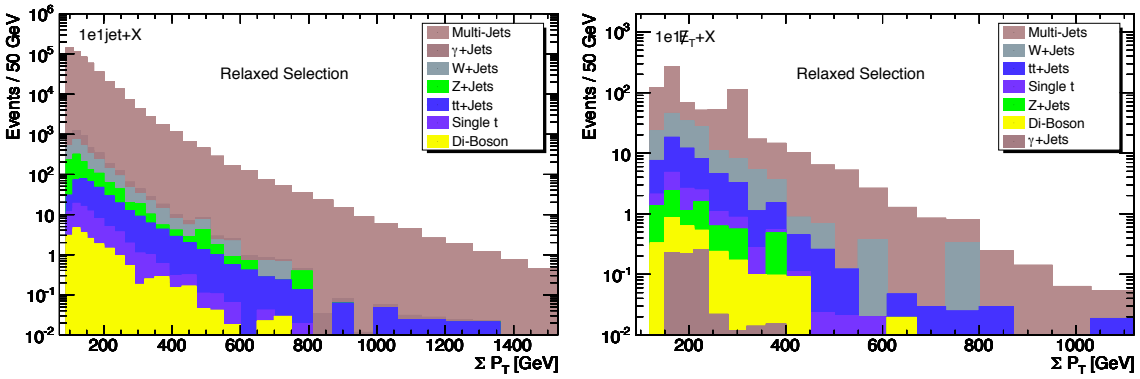

Figure 8.5: *Multi-jet Monte Carlo and other SM processes with relaxed cuts in comparison, for the two event classes used for normalization. The distributions are normalized to an integrated luminosity of 100* pb^{-1}.

- $90 - 170$ GeV in the $\sum p_{\mathrm{T}}$ distribution of the class $1e$ 1jet+X
- $130 - 210$ GeV in the $\sum p_{\mathrm{T}}$ distribution of the class $1e$ $\not\!\!E_{\mathrm{T}}$+X.

These two inclusive classes represent quite different corners of the phase space analysed with MUSiC, once requesting a lepton and a jet and once the combination of a lepton and missing transverse energy. In this way two independent estimates of the scale factor are obtained. Furthermore the two regions are both located at the very low p_{T} edge of the distributions, where a possible signal contamination from new physics is expected to be small and multi-jet plus other Standard Model processes dominate. From these control regions ($f_{\mathrm{QCD}} = 1.6$ for class $1e$ 1jet+X and $f_{\mathrm{QCD}} = 0.5$ for class $1e$ $\not\!\!E_{\mathrm{T}}$+X) one obtains the following scale factor with its uncertainty:

$$f_{\mathrm{QCD}} = \frac{\text{“data” } - \text{ SM MC without multi-jets}}{\text{relaxed “data” } - \text{ relaxed SM MC without multi-jets}} = 1.05 \pm 0.55 \qquad (8.1)$$

The relative uncertainty of 50% indicates that the estimation of multi-jet background from data for all event classes is not very precise. Nevertheless, since the multi-jet contribution in the regions of interest for possible deviations (e.g. from new physics) is not very large in most cases, even such a large uncertainty should have a minor impact on the search sensitivity. It is more vital to get a proper shape of the multi-jet background in all classes without the enormous single bin fluctuations of a Monte Carlo sample caused by the lack of MC statistics. Note that since this method is exercised only using a multi-jet Monte Carlo sample (as “data”) the subtraction of the other SM samples is not required. In any case the contribution in the denominator from relaxed SM MC without the multi-jet part is small since these mostly fulfill the tight electron ID.

Figure 8.5 illustrates this: Here the two event classes used for the normalization are shown, comparing the amount of multi-jet events which pass the cut relaxation to the rest of the SM processes. One can see that there is at least an order of magnitude between the multi-jet events with relaxed cuts and the other relaxed SM samples. Thus a possible

uncertainty on the subtraction of the relaxed SM samples without the multi-jet contribution, see denomiator of equation 8.3, is well absorbed by the overall 50% uncertainty of the multi-jet estimate.

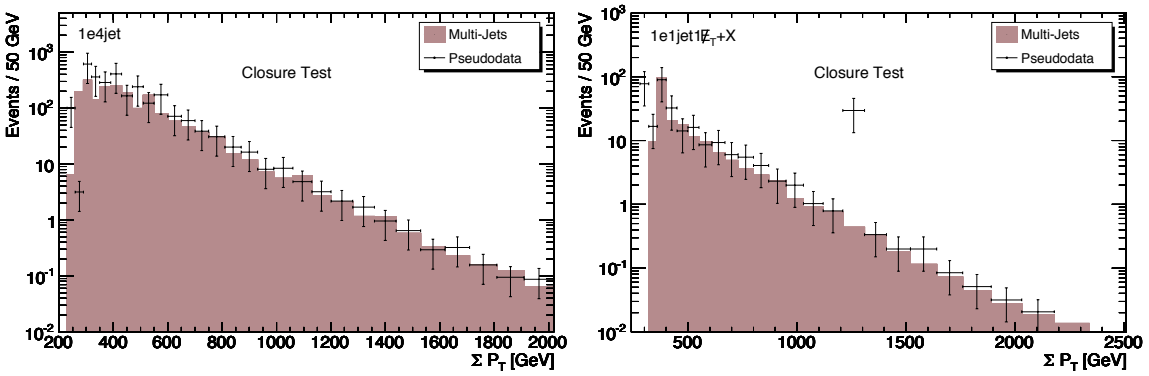

Figure 8.6: *Multi-jet Monte Carlo and estimate using cut relaxation in comparison, for two representative event classes. As the estimate from data will not be limited by statistics the displayed uncertainty is given by the uncertainty on the scale factor f_{QCD}. The single bin within the right distribution showing a discrepancy reflects the limited MC statistics for some of the multi-jet samples. The distributions are normalized to an integrated luminosity of 100* pb^{-1}.

Now all the ingredients are at hand to extent the method to all other event classes: the shape is taken from the relaxed distribution measured from data within each event class while the normalization factor determined from the two reference classes is assumed to be globally valid. Figure 8.6 shows the comparison of the multi-jet estimate from "data" with respect to the multi-jet Monte Carlo samples used to perform the cut relaxation. The uncertainties correspond to the uncertainty of the scaling factor. The sample with relaxed cuts and the one fulfilling all final selection cuts agree well in terms of the shape. Note that the event classes shown here do not contain the control regions, thus the agreement within the assumed uncertainties serves as a good indication that the extrapolation from one final state topology to another works reasonably well.

8.4 Early Searches for New Physics

Already with the first year of data and a still limited understanding of the detector, several prominent signatures for physics beyond the Standard Model might show up. These would be deviations which exceed the expectation by far beyond the expected large systematic uncertainties of a moderately calibrated and aligned detector. Such a signal might for example originate from theories containing new heavy gauge bosons or leptoquarks. Of course these theoretical models are already covered by dedicated searches. Here, the strength of MUSiC is in the variety of investigated final states. While dedicated searches scan only the invariant di-lepton or missing energy+lepton spectra, MUSiC will search for a deviation in any possible invariant mass spectrum also including final states with more than two physics objects. Therefore one should consider the following example of a heavy charged gauge boson as a benchmark with an application in a much broader context. The

availability of a dedicated analysis within CMS also allows to quantitatively compare the MUSiC results with the traditional approach.

The unexpected "discovery" of a new heavy boson was also the dress rehearsal for MUSiC in 2008. During the study of a SM cocktail (14 TeV centre of mass energy) provided by the CMS collaboration it turned out that the management had steered the inclusion of a Z' as a "hidden signal". The MUSiC analysis successfully detected this prominent deviation as one of the first (see figure 8.7 (left), details given in [121]). Here, the focus is on 10 TeV centre of mass energy and the related possible new heavy charged gauge boson W'.

8.4.1 New Heavy Charged Gauge Bosons

There are a several studies which have investigated the discovery potential of new heavy gauge bosons at a centre of mass energy of 14 TeV in the past [31–33]. The recent result given in figure 8.7 (right) estimates the discovery reach for a W' decaying into an electron and a neutrino at a centre of mass energy of 10 TeV as a function of the mass. In order to do so the distribution of the transverse invariant mass of the electron and the missing energy at 14 TeV from [33] has been reweighted to 10 TeV and fed into the same significance estimator (CL_S-method).

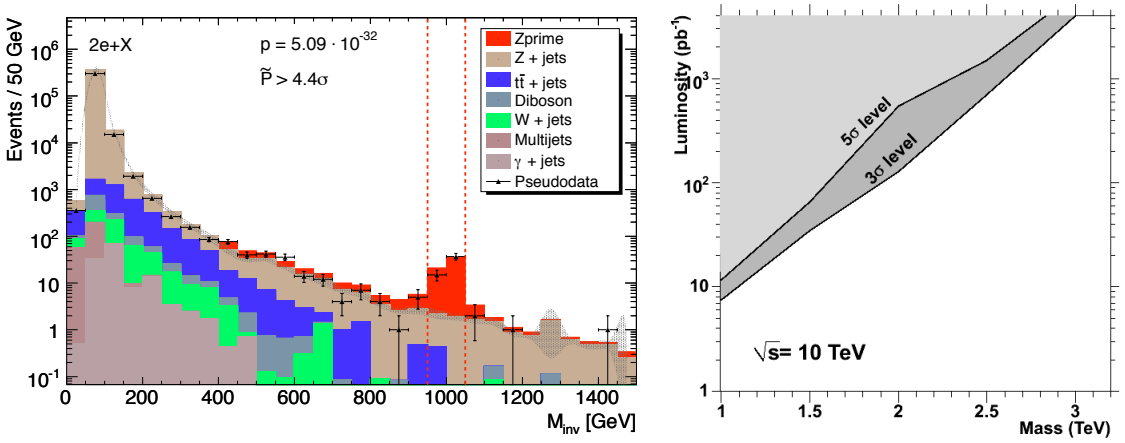

Figure 8.7: ***Left:*** *Z' "hidden signal" as dress rehearsal for the MUSiC analysis in 2008 (14 TeV centre of mass energy).* ***Right:*** *Required luminosity for the discovery of a potential W' in the electron plus neutrino channel as a function the W' mass at a centre of mass energy of 10 TeV.*

Inspired by these results, $W' \to e\nu$ samples with W' masses of 1 TeV, 1.5 TeV, and 2 TeV have been investigated within MUSiC at integrated luminosities of 10 pb^{-1}, 65 pb^{-1}, and 325 pb^{-1}. The luminosities are chosen according to the expected 5σ discovery reach of the dedicated analysis. The corresponding leading order cross sections times branching ratio are 1227 fb, 213 fb, and 50 fb, respectively. In order to be comparable with the dedicated analysis the global scan within MUSiC is restricted to the exclusive event classes (jet veto).

Scanning all exclusive event classes the biggest discrepancy between pseudo-data and SM expectation is found in the M_T distribution of the class $1e\,\not{E}_T$, which is also the final variable of the dedicated W' search. Figure 8.8 (left) shows this distribution for a single

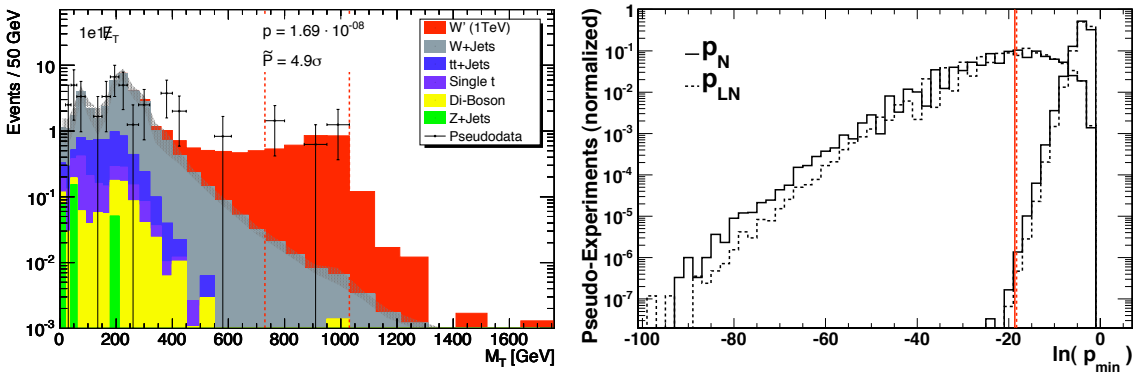

Figure 8.8: ***Left:*** *1 TeV W' signal within the invariant mass distribution of the 1e + $\not{E}_T$ class close to the discovery reach with an integrated luminosity of 10* pb^{-1} *($N_{\mathrm{data}} = 5$, $N_{\mathrm{MC}} = 0.07 \pm 0.01$). The double-peak structure is due to the general $\not{E}_T$ cut of 100* GeV. ***Right:*** *Corresponding p-value distributions for signal+background and background only comparing the two significance estimators Z_N and Z_{LN}.*

CMS experiment of a 1 TeV W' at an integrated luminosity of 10 pb^{-1}. The region of interest nicely selects the W'-peak at 1 TeV and the $\widetilde{P}$ of 4.9σ[1] indicates that the signal is close to a discovery.

Table 8.1 displays the detailed comparison results for the three W' masses. For the MUSiC analysis the significances have been determined with two different estimators. One is based on the Gaussian treatment of uncertainties (Z_N) while the other utilizes a lognormal approach (Z_N) (for details see 7.5.2). The results agree very well with the traditional analysis and also state that a discovery at these luminosities is possible. A priori this is quite surprising as no optimization with respect to any signal has been performed within MUSiC. In addition the application of the trial factor reduces the significance further. However, this kind of signal is quite specific as the expected Standard Model background is many orders of magnitude smaller than the signal. As only very few bins are actually populated with Standard Model events the effect of the trial factor is much less severe than in other distributions. The expected $p^{\mathrm{min}}_{\mathrm{data}}$ value of an average CMS experiment measuring data which contain a W' are of the order of 10^{-9}. This is consistent with the expected trial factor which usually lowers the p-value by a factor 100 – 1000 resulting in a $\widetilde{P}$ of 5σ in the case of the W'.

A remarkable feature of the algorithm is the region of interest which is picked i.e. the region where the discrepancy between pseudo-data and SM Monte Carlo is largest. In most cases it is just the bin containing the Jacobian peak. This is given by the fact that the signal as a function of the mass is roughly flat, but the background is exponentially decreasing. This distinguishes the model-independent search from the traditional approach which defines a broader region of interest a priori, while MUSiC defines the region when looking at the data. The price for this liberty is the trial factor.

This example also allows the comparison of the different significance estimators implemented within the search algorithm. Although their mathematical foundation and algo-

[1] To be consistent with the dedicated W' search all significances within this paragraph have been calculated as one-sided Gaussian deviations.

W' mass	$\mathcal{L}_{\text{int}}$	W' Analysis	MUSiC Z_N	MUSiC Z_{LN}	$p^{\text{min}}_{\text{data}}$ (expected) Z_N	$p^{\text{min}}_{\text{data}}$ (expected) Z_{LN}
1 TeV	10 pb^{-1}	$\approx 5\sigma$	$(5.04 \pm 0.08)\sigma$	$(5.12 \pm 0.06)\sigma$	$7.8 \cdot 10^{-9}$	$1.1 \cdot 10^{-8}$
1.5 TeV	65 pb^{-1}	$\approx 5\sigma$	$(5.09 \pm 0.08)\sigma$	$(5.5 \pm 0.3)\sigma$	$3.6 \cdot 10^{-9}$	$4.9 \cdot 10^{-9}$
2 TeV	325 pb^{-1}	$\approx 5\sigma$	$(5.11 \pm 0.08)\sigma$	$(5.3 \pm 0.1)\sigma$	$2.9 \cdot 10^{-9}$	$5.0 \cdot 10^{-9}$

Table 8.1: *Quantitative comparison of a dedicated search to the MUSiC approach for three different W' masses (stat. uncertainties only). The luminosity is chosen according to the 5σ-reach of the dedicated search. The following columns show the significances and expected mean signal+background p-values of the MUSiC search algorithm using two different estimators.*

rithmic implementation are fundamentally different their predicted significances agree well. Figure 8.8 (right) shows the comparison of the p-value distributions for signal+background and background only utilizing more than 10^7 pseudo-experiments. In agreement with the discussions in chapter 7.5 the p-values of the Z_{LN}-estimator are larger i.e. mark less significant deviations. As this effect (in this case) is more prominent in the background only distribution than in the signal+background (the mean value of the signal+background of the two estimators differ only slightly), the resulting $\widetilde{P}$ values of the Z_{LN}-estimator show a more liberal behaviour. Therefore the standard deviations of Z_{LN} given in the table have the trend to be larger than the corresponding Gaussian based estimator Z_{N}.

8.4.2 Negative Example: Higgs

Obviously, there are also possible signals where a model-independent approach is less successful. The classical LHC example is the search for a Standard Model Higgs boson which is very advanced and highly tuned for different Higgs masses and thus favourable decay channels. In this case all parameters except the mass of the expected signal are known. This provides the basis for a very solid tuning towards the different signatures and the potential of an efficient background suppression.

Consequently, a global search for a Standard Model Higgs with a mass of 160 GeV and a leading order cross section of 710 fb (decay into WW with di-leptonic final state) within MUSiC does not lead to any significant deviation (assumed integrated luminosity: 1 fb^{-1}). Figure 8.9 (left) shows a representative distribution, using the $1e\ 1\mu + X$ event class. One can see that in the kinematic region of the distribution where the Higgs contributes the SM is orders of magnitudes above the signal. The region of interest picked by the algorithm is not close to any signal, but only a fluctuation of the background. This is reflected by the $\widetilde{P}$ of 0.13σ, which means that the deviation found in the pseudo-data agrees with the SM expectation well within the assumed uncertainties. Looking at the $\widetilde{P}$ values of all event classes where the Higgs signal contributes (figure 8.9 (right)), one can see that the fluctuations are consistent with the background only hypothesis in all investigated classes. This proves that even in the case of a vanishing signal the MUSiC algorithm provides consistent and reasonable results. The dedicated CMS analysis [165] is able to establish a 5σ discovery for such a Higgs mass within an integrated luminosity of 1 fb^{-1}.

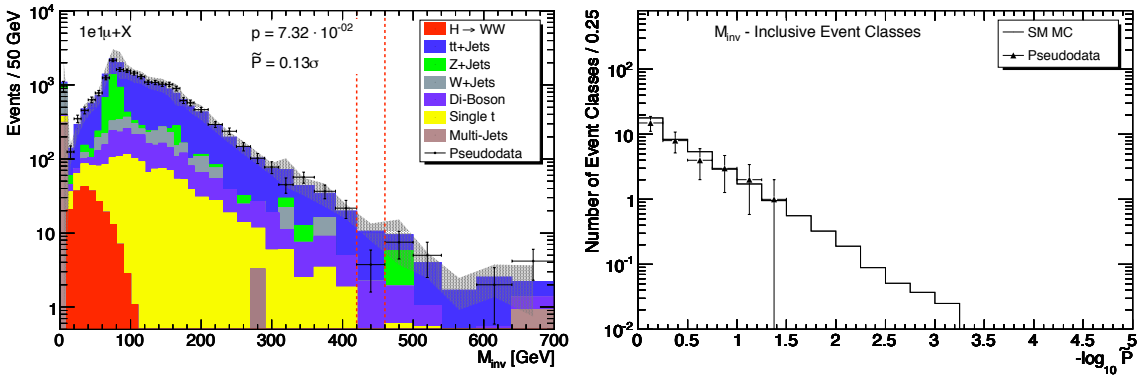

Figure 8.9: ***Left:*** *Event class* $1e\,1\mu + X$ *for a single pseudo-experiment assuming an integrated luminosity of* $1\,\mathrm{fb}^{-1}$*. The tiny Higgs signal is drawn in front of the SM background for better visibility. The* $\widetilde{P}$ *indicates a good agreement between pseudo-data and SM expectation and demonstrates that the algorithm is not sensitive to the Higgs signal in this distribution.* ***Right:*** *The distribution of the* $\widetilde{P}$ *of all inclusive event classes* (M_{inv}) *demonstrate that none of these classes show a deviation from the SM expectation.*

8.5 Signatures of New Physics with many Deviations

One of the benefits of a model-indepent search is its possibility to look at all possible final states at once. While in the previous examples deviations are only detected in one or a few prominent classes, MUSiC is able to provide an overall consistent picture for signatures of new physics within many final states.

Another reason why such a search strategy as presented here might be a good supplement to more conventional signal-driven searches is its generality. Most searches e.g. for supersymmetry or for theories with black holes are highly based on phenomenological models, with many assumptions. In the case of supersymmetry the soft symmetry breaking with the invention of the hidden sector is only introduced in an effective way missing a solid theoretical foundation. Within extra-dimension models predicting mini black holes the lack of knowledge of quantum-gravity is even more striking.

Searches for e.g. supersymmetry face another issue: the large number of unconstrained parameters leads to an almost unlimited parameter space where nature could have picked at most one point. Simplified models based on the minimal supersymmetric extension of the Standard Model like mSUGRA or GMSB reduce the parameter space using several well or not so well-founded physics assumptions. Typical SUSY search strategies at colliders pick some characteristic benchmark scenarios within these phenomenological models.

Thus it might be dangerous to rely solely on analyses optimized on specific SUSY points or black hole models. Model-independent search strategies are a well-suited supplement to overcome these drawbacks of the traditional signal-driven searches. In the following it is demonstrated how supersymmetry using a certain GMSB benchmark point and microscopic black holes within models with large extra dimensions would show up within MUSiC.

As mentioned in section 5.3.4 it might be desirable in the beginning to discard deviations found by the algorithm which suffer from poor statistics. In this context the following

examples are exercised in a statistical fail-safe way: Regions which have $N_{\text{data}} < 3$ or $N_{\text{MC}} < 3$ are discarded.

One should also notice that whenever representative pseudo-experiments are shown, the example is chosen as close as possible to the expected $p_{\text{min}}^{\text{data}}$ mean of many repeated toy experiments assuming signal plus background. The stated $\widetilde{P}$ values therefore represent the expected significance of an "average" CMS experiment.

8.5.1 Microscopic Black Holes

Theories with extra dimensions which effectively lower the Planck scale to the electroweak scale, might bring the phenomena of black holes from the cosmos into the lab. As mini black holes are expected to have masses of at least a TeV, they will leave spectacular signatures within the CMS detector. Due to their small size and thus their large temperature of several 100 GeV, they will immediately decay via Hawking radiation into a multitude of typically ten or more Standard Model particles, but also particles which escape undetected within an extra dimension (see figure 8.10 for a representative event display). Such particles like the graviton would only be indirectly visible as transverse missing energy.

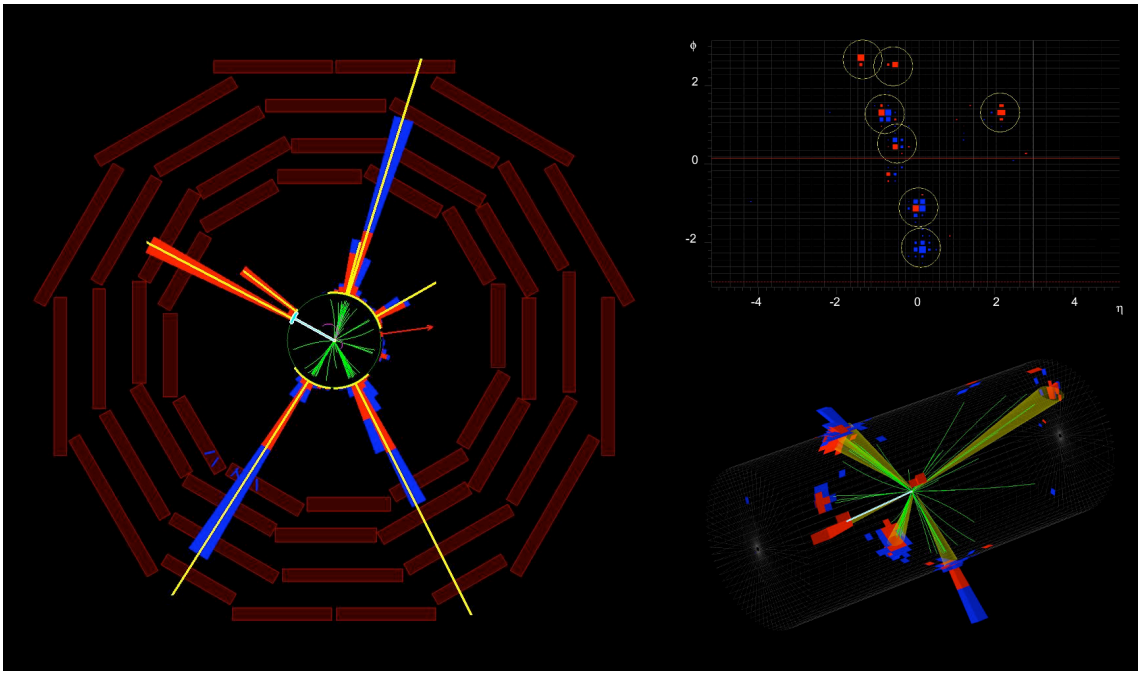

Figure 8.10: *Event display of a typical black hole event with a mass of 4* TeV *and 4 extra-dimensions at a reduced Planck scale of 1* TeV*. The event contains a 430* GeV *electron, a missing transverse energy of 140* GeV *and six high energy jets with momenta of 680, 670, 650, 180, 170, and 140* GeV*.*

As a benchmark point black hole events with a threshold mass of 4 TeV within 4 extra-dimensions and a reduced Planck scale of 1 TeV have been produced utilizing the BlackMax generator [44] fed into the full CMS detector simulation. The scenario has a cross section of 9.17 pb and an integrated luminosity of 100 pb^{-1} is assumed. Due to the large mass of the object and the decay via Hawking radiation the black hole leaves very spectacular

signatures within the detector. This is already reflected by the huge number of event classes being populated, but also by the numerous classes showing a striking deviation from the Standard Model expectation. From the 69 exclusive (46 with $\not{E}_{\mathrm{T}}$) and 239 inclusive event classes (116 with $\not{E}_{\mathrm{T}}$) which contain at least one black hole (to be compared with 95 exclusive (40 with $\not{E}_{\mathrm{T}}$) and 99 inclusive classes (41 with $\not{E}_{\mathrm{T}}$) which are populated with at least one SM event) within an integrated luminosity of 100 pb^{-1} the following classes have at least a 3σ deviation:

- 20 exclusive (43%) and 83 inclusive (43%) event classes in the $\not{E}_{\mathrm{T}}$-distribution
- 32 exclusive (46%) and 148 inclusive (62%) event classes in the M_{inv}-distribution
- 31 exclusive (45%) and 156 inclusive (65%) event classes in the $\sum p_{\mathrm{T}}$-distribution.

The distributions in figure 8.11 and 8.12 display a few representative event classes with significant deviations. Although the main characteristic of these events are the huge multiplicity of final state objects, the discrepancies are already visible within the inclusive one lepton plus one jet distributions. In addition to the lepton which serves as trigger object, the multitude of jets (up to nine jets with $p_{\mathrm{T}} > 60\,\mathrm{GeV}$) leads to at least one jet with a very large transverse momentum. Already these two objects in combination result in a prominent deviation in the tails of the $\sum p_{\mathrm{T}}$ distribution (see 8.11 (left)).

Of course in the case of inclusive classes deviations found are "duplicated" in some way since 1μ 5jet events contribute to 1μ 2jet $+ X$, 1μ 3jet $+ X$ and so on. Nevertheless, in general the inclusive classes are more sensitive to such a signal due to the spread of the events over a variety of exclusive event classes, which are themselves not as significant as the accumulated events within the inclusive classes. Still, there are also various exclusive event classes like the 1μ 5jet class shown in figure 8.11 (right) revealing a black hole signal on top of a small Standard Model background.

As the gravitons which are emitted by the black hole vanish undetected in the extra dimensions, the events contain also a sizeable fraction of missing transverse energy. Consequently, the black hole events can also be spotted within the missing transverse energy distributions. One can see in figure 8.12 (left) that already with a small number of additional particles the $\not{E}_{\mathrm{T}}$-distribution shows a prominent deviation.

Finally, due to the large mass of the black hole also the invariant mass distributions can serve as an indicator for black holes (see figure 8.12 (right)). Thus, black holes would really show up in all variables currently implemented within MUSiC. The number of event classes which reveal a black hole excess is overwhelming.

The manifold of deviations can also be summarized in the distribution of the $\widetilde{P}$ values of all exclusive and inclusive event classes within the $\sum p_{\mathrm{T}}$ distribution as given in figure 8.13. One can see that the pseudo-data with black holes globally disagree with the SM only expectation. Especially within the tails there are huge discrepancies which indicate that such a signal cannot be missed. In general the trend can be seen that the inclusive classes are more prominent than the exclusive classes. From the three considered variables most deviations are present in the $\sum p_{\mathrm{T}}$ variable which will also be the variable investigated first upon the arrival of LHC collision data.

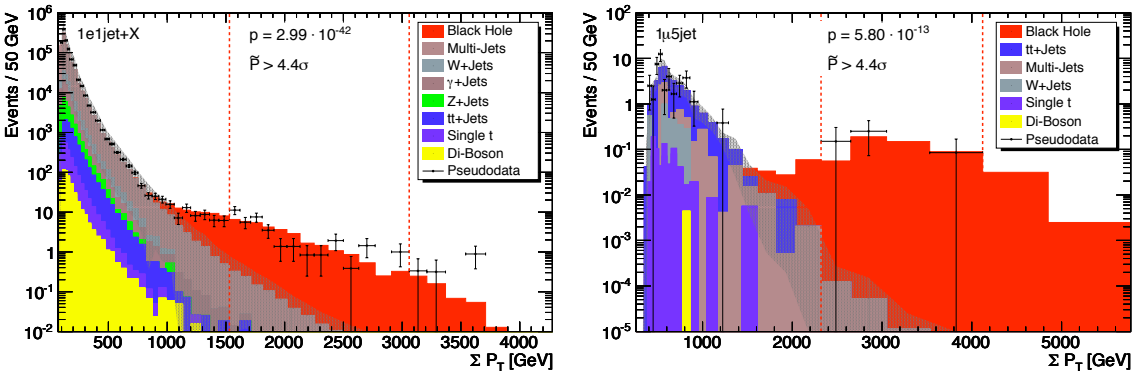

Figure 8.11: *Representative classes with a prominent black hole signal. Many classes show deviations within the $\sum p_{\mathrm{T}}$ distribution. Left: $N_{\mathrm{data}} = 75$, $N_{\mathrm{MC}} = 1.89 \pm 1.92$. Right: $N_{\mathrm{data}} = 4$, $N_{\mathrm{MC}} = 1.33 \cdot 10^{-3} \pm 0.87 \cdot 10^{-3}$.*

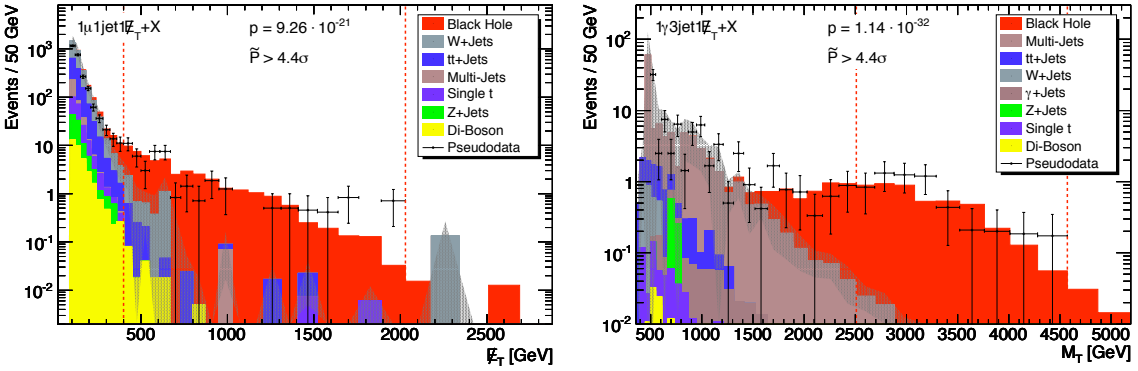

Figure 8.12: *Representative classes with a prominent black hole signal. Due to the graviton vanishing in extra dimensions and due to the huge mass of the black hole, deviations are also seen in the $\not{E}_{\mathrm{T}}$ and M_{inv} distribution. Left: $N_{\mathrm{data}} = 55$, $N_{\mathrm{MC}} = 6.35 \pm 1.78$. Right: $N_{\mathrm{data}} = 24$, $N_{\mathrm{MC}} = 0.25 \pm 0.10$.*

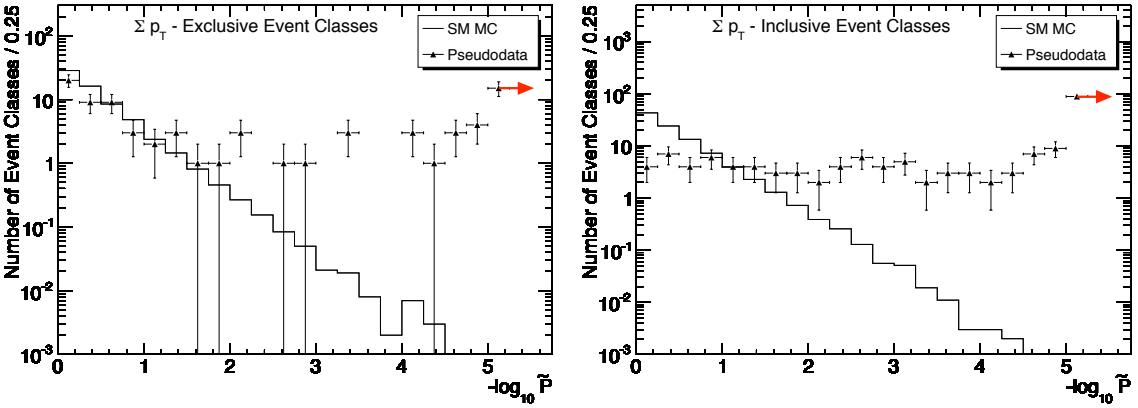

Figure 8.13: *$\widetilde{P}$ distribution of all exclusive (left) and inclusive (right) $\sum p_{\mathrm{T}}$ event classes which contain at least one black hole event. The pseudo-data globally disagree with the SM only expectation in both cases. The last bin (arrow) contains the classes which have a $\widetilde{P}$ of 10^{-5} or less.*

All in all, such a gold-plated signature will lead to alarms all over the place. With the variety of investigated final states, MUSiC obtains a coherent overall picture and could provide hints to disentangle the nature of such a deviation. It could help to discriminate between several models potentially leading to deviations in many distribution e.g. supersymmetry versus black holes.

One should however be aware of one drawback: In order to reconstruct the mass of a black hole all decay fragments need to be taken into account. Since the inclusive classes only consider a fraction of the objects to calculate the variables of interest, the exclusive event classes would be favourable in case of a (threshold) mass reconstruction. However, due to the large spread of the black hole signal over the numerous exclusive event classes the statistical power to reconstruct the mass is diluted. Therefore, once MUSiC would find such a smoking gun, a dedicated search would be initiated to overcome this limitation.

One should notice that this short-coming does not only appear in such a scenario with many particles in the final state. Also in cases with much lower final state multiplicities the mass reconstruction might not be possible in a generic way. Consider the case where two particles are pair-produced as in the case of leptoquark or top pair-production. Even with the detection of all final state variables, MUSiC would not be able to reconstruct the leptoquark or top mass, since the invariant mass of all final state objects would yield (at least for the s-channel) the invariant mass of the propagator and not of the leptoquark/top.

8.5.2 Gauge-mediated Supersymmetry

Another favoured candidate for physics beyond the Standard Model are supersymmetric theories. Within R-parity conserving variants SUSY particles are produced in pairs (mainly consisting of gluinos and squarks), which decay in possibly long chains. Therefore these events lead to spectacular cascades typically with high multiplicities of leptons/photons, jets and a large amount of $\not\!\!E_{\mathrm{T}}$ due to the lightest supersymmetric particles (LSP), which escapes undetected. So unlike single resonance production as for example $Z \rightarrow \mu\mu$, SUSY does not predominantly favour a single topology, but does contribute to a multitude of event classes within MUSiC. Therefore, a model-unspecific search can provide a consistent picture of SUSY particles appearing on top of the Standard Model prediction. The combination of significant deviations found in several classes could provide additional evidence and might help to establish the supersymmetric nature present in the data.

Since it would contradict the basic philosophy of MUSiC to really perform a large supersymmetry parameter scan the search results are highlighted using a typical benchmark point. The CMS point GM1c within a minimal gauge-mediated supersymmetric model with the following parameters is chosen:

- **GM1c:** $\Lambda = 100$ TeV; $M_m = 2\Lambda$; $\tan\beta = 15$; $N_5 = 1$; $\mathrm{sgn}(\mu) = 1$; $C_{\mathrm{G}} = 1$; σ (LO) $= 843$ fb;

The chosen point is characterized by the decay of the next to lightest supersymmetric particle (NLSP), which is in this case the neutralino $\tilde{\chi}_0^1$ with a mass of 140 GeV. Due to the smallness of C_{G} the neutralino decays always almost immediately into a photon and gravitino. The decay into a Z plus gravitino is possible, but suppressed to the per

mille level due to the bino-like nature of the NLSP. As the gravitino can only be detected indirectly, the typical signature of this scenario are two photons plus a significant amount of missing transverse energy. In addition further jets from the decay of the initial squarks or gluinos are present.

In the following this GMSB point will be discussed. A similar study which investigates a gravity mediated SUSY scenario (mSUGRA) including a more comprehensive analysis of different points as well as a comparison to more model-specific analyses can be found in [166]. All in all the results follow the expectations: With a dedicated search optimized for the specific point the expected significances are higher – or correspondingly, the required discovery luminosities are lower. MUSiC performs best when several channels are combined into a comprehensive review in accordance with its strategy.

The Global Search

A global scan has been performed on the 42 exclusive (29 with $\not\!\!E_{\mathrm{T}}$) and 198 inclusive event classes (96 with $\not\!\!E_{\mathrm{T}}$) which contain at least one GMSB event within 250 pb^{-1}. This compares to 225 (106 exclusive, 119 inclusive) event classes which are populated with at least one Standard Model event within the same amount of data. Already these numbers indicate that SUSY is present in a large part of the data, thus many different topologies could give rise to a SUSY signal. The following classes have an expected deviation of at least 3σ:

- 8 exclusive (28%) and 31 inclusive (32%) event classes in the $\not\!\!E_{\mathrm{T}}$-distribution
- 7 exclusive (17%) and 50 inclusive (25%) event classes in the M_{inv}-distribution
- 7 exclusive (17%) and 56 inclusive (28%) event classes in the $\sum p_{\mathrm{T}}$-distribution.

Basically all variables provide more or less similar results, although classes with missing transverse energy clearly dominate and thus also a relative large amount of them show a deviation in the $\not\!\!E_{\mathrm{T}}$ distribution. This is underlined by the fact that all significant exclusive classes contain $\not\!\!E_{\mathrm{T}}$ in addition to at least one photon (the full list of deviations is listed in table 8.2). Unfortunately experiences from past accelerator experiments show that $\not\!\!E_{\mathrm{T}}$ will be difficult to control and understand in the first data. Thus it might be desirable to also investigate classes without $\not\!\!E_{\mathrm{T}}$ which show indications of SUSY.

Due to the diversity of SUSY, there is not a single featured variable as in the case of resonances where the invariant mass is most suitable. As in the context of MUSiC all particles of the event class are combined to M_{inv} (M_{T}), not necessarily the correct particle combinations are found to produce resonance peaks. In addition to this the LSP distorts the picture such that only transverse masses can be constructed. The LSP in SUSY events leads to a considerable amount of missing transverse energy. Thus when analysing event classes with photons, jets and $\not\!\!E_{\mathrm{T}}$ the separation between SM and SUSY is prominent within the $\not\!\!E_{\mathrm{T}}$ variable. However, this variable does not include the momenta of the possibly many additional objects within the class. Therefore, the variable $\sum p_{\mathrm{T}}$ finally might be the golden mean between generality and sensitivity.

Discussion of Selected Event Classes

In the following a few representative event classes with significant discrepancies are highlighted. Table 8.2 lists all exclusive event classes which lead on average to a deviation of more than 3σ. As expected the classes contain at least one photon and missing energy, which reflect the decay of the neutralino as NLSP into a photon and a gravitino. Additional jets might stem from previous decays of the initially produced squarks and gluinos.

Event Class	Distribution	N_{GMSB}	N_{SM}	$p^{\mathrm{min}}_{\mathrm{data}}$ (expected)	$\widetilde{P}$ (expected)
2γ $\not\!\!E_{\mathrm{T}}$	$\sum p_{\mathrm{T}}$, M_{inv}, $\not\!\!E_{\mathrm{T}}$	3.71	0.11	3e-6	$\leq$ 4e-5
2γ 1jet $\not\!\!E_{\mathrm{T}}$	$\sum p_{\mathrm{T}}$, M_{inv}, $\not\!\!E_{\mathrm{T}}$	4.45	0.16	0.2e-6 – 7e-6	$\leq$ 1e-5
2γ 3jet $\not\!\!E_{\mathrm{T}}$	$\sum p_{\mathrm{T}}$, M_{inv}, $\not\!\!E_{\mathrm{T}}$	3.01	0.18	0.08e-4 – 1e-4	$\leq$ 5e-4
2γ 4jet $\not\!\!E_{\mathrm{T}}$	$\sum p_{\mathrm{T}}$, M_{inv}, $\not\!\!E_{\mathrm{T}}$	3.64	0.03	7e-6, 1e-5, 2e-11	$\leq$ 2e-5
2γ 5jet $\not\!\!E_{\mathrm{T}}$	$\sum p_{\mathrm{T}}$, M_{inv}, $\not\!\!E_{\mathrm{T}}$	3.77	0.01	5e-9,5e-9,1e-11	$<$ 1e-5
1γ 5jet $\not\!\!E_{\mathrm{T}}$	$\not\!\!E_{\mathrm{T}}$	7.87	10.40	8e-6	1e-4
1γ 6jet $\not\!\!E_{\mathrm{T}}$	$\sum p_{\mathrm{T}}$, M_{inv}, $\not\!\!E_{\mathrm{T}}$	6.57	0.56	3e-7, 2e-7, 6e-10	$\leq$ 3e-5
1γ 7jet $\not\!\!E_{\mathrm{T}}$	$\sum p_{\mathrm{T}}$, M_{inv}, $\not\!\!E_{\mathrm{T}}$	4.03	0.04	5e-6, 6e-6, 2e-10	$\leq$ 4e-5

Table 8.2: *List of all exclusive event classes with a deviation of at least 3σ for the GMSB SUSY point GM1c assuming 250 pb^{-1} of data. The last two columns state expected p-value and significance when repeating S+B and B hypotheses multiple times.*

Analysing the many event classes which are above the 3-sigma threshold there are of course some topologies with spectacular particle multiplicities, e.g. 1γ 7jet $\not\!\!E_{\mathrm{T}}$. Obviously one cannot expect the Monte Carlo prediction to perfectly match the measured data in these extreme kinematic regions. It is clear that such deviations found by the algorithm have to be taken with care. There is always a second (non-automated) step needed where the physicist with all his/her experience and bias interprets the results. This includes looking at the interesting class in detail, evaluating possible SM contributions missing so far, theoretical uncertainties of the MC prediction or possible detector effects causing the deviation.

However, also these "exotic" classes are worth looking at since the spectacular SUSY decays will populate them while the SM (including its uncertainties) is almost negligible. Thus one should not simply discard these classes arguing that Monte Carlo will never work here. If there are 100 events in the 6 muon channel where the SM is close to zero something interesting is going on which can very likely not be explained by the SM alone or by MC not working properly. Note that for bins with pseudo-data entries where (given the limited statistics) not a single MC event is predicted, a conservative 68% upper Poisson limit of 1.15 events is used. This upper limit is applied to all samples contributing to this specific region and then scaled according to the assumed luminosity.

In general the biggest discrepancies between pseudo-data and MC expectation can be found in the inclusive 2γ $\not\!\!E_{\mathrm{T}}$ $+X$ class shown in figure 8.14 (left). This class reveals a very prominent GMSB signal on top of a very small SM background over the whole range of the distribution. It is not surprising that this final state is also the most favoured within the dedicated searches hunting for GMSB.

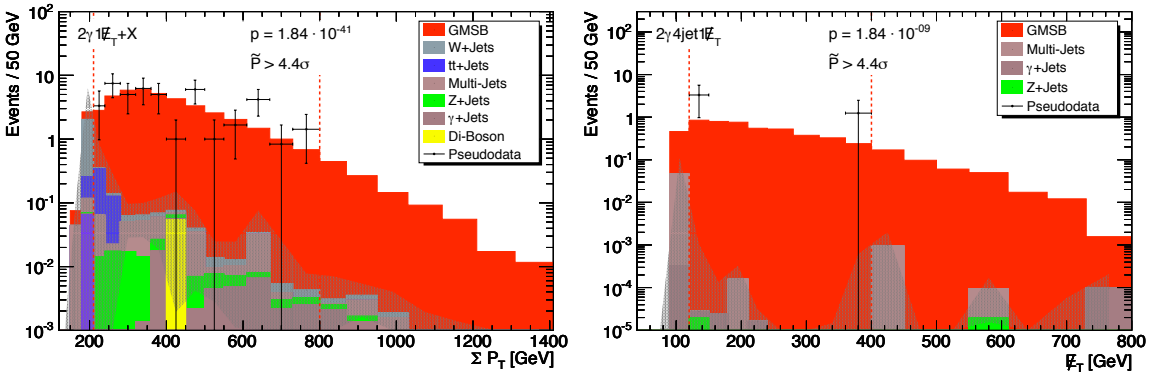

Figure 8.14: *Results of representative pseudo experiments which contain GM1c as signal assuming* 250 pb^{-1}, *using event classes with* $\not\!\!E_{\mathrm{T}}$. *Left:* $N_{\mathrm{data}} = 39$, $N_{\mathrm{MC}} = 0.68 \pm 0.27$. *Right:* $N_{\mathrm{data}} = 3$, $N_{\mathrm{MC}} = (1.53 \pm 1.17) \cdot 10^{-3}$.

As the statistical combination of inclusive classes might be difficult it is also worth mentioning that also the exclusive classes although not in such a prominent way are able to spot the GMSB signal. A representative distribution is given in figure 8.14 (right) with the 2γ 4jet $\not\!\!E_{\mathrm{T}}$ class. The main background left are multi-jet events whose contribution needs to be estimated from data. Of course the results need to be interpreted with care, but even if the amount of multi-jet events is larger by orders of magnitude the class is still worth to study.

The missing transverse energy distributions could be problematic, especially with the early data where this variable is not perfectly understood. In this context figure 8.15 shows two inclusive event classes which do not use $\not\!\!E_{\mathrm{T}}$ at all. The left plot refers to the 2γ 3jet$+X$ class. The right plot refers to the quite extraordinary 1μ 1γ 2jet$+X$ event class. Within the decay of the primarily produced squarks and gluinos also W-bosons might be emitted leading also to events with muons and electrons. Requiring a single photon the class might pick up events where only one of the photons is within the detector acceptance or where only one photon could be identified. As high energetic photons together with a muon do not appear very often within Standard Model processes there is a significant excess of GMSB events in the tails of the distributions. Therefore such classes might provide a promising alternative to the classical signature of photon plus missing transverse energy.

8.6 Possible MUSiC Extensions

There are various possibilities how to extend the MUSiC analysis. The number of investigated variables can be enlarged for example by looking at further kinematic variables such as angles between decay products. Additionally, further objects like τ's could be added or a more fine-grained object classification could be performed by e.g. distinguishing heavy flavour jets like b-jets from light quark or gluon jets. However, one should keep in mind that any additional object or distribution increases the trial factor. One should also note that several hundred distributions might still be looked at by eye, but this will miserably

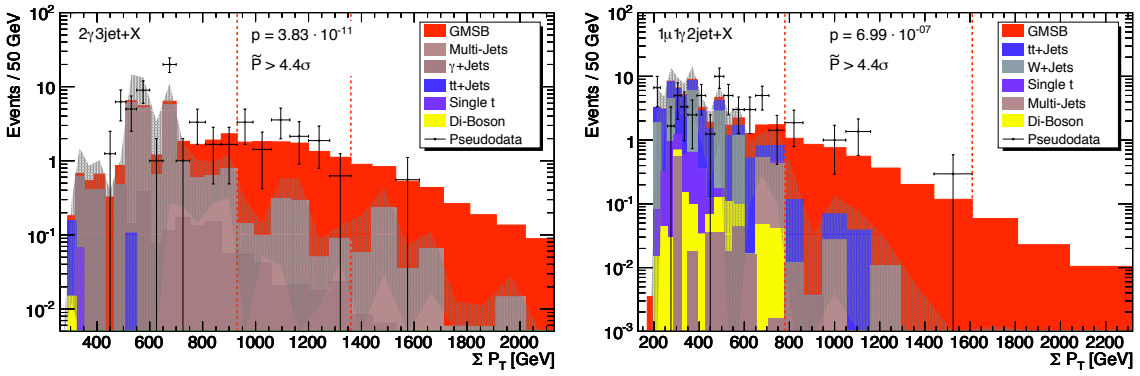

Figure 8.15: *Results of representative pseudo experiments which contain GM1c as signal assuming* 250 pb^{-1}, *using event classes without* $\not\!\!E_{\mathrm{T}}$. *Left:* $N_{\mathrm{data}} = 18$, $N_{\mathrm{MC}} = 1.37 \pm 0.56$. *Right:* $N_{\mathrm{data}} = 9$, $N_{\mathrm{MC}} = 0.44 \pm 0.39$.

fail for thousands of distributions. Following the MUSiC guidelines the goal should therefore be to have a minimum set of variables and distributions with the largest potential of spotting various deviations from a detector effect to physics beyond the Standard Model. Further, it is possible to implement different algorithms to measure the differences between data and MC expectation.

8.6.1 Charges and Hypothesis Ranking

One other possibility to achieve the goal of a minimal set of distributions with a maximal potential for spotting various deviations is to add only a very few selected event classes based on some characteristics of the Standard Model. For example one can exploit the fact that within the Standard Model only a very limited amount of events contain leptons with the same electric charge (same sign leptons). Processes which are dominant within the multi-lepton classes like Drell-Yan or WW-production decay into leptons with opposite charged leptons (opposite sign leptons). Therefore this difference can be used to increase the significance to certain detector effects related to the charge measurement or signatures beyond the Standard Model with same sign leptons within the final state.

In order to keep the addition of "charges" within MUSiC as simple as possible a new physics object q is introduced. It measures the absolute value of the sum of the lepton charges (electrons and muons). This reduces the number of additional classes to an absolute minimum following the concept of a hypothesis ranking. The single lepton classes, which contain most of the events and which are insensitive to e.g. same sign lepton signals stay unchanged, while the di- and tri-lepton classes are split into two sub-classes (di: $0q$ and $2q$, tri: $1q$ and $3q$).

The benefit of such an extension can be seen perfectly with supersymmetry as a benchmark signal with its long decay chains and resulting multi-lepton events. Here the CMS low mass LM0 point ($m_{1/2} = 160$ GeV, $m_0 = 200$ GeV, $A_0 = -400$ GeV, $\mu > 0$, $\tan\beta = 10$) with a leading order cross section of 110 pb has been utilized to perform a scan at an integrated luminosity of 100 pb^{-1}. Figure 8.16 shows the effect of the splitting of one typical event class when introducing charges as a new final state object.

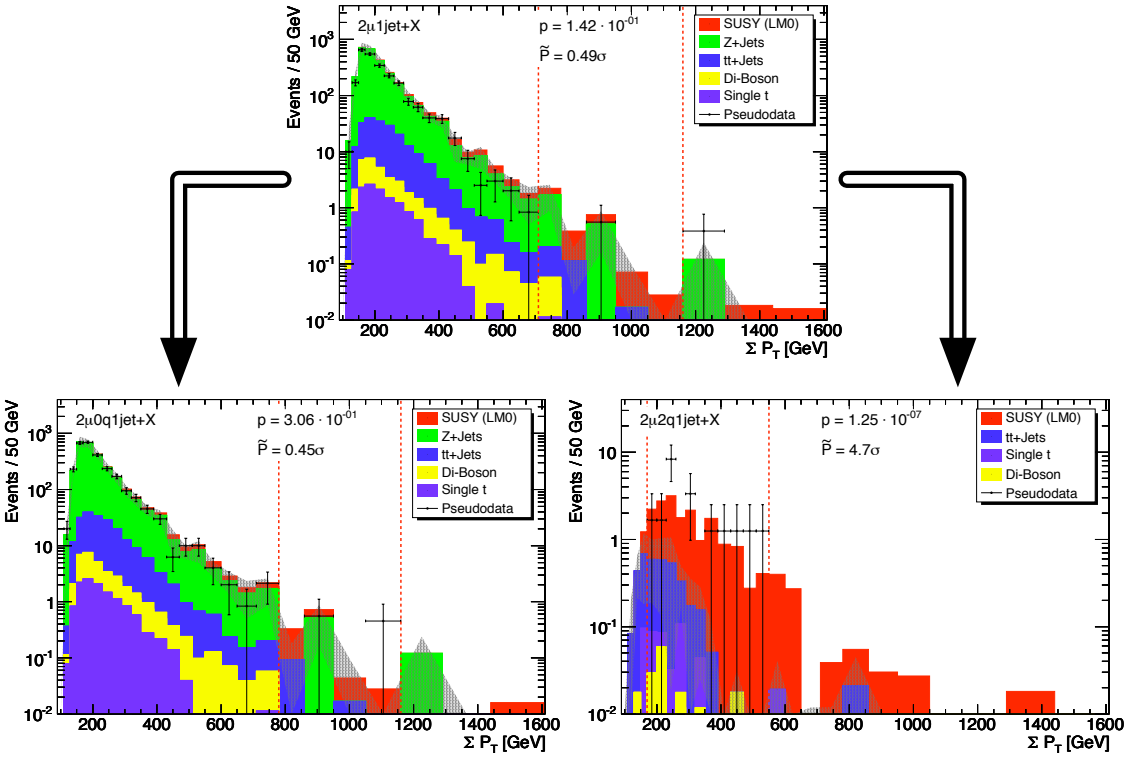

Figure 8.16: *Splitting of an event class into two classes with same-sign and opposite-sign leptons. While the signal (supersymmetric events from the LM0 benchmark point) is not visible in the plot without charge separation* ***(top)*** *and in the same-sign lepton plot* ***(left)****, a deviation of almost* 5σ *is present in the like-sign event class* ***(right)*** *(*$N_{\mathrm{data}} = 14, N_{\mathrm{SM}} = 1.64 \pm 0.54$*).*

The top plot shows the inclusive two muon plus one jet class, which is completely dominated by Z+Jets events. The low $\widetilde{P}$ value states that the pseudo-data are in perfect agreement with the MC expectation. The situation looks completely different when separating the events into same sign and opposite sign final states. All Z+Jets events end up in the opposite sign class, while a clear access of the supersymmetric signal over a very small Standard Model background can be spotted in the same sign class as indicated by the $\widetilde{P}$ value.

Chapter 9

Conclusions

This work describes the implementation and working principle of the model-independent search MUSiC which has been carried out as feasibility study within CMS. Being the first analysis of this kind which is ready to absorb data before the actual start-up of the experiment, it has a great potential to speed up the understanding of the detector, to re-discover the Standard Model and to reveal possible signatures of new physics within the data.

Without a focus on a specific signal within or beyond the Standard Model the approach is complementary to the traditional signal-driven analyses. Instead of an optimization of the event selection with respect to a certain signal, a model-independent analysis investigates all events without prejudice. Requiring a solid object identification, the events are classified into event classes according to their particle content. A broad data versus Standard Model MC comparison is performed by scanning variables which are sensitive for deviations from the Standard Model within each of the event classes. This general strategy is sensitive to a very broad spectrum of deviations which are illustrated within this study using benchmark scenarios. Representative examples demonstrate the feasibility to spot flaws in the reconstruction software, the sensitivity to spot detector malfunctions or the ability to reveal gold-plated signatures beyond the Standard Model. It might help in the tuning of the event generators and could aid in cross checking results from other groups like efficiencies in a broader context.

The manifold of possible origins for deviations underlines the fact that such an analysis tool cannot be used as a "standalone discovery machine", but needs careful steering and the results require a thought-full interpretation from physicists. Therefore, MUSiC can be seen as a physics alarm system similar to the data quality monitoring, but at a different level with a focus towards the probing of the Standard Model. Following the saying "Expect the Unexpected", the analysis covers a broad range of existing models, but also models not yet invented, and serves as an insurance not to miss anything.

Such a computing demanding analysis would not be possible without the developments in computer science. With the grid computing the LHC experiments start a new era of distributed and decentralized computing, served in a manner known from the power grid. A dedicated hierarchy of computing centres have been built up to allow fully distributed analysis chain. Data are taken, stored and reconstruction at CERN's Tier-0, further dis-

tributed to seven Tier-1s, where a second copy of the raw data and reconstructed data are stored. Tier-1 are also used for regular re-processing of the data with improved alignment and calibration and the extraction of skims which are further transferred to the Tier-2s where the ordinary physicist performs his grid-based analysis.

Within each of the MUSiC steps a grid based approach allows to parallelize the computing intensive tasks for a fast analysis turn-around. Thus for model-independent searches the grid is an irreplaceable tool.

The work demonstrates the readiness of MUSiC awaiting eagerly the arrival of the first proton-proton collisions to happen this year. An exciting time is ahead of us and likely the LHC and its experiments – hopefully with the help of MUSiC – will improve our understanding of particle physics significantly.

Appendix A

Units, Variables, and Coordinates

At this place units and conventions, which are used in this thesis, are stated. Instead of the *International System of Units* (SI-units) variables are given in the natural units of elementary particle physics by setting

$$\hbar \equiv 1 \qquad \text{and} \qquad c \equiv 1 \tag{A.1}$$

instead of

$$\hbar = 1.0546 \cdot 10^{-34}\ \text{Js} \qquad \text{and} \qquad c = 2.9979 \cdot 10^{8}\ \text{m/s}\ .$$

Since the energies in particles physics are tiny compared to daily life ones, physicists defined the unit of an "electron-volt", short eV. It is the energy gained by a particle carrying one elementary electric charge while moving through an electric field with a potential difference of one volt, thus

$$1\ \text{eV} \ = \ 1.6022 \cdot 10^{-19}\ \text{J}. \tag{A.2}$$

By convention (A.1) all units can be expressed in terms of electron-volt, like distances (eV^{-1}), times (eV^{-1}), masses (eV) or momenta (eV).

The global CMS coordinate system is introduced here, which is used when no other coordinate system is explicitly quoted. The cartesian system is defined with the x-axis pointing towards the center of the LHC ring and perpendicular, directed skywards to the surface, the y-axis. The z-axis completes a right-handed system along the beam axis (see figure A.1). The polar coordinates ϕ / θ are defined in the xy-plane / yz-plane referring to the x-axis / y-axis, respectively:

$$\tan\phi = \frac{y}{x} \qquad \text{and} \qquad \cos\theta = \frac{z}{\sqrt{x^2+y^2}}. \tag{A.3}$$

Within this coordinate system "transverse" variables, tagged by a subscript "T" like in p_{T} , are defined as the absolute value of the projection of the variable (as vector) onto the xy-plane. The "longitudinal" component, denoted by "L", is the absolute value of the projection along the z-axis.

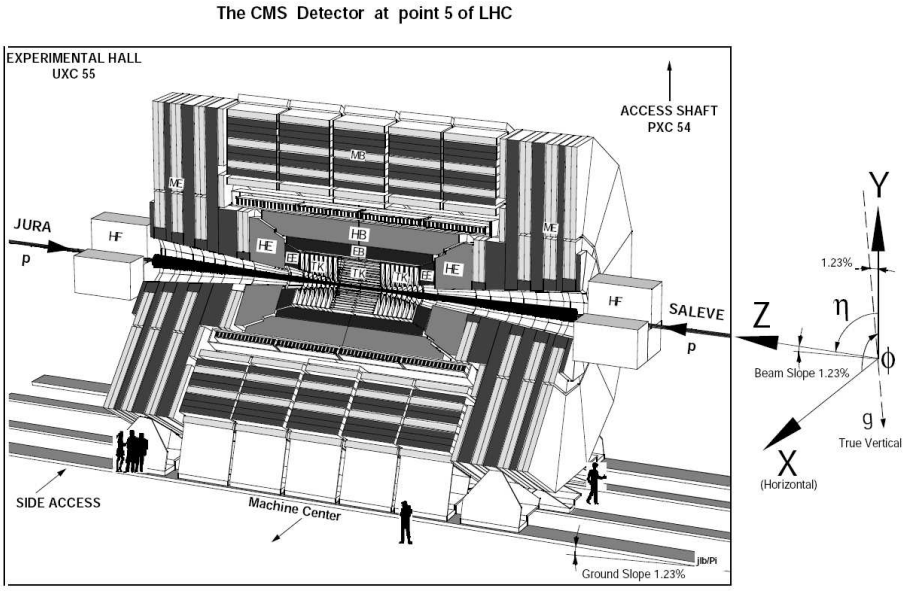

Figure A.1: *The CMS detector with the CMS global coordinate system [60].*

For a particle with a mass m, energy E and longitudinal momentum component p_{L} the "rapidity" y replaces as natural coordinate in high energy physics the polar angle θ in the following way

$$y := \frac{1}{2} \ln \left(\frac{E + p_{\mathrm{L}}}{E - p_{\mathrm{L}}} \right) . \tag{A.4}$$

It benefits from the fact that a difference in rapidity Δy is invariant under boosts along the z-axis, for example the distribution $\mathrm{d}N/\mathrm{d}y$ is unchanged. For practical issues the rapidity is approximated in the limit $m \ll E$ by the "pseudorapidity" η

$$\eta := -\ln \tan \left(\frac{\theta}{2} \right) . \tag{A.5}$$

Depending only on θ the pseudorapidity η can also be defined for particles with an unknown mass.

Appendix B

CMS Software & Datasets

The CMS software framework is used in order to process the simulated samples and to reconstruct the physics objects, using version CMSSW_2_2_9 [131]. The MUSiC framework is based on the official CMS Physics Analysis Toolkit (PAT). All samples are generated at a centre of mass energy of 10 TeV with the full detector simulation and originate from the MC production during the summer of 2008 and the winter of 2009. These samples have been generated and simulated using version CMSSW_2_1_17. For the digitization, the trigger-simulation and the reconstruction version CMSSW_2_2_3 is utilized. The physics objects are reconstructed assuming ideal conditions i.e. a perfectly aligned and calibrated detector.

The full list of datasets used within this thesis is given below. The signal datasets are in table B.1, while the backgrounds are listed in table B.2.

Process	$\sigma_{\mathrm{LO}}(fb)$	# events	path in dbs
SUSY LM0	110e3	2e5	/SUSY_LM0-sftsht/Summer08_IDEAL_V11_v1
SUSY GM1c	843	1e5	/GMSB_GM1c/Summer08_IDEAL_V11_redigi_v1
Higgs $\rightarrow WW$	710	1e5	/H160_WW_2l/Summer08_IDEAL_V11_redigi_v1
Black Hole (4 TeV)	9.2e3	2e4	private production (BlackMax)

Table B.1: *Used signal samples, together with their leading order cross sections, the number of produced events and the official CMS dataset path.*

Process	$\sigma_{\mathrm{LO}}(fb)$	# events	path in dbs
Photon+Jets	2.89e8	9e5	/PhotonJetPt15/Summer08_IDEAL_V12_redigi v1
	3.22e7	9e5	/PhotonJetPt30/Summer08_IDEAL_V12_redigi_v1
	1.01e6	8e5	/PhotonJetPt80/Summer08_IDEAL_V12_redigi_v1
	5.14e4	9e5	/PhotonJetPt170/Summer08_IDEAL_V12_redigi_v1
	4.19e3	1e6	/PhotonJetPt300/Summer08_IDEAL_V12_redigi_v1
	4.52e2	1e6	/PhotonJetPt470/Summer08_IDEAL_V12_redigi_v1
	2.00e1	1e6	/PhotonJetPt800/Summer08_IDEAL_V12_redigi_v1
	0.27	1e6	/PhotonJetPt1400/Summer08_IDEAL_V12_redigi_v1
	1.5e-3	1e6	/PhotonJetPt2200/Summer08_IDEAL_V12_redigi_v1
QCD	1.46e12	7e6	/QCDpt15/Summer08_IDEAL_V11_redigi_v3
	1.09e11	3e6	/QCDpt30/Summer08_IDEAL_V11_redigi_v1
	1.93e9	3e6	/QCDpt80/Summer08_IDEAL_V11_redigi_v1
	6.26e7	3e6	/QCDpt170/Summer08_IDEAL_V11_redigi_v1
	3.66e6	3e6	/QCDpt300/Summer08_IDEAL_V11_redigi_v1
	3.16e5	3e6	/QCDpt470/Summer08_IDEAL_V11_redigi_v1
	1.19e4	3e6	/QCDpt800/Summer08_IDEAL_V11_redigi_v2
	1.72e2	5e5	/QCDpt1400/Summer08_IDEAL_V11_redigi_v1
	1.42	2e6	/QCDpt2200/Summer08_IDEAL_V11_redigi_v2
	8.60e-3	5e5	/QCDpt3000/Summer08_IDEAL_V11_redigi_v1
Z+Jets	3.7e6	1e6	/ZJets-madgraph/Summer08_IDEAL_V11_redigi_v1
W+Jets	4.0e7	9e6	/WJets-madgraph/Summer08_IDEAL_V11_redigi_v1
TT+Jets	3.17e5	1e6	/TTJets-madgraph/Fall08_IDEAL_V11_redigi_v10
W+2 Photons	10.4	1e5	/Wgg-madgraph/Fall08_IDEAL_V11_redigi_v1
Z+2 Photons	5.1	1e5	/Zgg-madgraph/Fall08_IDEAL_V11_redigi_v1
WW inclusive	4.48e4	2e5	/WW/Summer08_IDEAL_V11_redigi_v1
ZZ inclusive	7.1e3	2e5	/ZZ/Summer08_IDEAL_V11_redigi_v1
WZ inclusive	1.74e4	2e5	/WZ_incl/Summer08_IDEAL_V11_redigi_v1
DrellYan $\mu\mu$	1.10e3	1e4	/DYmumuM200/Summer08_IDEAL_V11_redigi_v2
	44.88	1e4	/DYmumuM500/Summer08_IDEAL_V11_redigi_v2
	2.55	1e4	/DYmumuM1000/Summer08_IDEAL_V11_redigi_v2
	5.58e-2	1e4	/DYmumuM2000/Summer08_IDEAL_V11_redigi_v2
W $e\nu$		5e4	dedicated samples from the W' Working Group
SingleTop t	5.53e4	3e5	/SingleTop_tChannel/Summer08_IDEAL_V11_redigi_v3
SingleTop tW	2.73e4	2e5	/SingleTop_tWChannel/Summer08_IDEAL_V11_redigi_v3
SingleTop s	1.66e3	1e4	/SingleTop_sChannel/Summer08_IDEAL_V11_redigi_v3

Table B.2: *Used RECO background samples (mainly from Summer08 and Fall08 production) with their leading order cross sections, the number of produced events and the official CMS dataset path.*

Appendix C

Parton Distribution Function Uncertainty Determination

With the start of the Large Hadron Collider (LHC) high energy physics will enter a new regime: the energy frontier at the TeV-scale. As a proton-proton collider, providing a broad spectrum of parton-parton centre of mass energies, it is well-designed as a discovery machine revealing scenarios of new physics beyond the Standard Model (SM). However the discovery of new physics requires a detailed theoretical understanding of the Standard Model and its uncertainties. Besides knowing the cross sections only at a limited order of perturbation theory and a unphysical dependence on the factorization and renormalization scale, one of the major uncertainties is the limited knowledge of the distribution of partons within the proton. This distribution if described by Parton Distribution Functions (PDF). The enormous importance of the parton distribution functions is obvious by looking at the general cross section formula at a proton-proton collider

$$\sigma(pp \to X) = \sum_{i,j} \mathrm{PDF}_{i,p}(x_1, f_1, Q) \otimes \mathrm{PDF}_{j,p}(x_2, f_2, Q) \otimes \sigma_{ij\to X}(Q') \qquad \text{(C.1)}$$

The parton distribution functions $\mathrm{PDF}(x, f, Q)$ represent the probability to find a parton of the flavour f with a momentum fraction x at a given scale Q (factorization scale). In order to calculate a cross section at a hadron collider, this PDF needs to be folded with the partonic cross section $\sigma_{ij\to X}$. Due to the symmetric setup of the LHC the parton i with flavour f, momentum x might either stem from one or the other proton. The partonic cross section $\hat{\sigma}$ depends on a scale Q' referred to as renormalization scale. Due to the limited knowledge of the perturbation expansion (LO, NLO, ...) possible divergent terms might arise. Since these divergencies are unphysical, they need to be canceled by a suitable renormalization of the physical constants (e.g. couplings) at a certain (unphysical) scale, the renormalization scale Q'.

Any uncertainty on the PDFs propagates into an uncertainty on the cross section. Its determination within the CMS framework CMSSW will be outlined in this note. After a short reminder of how PDFs are obtained, the commonly used brute force method and an alternative reweighting method are described in detail. It will be demonstrated how the relevant inputs for those uncertainty determination methods are retrieved from the CMS

Monte Carlo data formats. Finally some uncertainties for the Summer08/Winter09 Monte Carlo production are quoted, comparing both methods. For further details see [157].

C.1 Best-Fit and Error PDFs

As pioneered by the CTEQ group [167–170] and adopted by the MRST/MSTW group [57], not only the "best-fit" PDF is provided, but in addition a set of "error PDFs", which can be used to propagate the experimental uncertainties in the global PDF determination. The so called Hessian method based on the linear error propagation is used: having a fit with n free parameters a_i and assuming that the goodness of fit χ^2 distribution can be expanded quadratically around the global minimum at a_i^0 one can write

$$\Delta\chi^2 = \chi^2 - \chi^2_{min} = \sum_{i,j}^{n} H_{ij}(a_i - a_i^0) \cdot (a_j - a_j^0). \tag{C.2}$$

Diagonalizing the Hessian matrix $H_{ij} = \frac{\partial^2\chi}{2\partial a_i \partial a_j}\Big|_{min}$ one can determine a set of n independent eigenvectors with their corresponding eigenvalues. By shifting the parameters along the eigenvectors one obtains $2n$ error PDFs representing the "up/down" variation of the n parameters.

Figures C.1 and C.2 show the PDF distributions as a function of the momentum fraction x for the latest global PDF fits at NLO obtained by the MSTW and the CTEQ group at two energy scales $Q = 100$ GeV and $Q = 1$ TeV . The relative uncertainties on the parton distribution functions for the different partons are also shown. These uncertainties have been calculated utilizing the set of $2n$ error PDFs and the best-fit PDF as described in section C.2.

The values for the different parton distribution functions provided by the PDF builders like CTEQ or MRST/MSTW can be accessed via the LHAPDF (Les Houches Accord Parton Distribution Function) library [158]. Although written in FORTRAN it is delivered with a C++ wrapper for direct calls within e.g. CMSSW code. LHAPDF is distributed with the CMS software stack and thus can be used not only by the event generators as in the brute force method, but also at end-user analysis level as described in the reweighting method.

C.2 PDF Uncertainty Determination

This section describes how the best-fit PDF in combination with the error-varied PDFs can be used to estimate the uncertainty on variables of interest such as cross sections, but also distributions of a final variable of an analysis.

C.2.1 The Brute Force Method

The method as suggested by its label relies on a decent amount of computing power: one generates the events of interest $M = 2n + 1$ times, where one varies the PDF from run to

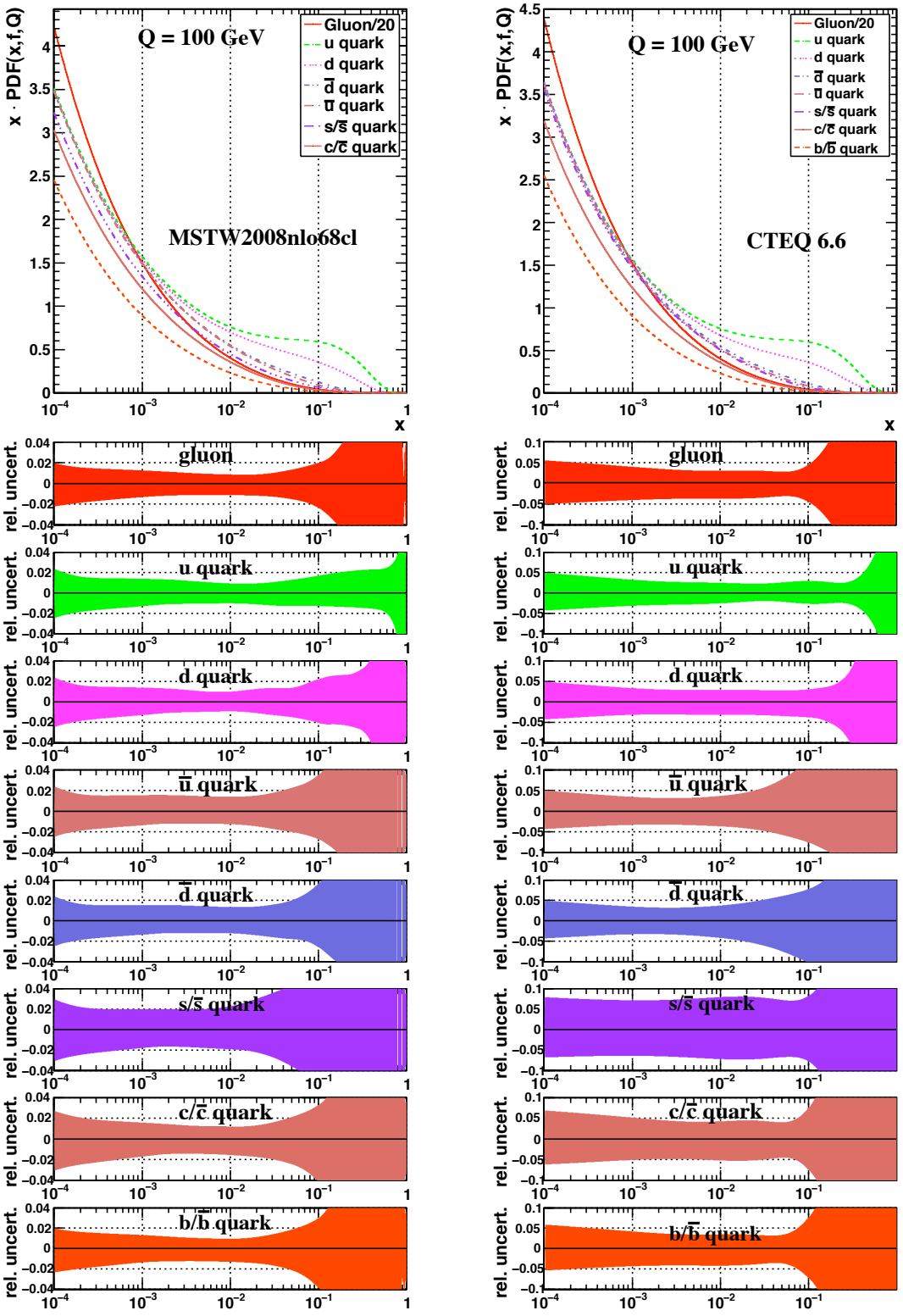

Figure C.1: *Parton distribution functions and their relative uncertainty at the factorization scale $Q = 100\,\text{GeV}$ obtained by the MSTW (left) and the CTEQ (right) group. The values from the error varied PDFs have been fed into the master formula to calculate the relative uncertainties. Note the differences in the y-axis range for the relative uncertainties.*

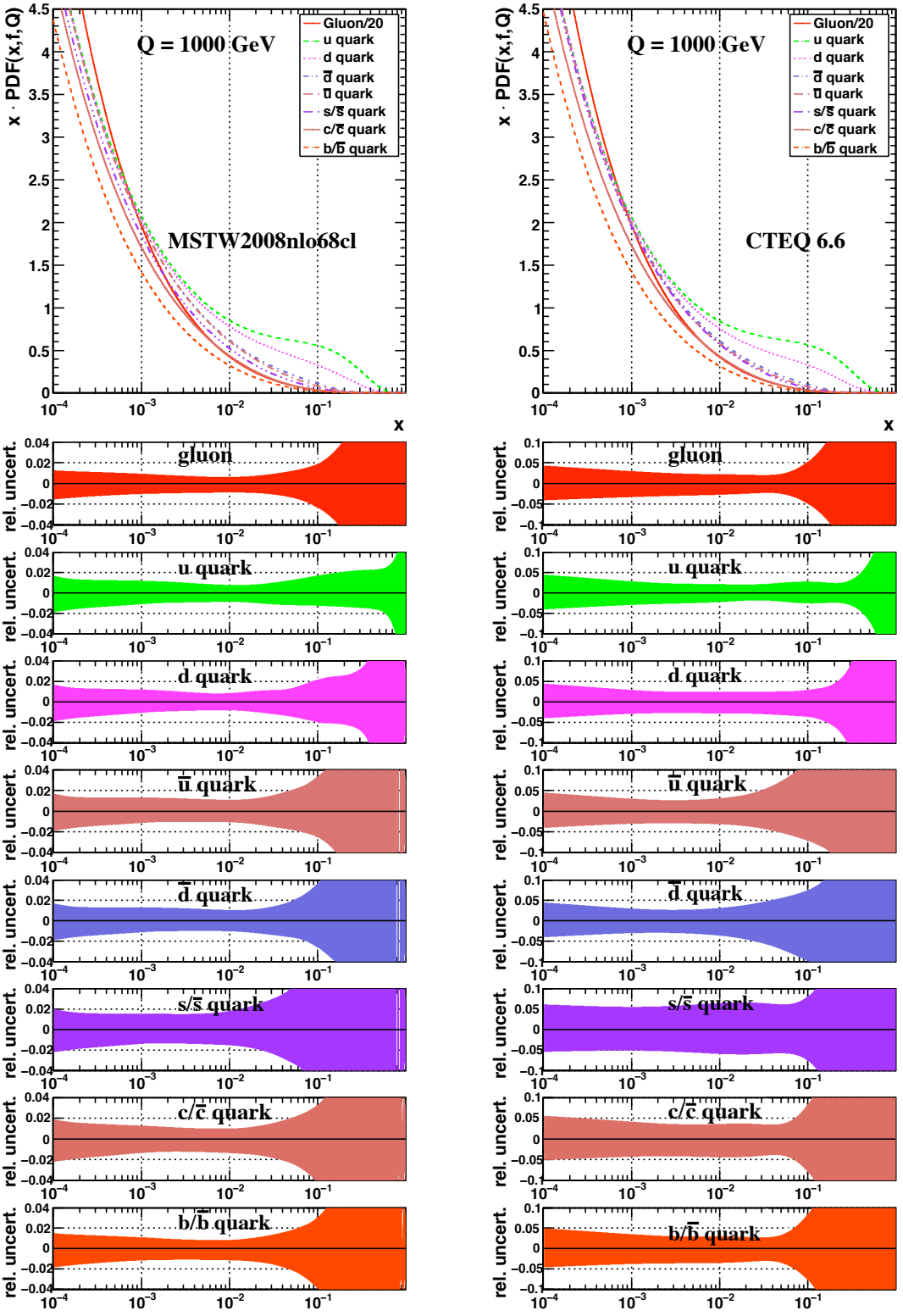

Figure C.2: *Parton distribution functions and their relative uncertainty at the factorization scale $Q = 1$ TeV obtained by the MSTW (left) and the CTEQ (right) group. The values from the error varied PDFs have been fed into the master formula to calculate the relative uncertainties. Note the differences in the y-axis range for the relative uncertainties.*

run i.e. once with the best-fit PDF and $2n$ times using the error varied PDFs for the n parameter used within the PDF global fits. Using the M different Monte Carlo samples one determines for each of them separately the variable X of interest. In the simplest case this could be the M cross sections directly given by running the event generator with the M different error PDFs. The uncertainty on the variable $X^{+\Delta X^{+}_{max}}_{-\Delta X^{-}_{max}}$ induced by the PDF uncertainty is then given by a so called **master formula** [57]:

$$\begin{aligned} \Delta X^{+}_{max} &= \sqrt{\sum_{i=1}^{N} \left[max\left(X^{+}_{i} - X_0, X^{-}_{i} - X_0, 0\right)\right]^2} \\ \Delta X^{-}_{max} &= \sqrt{\sum_{i=1}^{N} \left[max\left(X_0 - X^{+}_{i}, X_0 - X^{-}_{i}, 0\right)\right]^2} \end{aligned} \tag{C.3}$$

X_0 represents the value of the variable of interest X determined using the best-fit PDF, while $X^{\pm}_{i}$ denominate the up/down varied PDF corresponding to the variation of the i-th parameter[1]. Various other variations of the master formula exist in the literature (see e.g. [171] and references therein), but this formula is recommended by the groups performing the global PDF analysis [57], since it considers independently the maximal positive and negative variation of the observable of interest.

C.2.2 The Reweighting Method

The basic idea of the reweighting method is to factorize out the PDF part of the general cross section formula (C.1). Hence one assumes that, while varying the PDF within its uncertainties, the event itself does not change (no change in the available phase space, no topology change such as the jet multiplicity). Following this idea one can define for each event a set of $2n + 1$ weights, defined by the ratio of the PDF values evaluated for the different error PDFs with respect to the best fit PDF:

$$w^j := \frac{\mathrm{PDF}^j(x_1, f_1, Q) \cdot \mathrm{PDF}^j(x_2, f_2, Q)}{\mathrm{PDF}^0(x_1, f_1, Q) \cdot \mathrm{PDF}^0(x_2, f_2, Q)} \qquad \text{for} \qquad 0 \leq j \leq 2n \tag{C.4}$$

Following the definition w^0 is equal to unity, while all other values vary around this value. At this point one has for each generated event $2n + 1$ weights. In order to determine the uncertainty on an arbitrary variable X of interest, one calculates this variable $2n+1$ times utilizing once all weights w^0, once all weights w^1 and so on. On ends up with different values $X_0 ... X_{2n}$ of the variable of interest. These values can again be fed into the master formula to obtain the uncertainty on $X = X_0$.

[1] For the CTEQ and MRST/MSTW PDFs the error PDFs are ordered as up/down variation corresponding to the first parameter, up/down variation of the second parameter, ...

C.2.3 Discussion

Advantages of the reweighting method:

- Sample needs to be produced only once instead of M times. This is a huge advantage taking into account the large CPU requirements for generation and simulation of a sufficient number of events for studies at the LHC.
- It allows for PDF uncertainty determination after the full detector simulation and the restriction to events which pass the final selection cuts and thus enter the final variable distribution. Since the detector simulation requires order of magnitude more CPU time than the production of events at generator level, the application of the brute force method is not practical or even impossible at this stage. As shown in section C.3 the PDF uncertainty of a certain selected sub-sample might significantly differ from the uncertainty of the whole sample. An obvious example is Drell Yan: while the PDF uncertainty at the Z-pole is only about 2% it is as large as 5% for invariant Z masses above 2 TeV .
- Consisting only of event weights the method allows an easy application to a distribution of a variable. The principle is the same as described above. One creates $2n + 1$ distributions utilizing the best-fit and the error PDFs and finally applies the master formula (C.3) on a per bin basis.
- The PDF reweight method requires as input the values $x_{1/2}, f_{1/2}, Q$ which have been used in the generator for each event. Once these values are stored one can apply the reweight method to all PDFs one is interested in and can easily compare the uncertainty of PDFs from different groups (CTEQ vs MRST/MSTW) without regenerating the MC sample. It's even possible to use PDFs which are not on the market while the MC has been produced.
- Some event generators only provide the usage of a limited amount of the available modern parton distribution functions. In the case of MadGraph only the best-fit PDFs are available which make the application of the brute force method impossible.

Draw backs:

- One needs to take care of having enough statistics, but that is common to both the brute force and the reweighting method.
- The factorization assumed in the reweight method only holds to a certain degree. For example the PDFs are also used within the context of the shower evolution (Sudakov form factors). Thus varying the PDF might also change the topology of the event concerning the jet multiplicity.

Still, as it has been reported previously in [171] and as shown in the result section C.3, the agreement between both methods is remarkably good.

C.3 Results

The reweighting method has been applied to a subsample of the Summer08/Winter09 MC-production. Various event generators such as PYTHIA [172] (Di-Boson, LM1-4 and TTtauola supplemented with Tauola [173] for the τ-decays) and MADGRAPH [174] (W/Z+Jets, tt+Jets) are used. In Table C.1 one can see the PDF uncertainties for different physics channels and PDFs available within the LHAPDF distributed within the CMS software stack.

Dataset	LO cross section [pb]	CTEQ NLO				MSTW 2008		
		6	6.1	6.5	6.6	LO	NLO	NNLO
WJets	40000	$^{+3.5\%}_{-4.1\%}$	$^{+4.1\%}_{-5.0\%}$	$^{+3.4\%}_{-3.4\%}$	$^{+3.2\%}_{-3.2\%}$	$^{+1.0\%}_{-1.4\%}$	$^{+1.8\%}_{-1.4\%}$	$^{+1.5\%}_{-1.4\%}$
Z+Jets	3700	$^{+3.4\%}_{-3.9\%}$	$^{+3.8\%}_{-4.7\%}$	$^{+3.1\%}_{-3.3\%}$	$^{+2.9\%}_{-3.1\%}$	$^{+0.9\%}_{-1.4\%}$	$^{+1.8\%}_{-1.3\%}$	$^{+1.4\%}_{-1.3\%}$
TT+Jets	317	$^{+4.7\%}_{-4.4\%}$	$^{+4.9\%}_{-4.8\%}$	$^{+5.0\%}_{-4.3\%}$	$^{+4.7\%}_{-4.5\%}$	$^{+2.1\%}_{-2.2\%}$	$^{+1.9\%}_{-2.3\%}$	$^{+2.0\%}_{-2.0\%}$
TTtauola	241.7	$^{+4.0\%}_{-4.1\%}$	$^{+4.1\%}_{-4.4\%}$	$^{+4.4\%}_{-4.0\%}$	$^{+4.2\%}_{-4.2\%}$	$^{+2.1\%}_{-2.3\%}$	$^{+1.9\%}_{-2.4\%}$	$^{+2.0\%}_{-2.0\%}$
WWincl	44.8	$^{+3.6\%}_{-4.1\%}$	$^{+3.8\%}_{-4.7\%}$	$^{+3.1\%}_{-3.1\%}$	$^{+3.2\%}_{-3.2\%}$	$^{+1.0\%}_{-1.4\%}$	$^{+2.1\%}_{-1.5\%}$	$^{+1.8\%}_{-1.5\%}$
WZincl	19.3	$^{+3.7\%}_{-4.2\%}$	$^{+3.8\%}_{-4.7\%}$	$^{+3.1\%}_{-3.1\%}$	$^{+3.3\%}_{-3.4\%}$	$^{+1.1\%}_{-1.4\%}$	$^{+2.1\%}_{-1.5\%}$	$^{+1.7\%}_{-1.4\%}$
ZZincl	7.1	$^{+3.8\%}_{-4.1\%}$	$^{+3.9\%}_{-4.6\%}$	$^{+3.2\%}_{-3.2\%}$	$^{+3.2\%}_{-3.3\%}$	$^{+1.0\%}_{-1.4\%}$	$^{+2.0\%}_{-1.5\%}$	$^{+2.0\%}_{-1.8\%}$
SUSY LM1	16.1	$^{+7.6\%}_{-5.2\%}$	$^{+7.9\%}_{-6.1\%}$	$^{+6.8\%}_{-4.8\%}$	$^{+6.3\%}_{-5.3\%}$	$^{+1.8\%}_{-1.8\%}$	$^{+2.0\%}_{-1.6\%}$	$^{+1.8\%}_{-1.4\%}$
SUSY LM2	2.4	$^{+10.0\%}_{-6.2\%}$	$^{+10.2\%}_{-7.4\%}$	$^{+8.3\%}_{-5.6\%}$	$^{+7.6\%}_{-6.2\%}$	$^{+2.0\%}_{-1.9\%}$	$^{+2.4\%}_{-1.8\%}$	$^{+2.1\%}_{-1.6\%}$
SUSY LM3	11.8	$^{+8.6\%}_{-5.8\%}$	$^{+8.9\%}_{-6.8\%}$	$^{+7.6\%}_{-5.3\%}$	$^{+7.1\%}_{-5.9\%}$	$^{+1.9\%}_{-1.9\%}$	$^{+2.2\%}_{-1.8\%}$	$^{+1.9\%}_{-1.6\%}$
SUSY LM4	6.7	$^{+9.0\%}_{-5.8\%}$	$^{+9.2\%}_{-6.9\%}$	$^{+7.8\%}_{-5.3\%}$	$^{+7.2\%}_{-5.9\%}$	$^{+1.9\%}_{-1.9\%}$	$^{+2.2\%}_{-1.7\%}$	$^{+1.9\%}_{-1.6\%}$

Table C.1: *PDF Uncertainties for the various CMS datasets. All processes are generated at 10* TeV *centre of mass energy.*

Typically uncertainties of about 2% (3-7%) utilizing MSTW NLO PDFs (CTEQ6.6 NLO PDF) are obtained for (di-)boson or $t\bar{t}$ production as well as for the CMS low mass supersymmetry benchmark points. The comparison of the different CTEQ (NLO) PDFs show a clear evolution of precision with time representing the improving precision of the data used in the global PDF fit.

In general the uncertainties from the CTEQ distributions are roughly a factor of 2 larger in comparison to the MSTW PDFs. These differences are already visible within the PDF distributions in Figures C.1 and C.2 and originate from the differences in the global fitting procedure of the two groups. Although both are using the Hessian method for the construction of the error bands, their definition of how far one should move away from the optimal value is different. While the CTEQ group use a fixed value for the tolerance, the MSTW has adopted a more appropriate dynamical method. As discussed in [57] the MSTW group concludes that the estimated uncertainties of the CTEQ group are too conservative.

The reweighting method can also be applied to distributions as shown in Figure C.3. For each event of the Z+Jet sample which contains at least two muons the invariant mass

is calculated in addition to the 41 event weights from the MSTW NLO PDF. Using these values 41 distributions of the invariant mass are ploted. Finally the master formula C.3 has been applied to each bin.

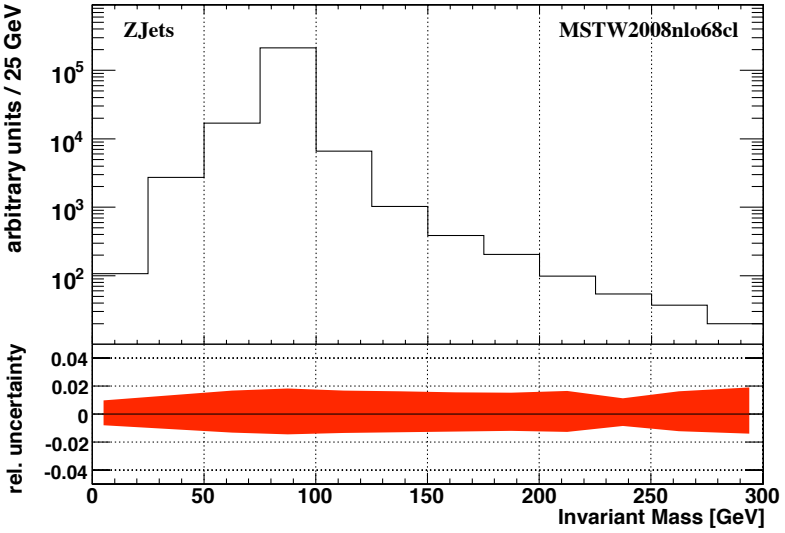

Figure C.3: *Application of the reweighting method to a distribution. The top plot shows the invariant mass distribution of the two leading muons from the Z+Jet sample. The relative uncertainty is obtained by creating 41 invariant mass distributions corresponding to the 41 (sub-)PDFs of the MSTW NLO fit and the subsequent use of the master formula on a bin by bin basis. The uncertainty in the shown range is flat and in agreement with the values of table C.1. The small dip above 200* GeV *reflects the limited statistics in that region.*

C.3.1 Heavy New Particles

The quoted PDF uncertainties for bosons are only valid in the dominant region of production i.e. at the electroweak scale. However, searches for new physics like supersymmetry have to deal with backgrounds like boson production far off the Breit-Wigner pole. Table C.2 shows that the uncertainty for Drell-Yan production increases from 2% (3%) at invariant γ/Z masses above 200 GeV to 6% (10%) for masses above 2 TeV utilizing the MSTW (CTEQ6.6) NLO PDFs. Similar uncertainties are expected for the production of massive particles such as the benchmark W' as a heavy carbon copy of the SM W boson.

The increasing uncertainty as a function of the invariant mass is caused by the reduced knowledge of the parton distribution functions for large momentum fractions ($x > 0.1$). One can see this while comparing the Z+Jets sample, where the main fraction of events are at the Z pole, with the W' of a mass of 1 TeV: While the smallest/largest momentum fraction x for Z+Jets is about 10^{-3} / 0.1 it is as large as 0.05/0.2 for the W' production of a mass of 1 TeV (see Figure C.5). Taking the flavour composition of the two processes into account (Figure C.4) and the distribution of the factorization scale Q (Figure C.6), one can use the PDF uncertainty plots (Figures C.2 and C.7) to qualitatively confirm the results of Table C.2.

Dataset	LO cross section [fb]	CTEQ 6	CTEQ 6.1	CTEQ 6.5	CTEQ 6.6	MSTW 2008 LO	MSTW 2008 NLO	MSTW 2008 NNLO
Z+Jets	3700	$^{+3.4\%}_{-3.9\%}$	$^{+3.8\%}_{-4.7\%}$	$^{+3.1\%}_{-3.3\%}$	$^{+2.9\%}_{-3.1\%}$	$^{+0.9\%}_{-1.4\%}$	$^{+1.8\%}_{-1.3\%}$	$^{+1.4\%}_{-1.3\%}$
W' ($m = 1$ TeV)	4090	$^{+5.5\%}_{-5.7\%}$	$^{+5.3\%}_{-5.8\%}$	$^{+4.9\%}_{-4.4\%}$	$^{+5.1\%}_{-5.4\%}$	$^{+2.2\%}_{-2.1\%}$	$^{+2.5\%}_{-1.9\%}$	$^{+2.4\%}_{-1.6\%}$
W' ($m = 1.5$ TeV)	710	$^{+7.5\%}_{-6.8\%}$	$^{+7.0\%}_{-6.8\%}$	$^{+6.2\%}_{-5.5\%}$	$^{+6.6\%}_{-7.0\%}$	$^{+3.2\%}_{-2.7\%}$	$^{+3.4\%}_{-2.2\%}$	$^{+3.4\%}_{-2.1\%}$
W' ($m = 2$ TeV)	350	$^{+9.7\%}_{-8.0\%}$	$^{+9.0\%}_{-8.1\%}$	$^{+7.3\%}_{-6.4\%}$	$^{+8.3\%}_{-8.4\%}$	$^{+4.2\%}_{-3.3\%}$	$^{+4.5\%}_{-2.9\%}$	$^{+5.1\%}_{-3.2\%}$
DY ($m > 0.2$ TeV)	1620	$^{+3.7\%}_{-4.6\%}$	$^{+3.8\%}_{-5.1\%}$	$^{+2.9\%}_{-3.1\%}$	$^{+3.4\%}_{-3.3\%}$	$^{+1.0\%}_{-1.4\%}$	$^{+1.8\%}_{-1.4\%}$	$^{+1.4\%}_{-1.2\%}$
DY ($m > 0.5$ TeV)	54.4	$^{+4.5\%}_{-5.0\%}$	$^{+4.4\%}_{-5.3\%}$	$^{+3.7\%}_{-3.6\%}$	$^{+4.7\%}_{-4.0\%}$	$^{+1.6\%}_{-1.8\%}$	$^{+2.2\%}_{-1.7\%}$	$^{+1.9\%}_{-1.4\%}$
DY ($m > 1$ TeV)	2.82	$^{+6.6\%}_{-6.6\%}$	$^{+6.1\%}_{-6.8\%}$	$^{+5.7\%}_{-5.1\%}$	$^{+7.1\%}_{-5.4\%}$	$^{+2.5\%}_{-2.4\%}$	$^{+3.0\%}_{-2.1\%}$	$^{+2.8\%}_{-1.8\%}$
DY ($m > 2$TeV)	0.06	$^{+9.8\%}_{-12.5\%}$	$^{+12.1\%}_{-10.5\%}$	$^{+8.5\%}_{-7.7\%}$	$^{+10.9\%}_{-8.9\%}$	$^{+5.0\%}_{-4.3\%}$	$^{+5.6\%}_{-3.4\%}$	$^{+5.7\%}_{-3.4\%}$

Table C.2: *PDF uncertainties for the various CMS datasets. All processes are generated at 10* TeV *centre of mass energy.*

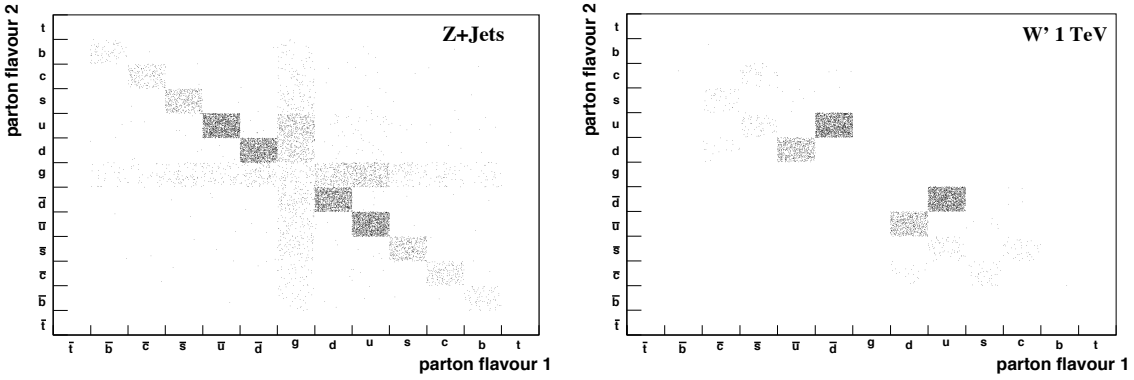

Figure C.4: *Flavour distribution of the Z+Jet (MadGraph) and the W' (1* TeV , PYTHIA*) sample. While* PYTHIA *is only able to generate $2 \to 2$ processes, the LO matrix element generator MadGraph is able to simulate $2 \to n$ ($n \leq 6$) diagrams. Therefore* PYTHIA *only allows for parton-parton combinations which can directly couple to a W' such as $u\bar{d}$ (dominant) and $\bar{u}d$ (predominant).* MADGRAPH *is also able to e.g. start with a gluon which splits into a quark-antiquark pair from which one parton interacts with a parton of the other proton to form a γ/Z, while the other parton is emitted as jet.*

C.3.2 Comparison of the Brute Force and the Reweighting Method

Table C.3 shows the comparison of the reweighting method with the brute force method for the Pythia channels. At least 100k events have been produced to obtain a reasonably precise cross section not limited by statistical uncertainties. The samples have been produced with the different error PDFs and the corresponding cross sections have been fed into the master formula for the uncertainty calculation. One can see that both methods agree well, in agreement with a previous analysis [171].

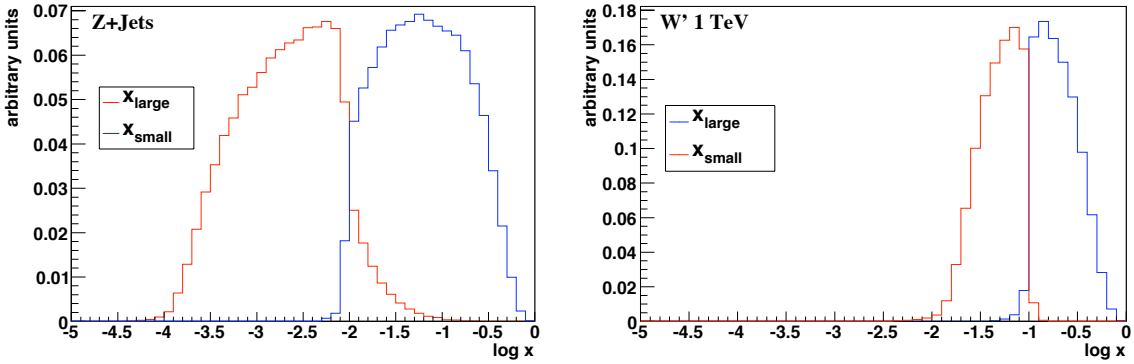

Figure C.5: *Momentum fraction distributions, separately for the parton with smallest/largest momentum fraction, for the Z+Jet (MadGraph) and for the W' (1* TeV, PYTHIA*) sample. As the Z is much lighter than the W', smaller fractions of the protons' momenta can be used for their production. Combining the flavour contribution distributions (Figure C.4) with the momentum distribution, the scale distribution (Figure C.6) and the uncertainties on the PDFs (Figures C.2 and C.7) one can qualitatively see that the expected uncertainty of the W' exceeds the one on Z+Jets.*

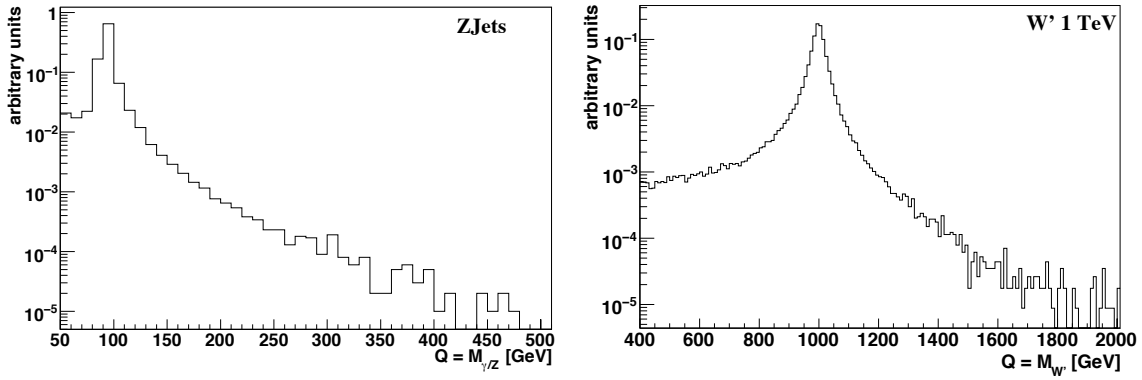

Figure C.6: *Distribution of the factorization scale Q for the Z+Jet and the W' sample. Both generators (*MADGRAPH/PYTHIA*) use the invariant mass of the resonantly produced boson as the choice of the scale leading to Breit-Wigner distributions around 90* GeV / *1* TeV .

C.4 Conclusion

Two methods for the calculation of the uncertainty induced by the imperfect knowledge of the parton distribution functions within the CMS context have been presented. While the brute force method is more profound it is impractical and in some cases even impossible to use due to the limited amount of computing resources. These drawbacks are circumvented by the reweighting method which estimates the PDF uncertainty using event weights. This allows the seamless integration of the PDF uncertainty determination into the analysis workflow. It allows to calculate the uncertainty restricted to only those events entering the final variable and the method is also applicable to distributions of variables. Both

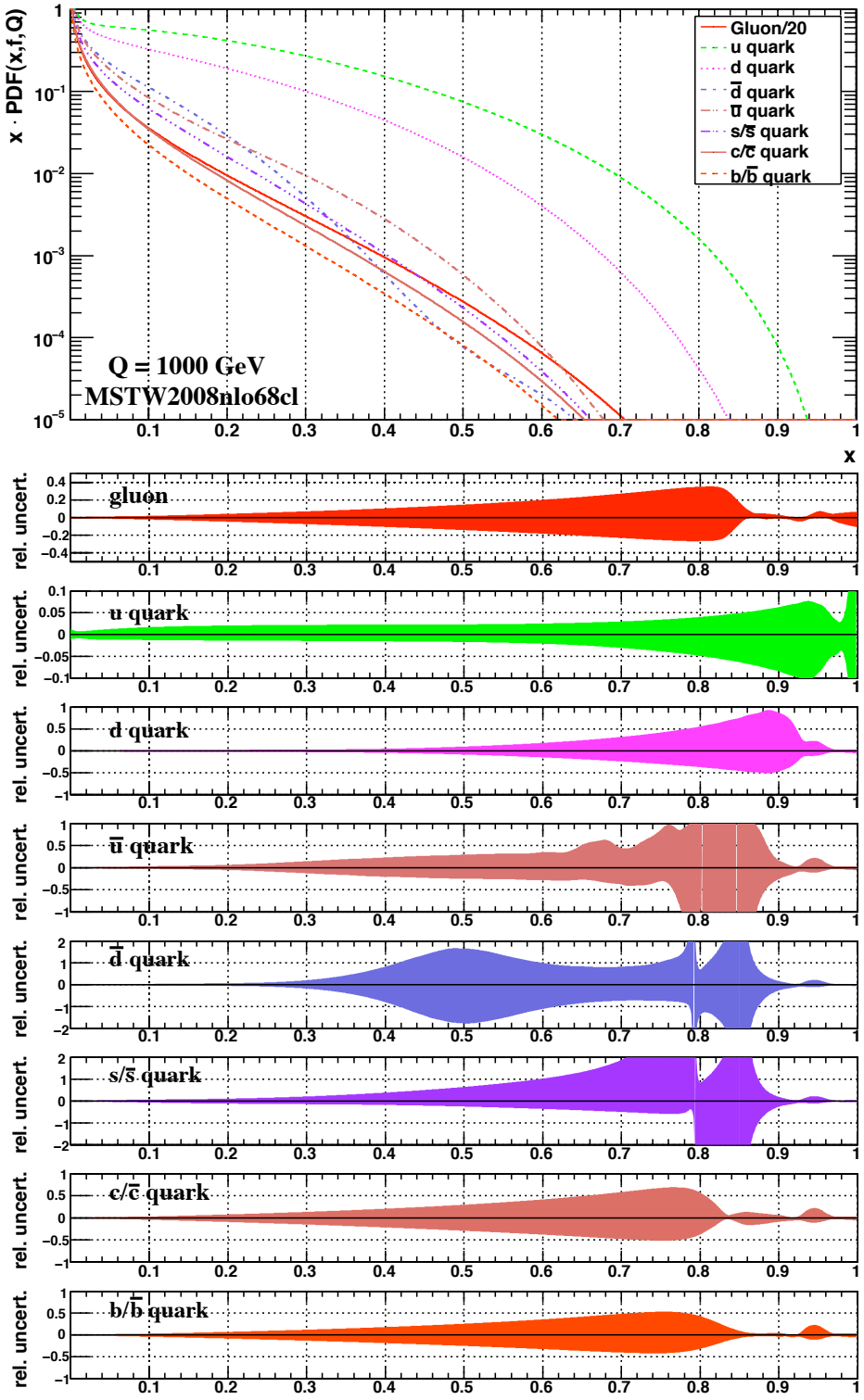

Figure C.7: *Parton distribution functions and their relative uncertainty at the factorization scale $Q = 1$ TeV for large momentum fractions x for the MSTW NLO PDF. The values from the error varied PDFs have been fed into the master formula to calculate the relative uncertainties.*

Dataset	Brute Force		Reweighting	
	CTEQ6.6	MSTW NLO	CTEQ6.6	MSTW NLO
TTtauola	$^{+5.0\%}_{-4.5\%}$	$^{+2.0\%}_{-2.5\%}$	$^{+4.2\%}_{-4.2\%}$	$^{+1.9\%}_{-2.4\%}$
WW	$^{+3.1\%}_{-3.1\%}$	$^{+2.1\%}_{-1.5\%}$	$^{+3.2\%}_{-3.2\%}$	$^{+2.1\%}_{-1.5\%}$
WZ	$^{+3.2\%}_{-3.3\%}$	$^{+2.1\%}_{-1.5\%}$	$^{+3.3\%}_{-3.4\%}$	$^{+2.1\%}_{-1.5\%}$
ZZ	$^{+3.1\%}_{-3.2\%}$	$^{+2.0\%}_{-1.5\%}$	$^{+3.2\%}_{-3.3\%}$	$^{+2.0\%}_{-1.5\%}$
LM1	$^{+8.5\%}_{-6.4\%}$	$^{+2.7\%}_{-2.0\%}$	$^{+6.3\%}_{-5.3\%}$	$^{+2.0\%}_{-1.6\%}$
LM2	$^{+9.5\%}_{-8.3\%}$	$^{+2.7\%}_{-2.4\%}$	$^{+7.6\%}_{-6.2\%}$	$^{+2.4\%}_{-1.8\%}$
LM3	$^{+9.2\%}_{-7.5\%}$	$^{+2.6\%}_{-2.4\%}$	$^{+7.1\%}_{-5.9\%}$	$^{+2.2\%}_{-1.8\%}$
LM4	$^{+9.3\%}_{-7.8\%}$	$^{+2.7\%}_{-2.3\%}$	$^{+7.2\%}_{-5.3\%}$	$^{+2.2\%}_{-1.7\%}$

Table C.3: *Comparison of the PDF uncertainty of the* PYTHIA *data sets determined by the brute force and by the reweighting method using the latest NLO PDFs of the CTEQ and MSTW group. Both methods show a good agreement. While the values for the $t\bar{t}$ and di-boson production agree perfectly the brute force method results in slightly larger uncertainties for the supersymmetry benchmark points. For the* MADGRAPH *samples the reference values from the brute force method cannot be calculated due to the lack of error PDFs within* MADGRAPH.

methods have been applied to the current CMS Monte Carlo production and are in good agreement.

Bibliography

[1] DESY. TESLA – An International, Interdisciplinary Center for Research, 2001. `http://tesla.desy.de/new_pages/TDR_CD/brochure/`.

[2] F. Halzen and A. D. Martin. *Quarks and Leptons: An introductory Course in Modern Particle Physics*. John Wiley & Sons, February 1984.

[3] M. Peskin and D. V. Schroeder. *An Introduction to Quantum Field Theory*. Westview Press, 1995.

[4] C. Berger. *Elementarteilchenphysik*. Springer, Juli 2001.

[5] P. Schmüser. *Feynman-Graphen und Eichtheorien für Experimentalphysiker*. Springer, August 1994.

[6] R. Mohapatra. *Unification and Supersymmetry*. Springer Verlag, second edition, 1996.

[7] R. Brandelik et al. Evidence for a Spin-1 Gluon in Three-Jet Events. *Physics Letters B*, 97(3-4):453 – 458, 1980.

[8] S. L. Glashow. Partial Symmetries of Weak Interactions. *Nucl. Phys.*, 22:579–588, 1961.

[9] A. Salam and J. C. Ward. Electromagnetic and Weak Interactions. *Phys. Lett.*, 13:168–171, 1964.

[10] S. Weinberg. A Model of Leptons. *Phys. Rev. Lett.*, 19:1264–1266, 1967.

[11] C. S. Wu, E. Ambler, R. W. Hayward, D. D. Hoppes, and R. P. Hudson. Experimental Test of Parity Conservation in Beta Decay. *Phys. Rev.*, 105:1413–1414, 1957.

[12] A. Djouadi, J. Kalinowski, and M. Spira. HDECAY: a Program for Higgs Boson Decays in the Standard Model and its Supersymmetric Extension. *Computer Physics Communications*, 108:56, 1998.

[13] G. Abbiendi et al. Search for the Standard Model Higgs Boson at LEP. *Physics Letters B*, 565:61, 2003.

[14] The Super-Kamiokande Collaboration. Evidence for an Oscillatory Signature in Atmospheric Neutrino Oscillation. *Physical Review Letters*, 93:101801, 2004.

[15] W. Grimus. Introduction to Left-Right Symmetric Models. *Invited talk at Lectures given at 4th Hellenic School on Elementary Particle Physics, Corfu, Greece*, 2-20 Sep, 1992.

[16] K. G. Begeman, A. H. Broeils, and R. H. Sanders. Extended Rotation Curves of Spiral Galaxies - Dark Haloes and Modified Dynamics. *Royal Astronomical Society*, 249:523–537, April 1991.

[17] G. Hinshaw et al. Five-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Data Processing, Sky Maps, and Basic Results. *The Astrophysical Journal*, 180:225, 2009.

[18] R. A. Knop et al. New Constraints on Ω_M, Ω_Λ, and w from an Independent Set of Eleven High-Redshift Supernovae Observed with HST. *ASTROPHYS. J*, 102, 2003.

[19] A. G. Riess et al. Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant. *The Astronomical Journal*, 116:1009–1038, September 1998.

[20] S. P. Martin. A Supersymmetry Primer, 1997. `http://www.citebase.org/abstract?id=oai:arXiv.org:hep-ph/9709356`.

[21] J. C. Pati and A. Salam. Lepton Number as the fourth Color. *Phys. Rev.*, D10:275–289, 1974.

[22] R. N. Mohapatra and J. C. Pati. Left-Right Gauge Symmetry and an 'Isoconjugate' Model of CP Violation. *Phys. Rev.*, D11:566–571, 1975.

[23] R. N. Mohapatra and G. Senjanovic. Neutrino Mass and Spontaneous Parity Nonconservation. *Phys. Rev. Lett.*, 44:912, 1980.

[24] M. Gell-Mann, P. Ramond, and R. Slansky. *Supergravity*. P. van Nieuwenhuizen and D. Z. Freedman, 1979.

[25] T. Yanagida. Proceedings of the Workshop on the Baryon Number of the Universe and Unified Theories, 1979.

[26] Z. Sullivan. Fully differential W' Production and Decay at next-to-leading Order in QCD. *Physical Review D*, 66:075011, 2002.

[27] T. Han, H. E. Logan, B. McElrath, and L.-T. Wang. Phenomenology of the Little Higgs Model. *Phys. Rev.*, D67:095004, 2003.

[28] Z. Sullivan. How to Rule Out Little Higgs (and Constrain Many Other Models) at the LHC. In *Proceedings of XXXVIIIth Rencontres de Moriond: QCD*, page 379. The Gioi Publishers, March 2003.

[29] G. Altarelli, B. Mele, and M. Ruiz-Altaba. Searching for New Heavy Vector Bosons in p anti-p Colliders. *Z. Phys.*, C45:109, 1989.

[30] DØ Collaboration. Search for W' Bosons Decaying to an Electron and a Neutrino with the DØ Detector. *Physical Review Letters*, 100:031804, 2008.

[31] C. Hof. Detection of New Heavy Gauge Bosons W' in CMS. Technical Report CMS-CR-2006-054, CERN, Geneva, Sep 2006.

[32] C. Hof, T. Hebbeker, and K. Hoepfner. Detection of New Heavy Charged Gauge Bosons in the Muon plus Neutrino Channel. Technical Report CMS-NOTE-2006-117, CERN, Geneva, Apr 2006.

[33] The CMS Collaboration. Search for $W' \rightarrow e\nu$. *CMS Physics Analysis Summary*, EXO 2008/004.

[34] CMS Collaboration. *CMS Physics – Technical Design Report*, volume II. CERN/LHCC 2006, 2005.

[35] N. Arkani-Hamed, S. Dimopoulos, and G. Dvali. Phenomenology, Astrophysics and Cosmology of Theories with Sub-Millimeter Dimensions and TeV Scale Quantum Gravity. *Physical Review D*, 59:086004, 1999.

[36] I. Antoniadis, N. Arkani-Hamed, S. Dimopoulos, and G. Dvali. New Dimensions at a Millimeter to a Fermi and Superstrings at a TeV. *Physics Letters B*, 436:257, 1998.

[37] N. Arkani-Hamed, S. Dimopoulos, and G. Dvali. The Hierarchy Problem and New Dimensions at a Millimeter. *Physics Letters B*, 429:263, 1998.

[38] L. Randall and R. Sundrum. An Alternative to Compactification. *Physical Review Letters*, 83:4690, 1999.

[39] L. Randall and R. Sundrum. A Large Mass Hierarchy from a Small Extra Dimension. *Physical Review Letters*, 83:3370, 1999.

[40] S. Hossenfelder. What Black Holes Can Teach Us, 2005. `http://www.citebase.org/abstract?id=oai:arXiv.org:hep-ph/0412265`.

[41] T. Appelquist and H.-U. Yee. Universal Extra Dimensions and the Higgs Boson Mass. *Phys. Rev. D*, 67(5):055002, Mar 2003.

[42] D. J. Kapner, T. S. Cook, E. G. Adelberger, J. H. Gundlach, B. R. Heckel, C. D. Hoyle, and H. E. Swanson. Tests of the Gravitational Inverse-Square Law below the Dark-Energy Length Scale. *Physical Review Letters*, 98(2):021101, 2007.

[43] G. Landsberg. Collider Searches for Extra Dimensions. *ECONFC040802*, MOT006, 2004.

[44] D.-C. Dai, G. Starkman, D. Stojkovic, C. Issever, E. Rizvi, and J. Tseng. BlackMax: A Black-Hole Event Generator with Rotation, Recoil, Split Branes and Brane Tension. *Physical Review D*, 77:076007, 2008.

[45] S. B. Giddings and S. Thomas. High Energy Colliders as Black Hole Factories: The End of Short Distance Physics. *Phys. Rev. D*, 65(5):056010, Feb 2002.

[46] G. L. Kane, C. Kolda, L. Roszkowski, and J. D. Wells. Study of Constrained Minimal Supersymmetry. *Physical Review D*, 49:6173, 1994.

[47] G. F. Giudice and R. Rattazzi. Theories with Gauge-Mediated Supersymmetry Breaking. *Physics Reports*, 322:419, 1999.

[48] V. M. Abazov et al. Search for Supersymmetry in Di-Photon Final States at $\sqrt{s} =$ 1.96 TeV, 2007.
`http://www.citebase.org/abstract?id=oai:arXiv.org:0710.3946`.

[49] V. M. Abazov et al. Search for Squarks and Gluinos in Events with Jets and Missing Transverse Energy using 2.1 fb^{-1} of $p\bar{p}$ Collision Data at $\sqrt{s} = 1.96$ TeV. *Phys. Lett.*, B660:449–457, 2008.

[50] L. Evans (ed.) and P. Bryant (ed.). LHC Machine. *JINST*, 3:S08001, 2008.

[51] W. J. Stirling. Private communications, 2009.

[52] O. Bruning (eg.) et al. *LHC Design Report. Vol. I: The LHC Main Ring.* CERN-2004-003, 2004.

[53] CERN. CERN Photo Data Base. `http://cdsweb.cern.ch`.

[54] Interim Summary Report on the Analysis of the 19 September 2008 Incident at the LHC. https://edms.cern.ch/document/973073/1.

[55] L. Evans. LHC Status Report. 95th LHCC Meeting.

[56] R. Aymar. Status of CERN Activities. Talk given at the European Commitee for Future Accelerators Meeting, November 2008.

[57] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt. Parton Distributions for the LHC, 2009.

[58] M. Pimia et al. Compact Muon Solenoid. In G. Jarlskog and D. Rain, editors, *Proc. Large Hadron Collider Workshop, Aachen*, volume III, page 547. CERN 90-10, October 1990.

[59] CMS Collaboration. *The Tracker System Project – Technical Design Report.* CERN/LHCC 94-38, December 1994.

[60] CMS Collaboration. *The Electromagnetic Calorimeter Project – Technical Design Report.* CERN/LHCC 97-33, December 1997.

[61] CMS Collaboration. *The Hadron Calorimeter Project – Technical Design Report.* CERN/LHCC 97-31, June 1997.

[62] CMS Collaboration. *The Magnet Project – Technical Design Report.* CERN/LHCC 97-10, May 1997.

[63] CMS Collaboration. *The Muon Project – Technical Design Report.* CERN/LHCC 97-32, December 1997.

[64] CMS Collaboration. *The Level-1 Trigger – Technical Design Report*, volume I. CERN/LHCC 2000-038, December 2000.

[65] CMS Collaboration. *The Trigger and Data Acquisition Project – Technical Design Report*, volume II. CERN/LHCC 2002-026, December 2002.

[66] R. Adolphi et al. The CMS Experiment at the CERN LHC. *JINST*, 0803:S08004, 2008.

[67] CMS Collaboration. *CMS Physics – Technical Design Report*, volume I. CERN/LHCC 2006, 2005.

[68] CMS Informations. `http://cmsinfo.cern.ch`.

[69] N. Marinelli, T. Kolberg, C. Jessop, and R. Ruchti. Track Finding and Identification of Converted Photons with the CMS Tracker and ECAL. *CMS Analysis Note*, 2008/102.

[70] CMS Collaboration. *Addendum to the CMS Tracker TDR.* CERN/LHCC 2000-016, February 2000.

[71] W. Funk. The Electromagnetic Calorimeter of CMS. *CMS Note*, CMS CR-2004/040, 2002.

[72] D. Froidevaux and P. Sphicas. General-Purpose Detectors for the Large Hadron Collider. *Ann. Rev. Nucl. Part. Sci.*, 56:375–440, 2006.

[73] The TOTEM Collaboration. *TOTEM - Total Cross Section, Elastic Scattering and Diffraction Dissociation at the LHC.* CERN/LHCC 97-49, August 1997.

[74] Iguanacms. `http://iguanacms.web.cern.ch/iguanacms/`.

[75] M. Bontenackels and H. Reithler, 2007. Private Communications.

[76] C. Albajar et al. Test Beam Analysis of the First CMS Drift Tube Muon Chamber. *Nucl. Instrum. Meth.*, A525:465–484, 2004.

[77] G. Bruno. Resistive Plate Chambers in Running and Future Experiments. *Eur. Phys. J.*, C33:1032–1034, 2004.

[78] CMS Collaboration. *CMS High Level Trigger.* CERN/LHCC 2007-021, 2007.

[79] CERN Document Server. Beam Splash in CMS. `http://cdsweb.cern.ch/record/1125999`.

[80] I. Foster. What is the Grid? – A Three Point Checklist. *GRIDtoday*, 1(6), July 2002.

[81] C. Kesselman and I. Foster. *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann Publishers, November 1998.

[82] I. Foster, C. Kesselman, and S. Tuecke. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. *International Journal of Supercomputer Applications*, 15:2001, 2001.

[83] A. Gellrich. The Grid – An Introduction from a Personal Perspective, GridKa School 2008. `http://gks08.fzk.de`.

[84] The World Wide LHC Computing Grid. `http://www.cern.ch/lcg`.

[85] E. Laure and B. Jones. Enabling Grids for e-Science: The EGEE Project. Technical Report EGEE-PUB-2009-001. 1, CERN, Sep 2008.

[86] The Open Science Grid. `http://www.opensciencegrid.org/`.

[87] P. J. W. Faulkner et al. GridPP: Development of the UK Computing Grid for Particle Physics. *J. Phys.*, G32:N1–N20, 2006.

[88] The INFN Grid Project. `http://grid.infn.it/`.

[89] The NorduGrid. `http://www.nordugrid.org/`.

[90] CASTOR - CERNs Advanced Storage Manager. `http://castor.web.cern.ch/castor/`.

[91] dCache. `http://www.dcache.org/`.

[92] Disk Pool Manager. `https://twiki.cern.ch/twiki/bin/view/LCG/DpmGeneralDescription`.

[93] Berkeley Storage Manager. `http://datagrid.lbl.gov/bestman/`.

[94] Storage Resource Manager. `http://storm.forge.cnaf.infn.it/`.

[95] gLite - Lightweight Middleware for Grid Computing. http://glite.web.cern.ch/glite/.

[96] F. Würthwein. Science on the Grid with CMS at the LHC. *J. Phys. Conf. Ser.*, 125:012073, 2008.

[97] A. Duarte, P. Nyczyk, A. Retico, and D. Vicinanza. Monitoring the EGEE/WLCG Grid Services. *J. Phys. Conf. Ser.*, 119:052014, 2008.

[98] GGUS - The Grid Global User Support. `http://www.ggus.org`.

[99] The CMS Collaboration. *The Computing Project – Technical Design Report.* CERN/LHCC 2005-023, June 2005.

[100] A. Anzar et al. The CMS Dataset Bookkeeping Service. *J. Phys. Conf. Ser.*, 119:072001, 2008. `https://cmsweb.cern.ch/dbs_discovery/`.

[101] PhEDEx – CMS Data Transfers. `http://cmsweb.cern.ch/phedex/`.

[102] J. Rehn et al. PhEDEx High-Throughput Data Transfer Management System. In *Proc. Int. Conf. on Computing in High Energy and Nuclear Physics (Mumbai)*, volume 2, page 1027. Macmillan India ltd., 2006.

[103] FTS - The gLite File Transfer System. http://cern.ch/egee-jra1-dm/FTS.

[104] S. Kosyakov et al. Frontier: High Performance Database Access using Standard Web Components in a Scalable Multi-Tier Architecture. Presented at Computing in High-Energy Physics (CHEP '04), Interlaken, Switzerland, 27 Sep - 1 Oct 2004.

[105] J. M. Hernandez et al. Bringing the CMS Distributed Computing System into Scalable Operations. Talk given at CHEP'09 (proceedings to be published).

[106] F. Fanzago et al. CRAB: A Tool to Enable CMS Distributed Analysis. In *Proc. Int. Conf. on Computing in High Energy and Nuclear Physics (Mumbai)*, volume 2, page 1110. Macmillan India ltd., 2006.

[107] J. Andreeva et al. Dashboard for the LHC Experiments. *J. Phys. Conf. Ser.*, 119:062008, 2008.

[108] MonALISA - Monitoring Agents using a Large Integrated Services Architecture. http://monalisa.caltech.edu/.

[109] The CMS Job Robot. http://jobrobot.web.cern.ch/JobRobot/.

[110] The CMS Site Status Board. http://dashb-ssb.cern.ch/dashboard/request.py/siteviewhome.

[111] The Happy Face Project. https://ekptrac.physik.uni-karlsruhe.de/HappyFace/wiki.

[112] The Quattor System Administration Toolsuite. `http://www.quattor.org`.

[113] LEMON - LHC Era Monitoring. `http://cern.ch/lemon`.

[114] The CMS Collaboration. CMS Times, October 2006.

[115] P. Biallass, T. Hebbeker, and K. Hoepfner. First Measurements of Cosmic Muons with Magnetic Field in CMS. *J. Phys. Conf. Ser.*, 110:122004, 2008.

[116] P. Biallass, T. Hebbeker, and K. Hoepfner. Simulation of Cosmic Muons and Comparison with Data from the Cosmic Challenge using Drift Tube Chambers. *CMS Note*, 024, 2007.

[117] The National Analysis Facility. `http://naf.desy.de/`.

[118] Helmholtz Alliance - Physics at the Terascale. `http://www.terascale.de/`.

[119] Lustre a Network Clustering Filesystem. http://www.lustre.org/.

[120] B. Bellenot et al. PROOF - The Parallel ROOT Facility. In *HPDC*, pages 379–380, 2006.

[121] P. Biallass. *Commissioning of the CMS Muon Detector and Development of Generic Search Strategies for New Physics*. PhD thesis, RWTH Aachen University, 2009.

[122] The DØ Collaboration. Search for New Physics in e mu X Data at DØ Using Sleuth: A Quasi-Model-Independent Search Strategy for New Physics. *Physical Review D*, 62:092004, 2000.

[123] B. Abbott et al. A Quasi-Model-Independent Search for New Physics at Large Transverse Momentum. *Phys. Rev.*, D64:012004, 2001.

[124] B. Abbott et al. Quasi-Model-Independent Search for New High p_T Physics at DØ. *Phys. Rev. Lett.*, 86(17):3712–3717, Apr 2001.

[125] A. Aktas et al. A General Search for New Phenomena in e p Scattering at HERA. *Phys. Lett.*, B602:14–30, 2004.

[126] H1 Collaboration. A General Search for New Phenomena at HERA, 2009. http://www.citebase.org/abstract?id=oai:arXiv.org:0901.0507.

[127] The CDF Collaboration. Model-Independent and Quasi-Model-Independent Search for New Physics at CDF. *Physical Review D*, 78:012002, 2008.

[128] T. Hebbeker. A Global Comparison between L3 Data and Standard Model Monte Carlo – a First Attempt. http://web.physik.rwth-aachen.de/∼hebbeker/l3note_2305.pdf, 1998.

[129] P. Biallass. Model Independent Search for Deviations from the Standard Model at the Tevatron: Final States with Missing Energy, Diploma Thesis, RWTH Aachen University, 2004.

[130] F. Fabozzi, C. D. Jones, B. Hegner, and L. Lista. Physics analysis tools for the cms experiment at lhc. *Nuclear Science, IEEE Transactions on*, 55(6):3539–3543, Dec. 2008.

[131] C. Jones. The New CMS Data Model and Framework. In *CHEP'06 Conference Proceedings*, 2007.

[132] M. Erdmann, G. Müller, and J. Steggemann. Physics eXtension Library 2.0. http://pxl.sourceforge.net.

[133] R. Brun and F. Rademakers. ROOT: An Object Oriented Data Analysis Framework. *Nucl. Instrum. Meth.*, A389:81–86, 1997.

[134] The CMS Processing Model. https://twiki.cern.ch/twiki/bin/view/CMS/WorkBookCMSSWFramework.

[135] S. Agostinelli et al. G4 – A Simulation Toolkit. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 506(3):250 – 303, 2003.

[136] The CMS Collaboration. Towards a Measurement of the Inclusive $W \to \mu\nu$ and $Z \to \mu^+\mu^-$ Cross Sections in pp Collisions at $\sqrt{s} = 14$ TeV . *CMS Physics Analysis Summary*, EWK 2007/002.

[137] The CMS Collaboration. Measuring Electron Efficiencies with Early Data. *CMS Physics Analysis Summary*, EWK 2007/001.

[138] W. Andrews et al. Data-Driven Methods to Estimate the Electron and Muon Fake Contributions to Lepton Analyses. *CMS Analysis Note*, 2009/041.

[139] G. Abbiendi et al. Muon Reconstruction in the CMS Detector. *CMS Analysis Note*, 2008/097.

[140] R. Frühwirth. Application of Kalman Filtering to Track and Vertex Fitting. *Nucl. Instrum. Meth.*, A262:444–450, 1987.

[141] M. Mulders. Muon Identification in CMS. *CMS Analysis Note*, 2008/098.

[142] I. Altsybeev et al. Search for New High-Mass Resonances Decaying to Muon Pairs in the CMS Experiment. *CMS Analysis Note*, 2007/038.

[143] S. Bafoni et al. Electron Reconstruction in CMS. *CMS Note*, 2006/40.

[144] R. Frühwirth. Track Fitting with Non-Gaussian Noise. *Computer Physics Communications*, 100:1–16, 1997.

[145] J. Branson et al. A Cut Based Method for Electron Identification in CMS. *CMS Analysis Note*, 2008/082.

[146] J. Nysten. Photon Reconstruction in CMS. *Nuclear Instruments and Methods in Physics Research A*, 534(1-2):194 – 198, 2004. Proceedings of the IXth International Workshop on Advanced Computing and Analysis Techniques in Physics Research.

[147] N. Marinelli. Track Finding in Gamma Conversions in CMS, 2007. `http://www.citebase.org/abstract?id=oai:arXiv.org:0710.2818`.

[148] G. P. Salam and G. Soyez. A Practical Seedless Infrared-Safe Cone Jet Algorithm. *JHEP0705*, 086, 2007.

[149] The CMS Collaboration. Performance of Jet Algorithms in CMS. *CMS Physics Analysis Summary*, JME 2007/003.

[150] The CMS Collaboration. Plans for Jet Energy Corrections at CMS. *CMS Physics Analysis Summary*, JME 2007/002.

[151] The CMS Collaboration. $\not\!\!E_{\mathrm{T}}$ performance in CMS. *CMS Physics Analysis Summary*, JME 2007/001.

[152] R. D. Cousins. Why isn't every Physicist a Bayesian? *American Journal of Physics*, vol. 63(5):398–410, 1995.

[153] R. D. Cousins, J. T. Linnemann, and J. Tucker. Evaluation of Three Methods for Calculating Statistical Significance when Incorporating a Systematic Uncertainty into a Test of the Background-Only Hypothesis for a Poisson Process. *ArXiv Physics e-prints*, February 2007.

[154] Alexander L. Read. Presentation of Search Results: The CL(s) Technique. *J. Phys.*, G28:2693–2704, 2002.

[155] S. Schmitz. Model Unspecific Search for New Physics with High pT Photons in CMS. Master's thesis, RWTH Aachen University, 2009 (thesis in preparation).

[156] S. D. Biller. Hypothesis Ranking and the Context of Probabilities in an Open-Ended Search. *Astroparticle Physics*, 4:285–291, February 1996.

[157] C. Hof et al. Parton Distribution Uncertainty Determination within CMSSW. *CMS Analysis Note*, 2009/048.

[158] M. R. Whalley, D. Bourilkov, and R. C. Group. The Les Houches Accord PDFs (LHAPDF) and Lhaglue, 2005.

[159] The CMS Collaboration. Plans for Jet Energy Corrections at CMS. *CMS Physics Analysis Summary*, JME 2007/002.

[160] E. Barberis et al. Trigger and Reconstruction Studies with Beam Halo and Cosmic Muons. *CMS Note*, 012, 2006.

[161] The CMS Collaboration. MUSiC – An Automated Scan for Deviations between Data and Monte Carlo Simulation. *CMS Physics Analysis Summary*, EXO 2008/005.

[162] The CMS Collaboration. Identification and Treatment of Anomalous Signals in the CMS Hadronic Calorimeter. *CMS CRAFT Paper*, CFT 2009/019.

[163] The CMS Collaboration. Calorimeter Jet Quality for the First CMS Collision Data. *CMS Physics Analysis Summary*, JME 2009/008.

[164] The methods are discussed in numerous DØ and CDF publications. A recent example is: D0 Collaboration. Evidence for Production of Single Top Quarks. *Physical Review D*, 78:012005, 2008.

[165] The CMS Collaboration. Search Strategy for a Standard Model Higgs Boson Decaying to Two W Bosons in the Fully Leptonic Final State. *CMS Physics Analysis Summary*, HIG 2008/006.

[166] P. Biallass et al. Search Strategies for mSUGRA in the Muons+Jets+MET Final State. *CMS Analysis Note*, 2008/034.

[167] J. Pumplin, D. R. Stump, J. Huston, H. L. Lai, P. Nadolsky, and W. K. Tung. New Generation of Parton Distributions with Uncertainties from Global QCD Analysis. *JHEP0207*, 012, 2002.

[168] D. Stump, J. Huston, J. Pumplin, W. Tung, H. L. Lai, S. Kuhlmann, and J. F. Owens. Inclusive Jet Production, Parton Distributions, and the Search for New Physics. *JHEP*, 0310:046, 2003.

[169] W. K. Tung, H. L. Lai, A. Belyaev, J. Pumplin, D. Stump, and C. P. Yuan. Heavy Quark Mass Effects in Deep Inelastic Scattering and Global QCD Analysistering and Global QCD AnalysisAnalysis. *JHEP0702*, 053, 2007.

[170] P. M. Nadolsky, H. L. Lai, Q. H. Cao, J. Huston, J. Pumplin, D. Stump, W. K. Tung, and C. P. Yuan. Implications of CTEQ Global Analysis for Collider Observables. *Physical Review D*, 78:013004, 2008.

[171] D. Bourilkov, R. C. Group, and M. R. Whalley. LHAPDF: PDF Use from the Tevatron to the LHC, 2006.

[172] T. Sjostrand, S. Mrenna, and P. Skands. PYTHIA 6.4 Physics and Manual. *JHEP*, 05:026, 2006.

[173] Z. Wäs and P. Golonka. TAUOLA as Tau Monte Carlo for Future Applications. *Nuclear Physics B - Proceedings Supplements*, 144:88 – 94, 2005. TAU 2004.

[174] F. Maltoni and T. Stelzer. MadEvent: Automatic Event Generation with MadGraph. *JHEP0302*, 027, 2003.

List of Figures

List of Tables

Acknowledgements

The last pages are devoted to all the people, who have supported and accompanied me in my PhD time. The work at hand would not be possible without you!

First of all, very special thanks to Philipp Biallass for the fun time and fruitful collaboration in the whole MUSiC project from its birth three years ago up to its readiness to absorb data. It was a great pleasure to work together and to push each other to the limits. Within the countless discussions, coding sessions and approval marathons we learned a lot from each other and could benefit from the others expertise.

I would like to express my highest regards to Prof. Dr. Thomas Hebbeker, who – as a godfather of model-independent searches – introduced me to this fascinating research topic. I really enjoyed the years in his group and I have been glad for his guidance and advices, for the freedom and the trust and for always asking the right questions. Many thanks to Prof. Dr. Christopher Wiebusch for immediately agreeing to be my second referee.

Thanks to Stefan Schmitz who joined the MUSiC team in the last year with significant contributions especially with the inclusion of photons and outstanding work on the statistics part. A warm welcome also to Holger Pieta. He has the honour, but also the burden to run MUSiC on the first *pp*-collisions. Have fun and ... Good luck! Not to forget Klaas Padeken and Erik Dietz-Laursonn who recently joined the project. Special thanks to Kerstin Hoepfner and Arnd Meyer for their advises, for the physics discussions and for the proof-reading of numerous talks and papers.

Many thanks to the EXOTICA group for the continuous exchange of information and help. Very special thanks to the conveners, Sarah Eno, Albert De Roeck, and Greg Landsberg for the outstanding support in this project and for believing in its physics potential; without this MUSiC would have fallen silent for sure. Thanks to the ARC, Claire Shepherd-Themistocleous, Vivek Sharma and Bob Cousins for carefully reviewing the analysis and helping to shape the message it should deliver. Thanks also for the useful comments and questions from various members of the CMS collaboration during the review process. Many thanks to Arnd Meyer and Thomas Hebbeker for the supervision and continuous help on this project. Thanks to Sascha Caron and Georgios Choudalakis for sharing their experiences on model independent searches from past experiments. Many thanks to Daniel Teyssier and Holger Pieta for the collaboration on the comparison of MUSiC with a dedicated SUSY analysis. Thanks to Patrick Tsang and Greg Landsberg for their advices on Black Hole production. Finally, since MUSiC strongly depends on the input and experiences from various dedicated analyses and object identification groups, many thanks to all the people involved in these countless studies. The work has been supported by BMBF and DFG.

During my computing service I worked together with an almost countless number of people. Forgive me if I miss someone! First of all I'd like to thank the RWTH Grid Team

lead by Thomas Kress and Andreas Nowack. Especially Andreas taught me a lot not only about grid computing, but also about operating systems and bash scripting. Concerning the official CMS MC Production, ProdAgent operation, testing and development I would like to thank the project leaders Alexandra Fanfani, Dave Evans and Peter Elmer for their enthusiasm and support. A big thanks also to the MC production operators Alexander Flossdorf, Ajit Mohapatra, Oliver Gutsche, David Mason, Maarten Thomas, Guillelmo Gomez Ceballos, and Jose Hernandez: I was happy to produce and process gazillion of MC events all over the world with you guys. Further I'd like to express my gratitude to the CRAB and DBS teams for their support in various computing questions and to the PAT team in person of Benedikt Hegner for help in various implementation issues. Thanks also to the NAF user committee, where I was able to follow the birth of the NAF analysis centre. As the MUSiC analysis is built upon the objects and expertise of PXL I would like to thank the developers especially to Steffen Kappler for his tips concerning the design and the implementation of various PXL extensions matching the MUSiC requirements.

Thanks to Michael Bontenackels and Oleg Tsigenov for being my first (from day one of my diploma thesis) and also last (during endless evenings and nights) office mates. Probably the only office with 24/7 opening hours and non-stop support in all situations (customer guidance system has been ordered), the best atmosphere, 9 pm coffee breaks (with the inventor Daniela Käfer), and lots of sweets and unhealthy food.

After more than 4 years at the institute IIIA there are of course countless people I met and which helped me during this time. Many thanks to the members of our local CMS group – Metin Ata, Walter Bender, Michael Bontenackels, Prof. Martin Erdmann, Andreas Hinzmann, Hendrik Jansen, Tatsiana Klimkovich, Peter Kreuzer, Markus Merschmeyer, Paul Papacz, Holger Pieta, Hans Reithler, Stefan Schmitz, Michael Sowa, Daniel Teyssier, Oleg Tsigenov, Lars Sonnenschein and Jan Steggemann – for all the useful discussions on physics and the countless help in technical problems. Also many thanks to our colleagues from IIIB for the nice times skiing in Saas Grund and for the coffee machine.

Not to forget all the little helpers for proof-reading of this and other work: Thomas Hebbeker, Amelia Schultheis, Arnd Meyer, Philipp Biallass and Michael Bontenackels (thanks also for the nice MUSiC logo).

Thanks to all of my friends who filled my free time with everything except physics. Thanks to the Doppel-Kopf and cooking circle (Christin, Martina, Heike, Jan, Dirk, Michael), the climbers, the skiers and the mountaineers (Ulla, Martina, Daniela, Amelia, Fred, Michi, Ulli, Dirk and Jan) and of course the dancers (especially to my long-time "Tanzsportgerät" Martina and the coaches Alex and Martin).

Zu guter Letzt möchte ich mich ganz herzlich bei meinen Eltern und natürlich bei Amelia bedanken. Vielen Dank für die ununterbrochene Unterstützung und den moralischen Rückhalt. Vor allem aber Danke für Eure Liebe und Euer Vertrauen.

Carsten Hof

Personal Data

Address	Süsterfeldstr. 36 52072 Aachen Germany
Phone	+49 160 8014812 (cellular)
Email	Carsten_Hof@web.de
Birth Date	18.05.1981
Birth Place	Siegen
Citizenship	German

Education

01/2006 - 12/2009	Doctoral studies as a member of the graduate school "Elementarteilchenphysik an der TeV-Skala", RWTH Aachen University Dr. rer. nat. (Ph. D.) in Physics, grade: (grades on a scale of 1–4 where 1 is best)
05/2003 - 12/2005	Studies of Physics, RWTH Aachen University Diploma in Physics, grade: 1.1
07/2001 - 04/2003	Studies of Physics and Mathematics, TU Kaiserslautern Pre-diploma in Physics, grade: 1.5 Pre-diploma in Mathematics, grade: 1.2
08/2000 - 06/2001	Civil service at the protestant parish administration, Betzdorf
During Civil Service	Correspondence course "Früheinstieg ins Physik-Studium" First two semesters of the regular Physics studies
08/1991 - 07/2000	Freiherr-vom-Stein Gymnasium, Betzdorf Abitur (university entrance diploma), grade: 1.3

Research Experience

Ph. D. Thesis	Title: "Implementation of a Model-Independent Search for New Physics with the CMS Detector exploiting the World-Wide LHC Computing Grid" **Main Focus:** – Feasibility study of a generic physics data analysis – Search for deviations from the Standard Model expectation caused by new physics signals – Grid Computing
The CMS Experiment	– Organization of the CMS model-independent search group – Installation and operation of the RWTH grid cluster – Coordination of the successful participation of the Aachen grid site within the "CMS Software, Analysis and Computing Challenges" – Leader of the CMS Europe grid-based Monte Carlo event production team – Development of experiment and analysis software (CMSSW) – CMS representative within the user committee of the German National Analysis Facility located at DESY, Hamburg – Regular research visits to CERN, Geneva
Conferences	– Poster presentation, Physics at the LHC, Cracow, 2006 – Talk at the Aspen Winter Conference, 2007 – Oral presentation, Physics at the LHC, Split, 2008 – Talk at the APS Spring Conference, Denver, 2009 – Various talks at the German Physics Society conferences (DPG)
Schools & Courses	– CERN summer student (3 months, summer 2004) C++ software development within the ROOT team – "Herbstschule für Hochenergiephysik", Maria Laach, 2005 – Fermilab Hadron Collider Summer School, Chicago, 2008 – Courses of the graduate school (e.g. Statistics (G. Cowan), C++ (P. Kunz), Supersymmetry (P. Zerwas))
Diploma Thesis	Title: "Search for New Heavy Charged Gauge Bosons with the Future CMS Detector" feasibility study for the search of new resonances at the LHC collider
Student Jobs	– Supervision and teaching of students (since 2001) – Metal workshop, AMS Elkenroth (repeatedly, several months)

Honors & Awards

2006: Physics diploma awarded with Schöneborn-Preis for outstanding study achievements

2000: Abitur awarded by the "Verein der Ehemaligen" for graduating as best of the year

Selected Publications

C. Hof *et al.*, "MUSiC – An Automated Scan for Deviations between Data and Monte Carlo Simulation", CMS PAS EXO-2008-005

C. Hof, "Searches for Supersymmetry in All-Hadronic and Lepton + Jets + MET Final States in CMS", CERN-CMS-CR-2008-106

C. Hof *et al.*, "CMS Monte Carlo Production in the WLCG Computing Grid", J. Phys. Conf. Ser. 119, 2008

C. Hof *et al.*, "CMS Monte Carlo Production Operations in a Distributed Computing Environment", Nucl. Phys. Proc. Suppl. 177-178, 2008

C. Hof, "Detection of New Heavy Gauge Bosons W' in CMS", Acta Phys. Polon. B38, 2007

I. Antcheva, R. Brun, C. Hof, F. Rademakers, "The Graphics Editor in ROOT", Nucl. Instrum. Meth. A559, 2006

Other Skills

Languages	German (mother tongue), English (fluent), French (basic)
Programming	C/C++ (expert), Turbo Pascal (advanced), Python (advanced), Bash (advanced), Java (basic), Fortran (basic)
Operating Systems	Mac OS, Windows, Unix/Linux
Miscellaneous	Grid Middleware, ROOT, Maple, Origin, MS Office Professional, Open Office, LaTeX, HTML, PHP
Teaching & Supervision	Tutoring of the lectures Physics 1–4, supervision of various diploma theses, student lab course assistance
Hobbies	Mountain hiking, climbing, skiing, ballroom dancing

Aachen, December 16, 2009