

Search for Supersymmetric Top-Quark Partners Using Support Vector Machines and Upgrade of the Hadron Calorimeter Front-End Readout Control System at CMS

Dissertation

zur Erlangung des Doktorgrades

des Fachbereichs Physik

der Universität Hamburg

vorgelegt von

Mehmet Özgür Şahin

aus Gaziantep



Hamburg

2016

Gutachter der Dissertation: Dr. Isabell-Alissandra Melzer-Pellmann
Prof. Dr. Peter Schleper

Vorsitzender des Prüfungsausschusses: Prof. Dr. Jan Louis

Vorsitzende des Promotionsausschusses: Prof. Dr. Daniela Pfannkuche

Leiter des Fachbereichs Physik: Prof. Dr. Peter Hauschildt

Dekan der Fakultät für Mathematik,
Informatik und Naturwissenschaften: Prof. Dr. Heinrich Graener

“Not all those who wander are lost”

— J. R. R. Tolkien

Abstract

In this thesis a search for direct pair production of supersymmetric top-quark partners as well as work on the upgrade of the front-end readout controller of the Hadron Calorimeter (HCAL) of the Compact Muon Solenoid (CMS) experiment are presented.

The most appealing extension of the Standard Model (SM) is supersymmetry (SUSY), relating the integer spin (bosons) and half-integer spin elementary particles (fermions). Supersymmetric top-quark partners ($\tilde{t}$) around and below the TeV energy scale offer a solution to the hierarchy problem. Furthermore, R-parity conserving SUSY models propose a cold dark matter candidate in the form of stable lightest supersymmetric particles, e.g. lightest neutralinos ($\tilde{\chi}^0$).

The analysis performed in this thesis is a search for top-squark pair production in a final state consisting of a single isolated lepton, jets, among which at least one is tagged as bottom-quark jet, and large missing transverse energy at the CMS experiment at the CERN Large Hadron Collider (LHC) with 8 TeV center-of-mass energy. A new Support Vector Machines (SVM) High-Energy Physics interface (SVM-HINT) software is introduced to classify signal events originating from new physics processes and the SM background. SVM-HINT is enhanced with a novel statistical significance based optimization algorithm providing a state-of-the-art classification power. Monte Carlo simulations are used in the training and optimization procedure, and high signal purity search regions are determined in the search for top-squark pair production. The background event yields in each search region are predicted using a data-driven background estimation method. The results are interpreted within a simplified model assuming a branching ratio of 100% to $\tilde{t} \rightarrow t \tilde{\chi}^0$. No significant discrepancy between the data and the SM predictions has been observed. Exclusion limits were derived to constrain the $m_{\tilde{t}}$ and $m_{\tilde{\chi}^0}$ of the investigated simplified model. The sensitivity of the previous searches with the 8 TeV center-of-mass energy in the single lepton final states is extended to $m_{\tilde{t}} = 675$ GeV and $m_{\tilde{\chi}^0} = 225$ GeV.

The results of the present analysis once again verified the necessity to reach higher center-of-mass energies and luminosities at the LHC. Such an upgrade will increase the radiation exposure of the readout electronics. A reliable operation of the detector electronics under these harsh conditions is absolutely crucial. Therefore, a new front-end readout control system has been integrated to the upgraded electronics infrastructure of the CMS HCAL, which simultaneously sets up and controls all front-end modules. Furthermore, it recovers diagnostic information and responses immediately in case of unexpected events. A firmware for the next-generation Front-End-Control module helping to accomplish these tasks has been developed. Consistency and reliability of the control system is successfully tested in the test-stands and irradiation beam tests.

Zusammenfassung

In der vorliegenden Dissertation wird eine Suche nach der direkten Paarproduktion von supersymmetrischen Top-Quark-Partnern sowie Arbeiten am Upgrade der Front-End-Controller des hadronischen Kalorimeters (HCAL) des Compact Muon Solenoid-Experiments (CMS) vorgestellt.

Eine der attraktivsten Erweiterungen des Standard-Modells der Teilchenphysik (SM) stellen supersymmetrische Modelle (SUSY) dar, die auf einer Symmetrie zwischen Elementarteilchen mit ganzzahligem Spin (Bosonen) und halbzahligem Spin (Fermionen) beruhen. Supersymmetrische Partner des Top Quarks ($\tilde{t}$) mit einer Masse um oder unter der TeV-Energieskala erlauben eine Lösung des Hierarchieproblems. Zudem ermöglichen SUSY-Modelle mit erhaltener R-Parität eine Erklärung von kalter dunkler Materie mit Hilfe des leichtesten supersymmetrischen Teilchens, des Neutralinos ($\tilde{\chi}^0$).

Die in dieser Arbeit vorgestellte Analyse sucht nach Anzeichen der Produktion von Top-Squark-Paaren mit dem CMS-Experiment am CERN Large Hadron Collider (LHC) bei einer Kollisionsenergie von 8 TeV. Die Analyse basiert auf Ereignissen mit einer Signatur bestehend aus einem einzelnen, isolierten Lepton, mehreren Jets von denen mindestens einer von einem b-Quark stammt, sowie fehlende Transversalenergie. Im Rahmen dieser Analyse wurde eine neue Schnittstelle für Support-Vektor-Maschinen in der Hochenergiephysik (SVM-HINT) entwickelt, um die ausgewählten Ereignisse als Signal oder SM-Untergrund zu klassifizieren. SVM-HINT wird zudem um einen neuartigen, auf statistischer Signifikanz basierten Algorithmus erweitert, der eine hohe Klassifikationsleistung aufweist. Es werden Monte Carlo-Simulationen verwendet, um den Algorithmus zu trainieren und zu optimieren sowie um Regionen des Phasenraums der Top-Squark-Paarproduktion mit großem Signal-zu-Untergrund-Verhältnis abzuschätzen. Der Anteil von Untergrundereignissen in jeder dieser identifizierten Suchregionen wird mit Daten-basierenden Methoden abgeschätzt. Die Ergebnisse werden in einen vereinfachten Modell unter der Annahme eines Verzweigungsverhältnisses von $\tilde{t} \rightarrow t \tilde{\chi}^0$ von 100% interpretiert. Es werden keine signifikanten Abweichungen zwischen der Messung und den Vorhersagen des SM beobachtet. Eine Ausschlussgrenze für die Massen $m_{\tilde{t}}$ und $m_{\tilde{\chi}^0}$ dem vereinfachten Modell wird abgeleitet. Die Sensitivität früherer Analysen bei einer Kollisionsenergie von 8 TeV mit einzelnen Leptonen im Endzustand wird auf $m_{\tilde{t}} = 675$ GeV sowie $m_{\tilde{\chi}^0} = 225$ GeV erweitert.

Das Ergebnis dieser Analyse stellt eine der Motivationen für eine höhere Kollisionsenergie sowie Luminosität des LHC dar. Ein Betrieb bei entsprechenden Parametern des Beschleunigers hat eine erhöhte Strahlenbelastung der Detektoren und Ausleseelektronik zur Folge. Der zuverlässige Betrieb der Elektronikkomponenten unter diesen Bedingungen ist von entscheidender Wichtigkeit. Deshalb wird ein neues Kontrollsystem in die erneuerte Front-End-Ausleseelektronik des CMS HCAL integriert, welches ein simultanes Konfigurieren und Steuern aller Front-End-Module ermöglicht. Zudem erlaubt es im Falle eines unvorhergesehenen Ereignisses den unmittelbaren Zugriff auf Diagnoseinformationen. Im Rahmen dieser Arbeit wurde die Firmware für diese next-generation-Front-End-Controller entwickelt und die nötigen Routinen für diese Funktionalität implementiert. Die Konsistenz und Zuverlässigkeit des neuen Kontrollsystems wurde erfolgreich in Laborexperimenten sowie Teststrahlungsmessungen bestätigt.

Contents

1	Introduction	1
2	Theoretical Framework	3
2.1	Standard Model	3
2.1.1	Elementary particles of the Standard Model	4
2.1.2	The Lagrange formalism and the gauge transformations	5
2.1.3	Quantum Chromodynamics	6
2.1.4	Electroweak theory	6
2.1.5	Shortcomings of the Standard Model	9
2.2	Beyond the Standard Model	11
2.2.1	Supersymmetry	11
2.2.2	The Minimal Supersymmetric extension of the Standard Model	13
2.2.3	Searches for Supersymmetry at colliders	16
3	Experimental Setup	19
3.1	Large Hadron Collider	20
3.2	The Compact Muon Solenoid detector	23
3.2.1	Inner tracking detector	25
3.2.2	Electromagnetic calorimeter	27
3.2.3	Hadron calorimeter	28
3.2.4	Forward detector	29
3.2.5	Muon system	30
3.2.6	Trigger and data acquisition	30
3.2.7	Luminosity determination	32
3.2.8	Offline computing	32
3.3	Monte Carlo simulations	32
4	The CMS HCAL Front-end Readout Control System Upgrade	35
4.1	μ TCA	36
4.1.1	CMS HCAL μ TCA	40

4.2	The control system structure	42
4.2.1	CCM server, TCDS, TTC and AMC13	42
4.2.2	ngFEC-FC7	45
4.2.3	Front-end crate	53
4.3	System test	55
4.3.1	ngFEC test	55
4.3.2	Control system test	56
4.4	Outlook	58
5	Object Reconstruction and Variable Definitions	59
5.1	Object reconstruction	60
5.1.1	Particle-flow algorithm	60
5.1.2	Primary vertex	60
5.1.3	Muons	61
5.1.4	Electrons	63
5.1.5	Taus	64
5.1.6	Jets	65
5.1.7	Missing transverse energy E_T	69
5.2	Event and object selection	71
5.2.1	Event cleaning	73
5.2.2	Preselection	74
5.3	Variable definitions	76
5.3.1	Low-level variables	76
5.3.2	High-level variables	78
6	SVM	83
6.1	A brief introduction to machine learning	84
6.2	Support vector machines	85
6.2.1	Linearly separable distributions	86
6.2.2	Overlapping distributions	89
6.2.3	Non-linear distributions	90
6.2.4	Computing SVM	92
6.2.5	Probability output	93
6.2.6	Hyper-parameter tuning	94
6.3	Binary Decision Trees	96
6.3.1	Boosting	97
6.4	Classifier implementations	98
6.4.1	Method of comparing performance of classifiers	98
6.4.2	SVM hyper-parameter optimization algorithm	99
6.4.3	Parallelization	100
6.5	Case studies	101
6.5.1	Performance comparison on a toy model	101

6.5.2	Third generation supersymmetric partner search	102
6.6	Summary	107
7	Analysis strategy	109
7.1	SVM hyperparameter tuning	110
7.2	Simulation validation	115
7.2.1	$W + \text{jets}$ and $1l \text{ top}$ validation in CR1: $0b$ jet final state	116
7.2.2	$t\bar{t} \rightarrow l\bar{l}$ validation in CR2: $2l$ final states and CR3: $l\tau$ final states	117
7.3	Background Estimation	119
8	Results and Interpretation	123
8.1	Systematic uncertainties	123
8.1.1	Systematic uncertainties on the background estimation	124
8.1.2	Systematical uncertainties on the signal samples	126
8.2	Interpretation and comparison	129
8.2.1	Comparison with similar studies	131
9	Conclusion nad Outlook	133
	Appendices	137
A	Gigabit Transreceiver Protocol	139
B	Lepton Identification and Isolation Efficiency	143
C	Statistical Inference	147
	Bibliography	151

Introduction

The human curiosity stimulated by the fascination of Nature and Universe can be seen as the origin of the scientific conception of Physics. Physics' range extends over a spatial distance starting from the smallest *Planck scale* to the diameter of the observed Universe. In order of magnitude, the *range* exceeds the letter count of this sentence¹. What is more fascinating than this range is that how the smallest and the biggest are connected to each other with experimental observations of Nature. Relating these two ends in a unified, descriptive theory is one of the utmost objectives of the field of Physics. The Standard Model (SM) of particle physics is a great leap forward in explaining the smallest, while it opens a descriptive window to the physics of large scales.

The SM is a theory framework describing the elementary particles and fundamental forces of Nature: the half-integer-spin fermions consisting three lepton and quark families, and their electroweak and strong interactions mediated by bosons with integer spins. The SM has been meticulously constructed and tested, resulting in the one of the most successful particle physics theories ever considered. However, it fails to explain a fundamental force of Nature, gravity, and its particle content accounts for only less than 5% of the Universe's energy and mass. Despite its fascinating predictions, some observations can only be added in an ad-hoc way, which puzzles particles physicists ever since, e.g. the mass of the recently discovered Higgs boson is required to be immensely fine tuned within the SM to explain the experimental observations, this *unnatural* fine-tuning is also referred to as the *hierarchy* problem.

The most prominent extension of the SM is the Supersymmetry (SUSY) relating bosons and fermions. The supersymmetric models respecting this space symmetry account for most of the questions left unanswered by the SM, while providing an aesthetically appealing structure. Supersymmetric top-quark partners ($\tilde{t}$) around and below the TeV energy scale, for instance, offer a solution to the hierarchy problem. Furthermore, R-parity conserving SUSY models propose a cold dark matter candidate in the form of stable lightest supersymmetric particles as neutralinos ($\tilde{\chi}^0$).

The experimental observations in the particle colliders and astrophysical probes show no indi-

¹The order of magnitudes are considered in metric units. The order of magnitude of the Planck length is about 10^{-35} m, and the diameter of the observed universe is taken approximately 10^{27} m [1–3].

cation of SUSY. Therefore, the Large Hadron Collider (LHC) as the most powerful particle collider to date, is the most capable tool for searching new physics and hence SUSY. The Compact Muon Solenoid (CMS) is one of two multi-purpose detectors collecting the data delivered by the LHC.

The analysis performed in this thesis is a search for top-squark pair production in a final state consisting of a single isolated lepton, jets, among which at least one is tagged as bottom-quark jet, and large missing transverse energy at the CMS experiment at the CERN LHC with 8 TeV center-of-mass energy. With an increasing top-squark mass, these events become extremely rare to observe in CMS. By drawing an analogy to the famous idiom "searching for a needle in haystack", the search for the top-squark partner can be described as "searching for a slightly different and rare hay in a haystack". Therefore, advanced machine learning algorithms can be used to exploit nuances between the kinematic distributions of SM and SUSY processes. With this purpose, a new Support Vector Machines (SVM) High-Energy Physics interface (SVM-HINT) software is introduced to classify signal events originating from new-physics processes and the SM background. SVM-HINT is enhanced with a novel statistical significance based optimization algorithm providing a state-of-the-art classification power. Monte Carlo simulations are used in the training and optimization procedure, and high signal purity search regions are estimated in the search for top-squark pair production. The background event yields in each search region are predicted using a data-driven background estimation method. The results are interpreted within a simplified model assuming a branching ratio of 100% to $\tilde{t} \rightarrow t \tilde{\chi}^0$.

An upgrade of the LHC and the detectors has been planned to extend the physics reach of the experiments. Within this upgrade effort, the electronics infrastructure of the CMS detector is completely redesigned to cope with the higher energy and luminosity. Such an upgrade will also increase the radiation exposure of the readout electronics. A reliable operation of the detector electronics under these harsh conditions is absolutely crucial for the continuous data taking which is required to observe extremely rare new-physics events. Therefore, a new front-end read-out control system has been developed and integrated into the upgraded electronics infrastructure of the CMS HCAL, which simultaneously sets up and controls all front-end modules. In addition, it recovers diagnostic information and responses immediately in case of unexpected events.

The thesis organized as follows: A brief overview of the SM and its supersymmetric extension is given in Chapter 2. After introducing the experimental setup consisting of the LHC and the CMS detector in Chapter 3, the CMS-HCAL front-end-readout control upgrade is discussed in Chapter 4. Chapter 5 describes the reconstruction of the physics objects, and their features that are used in the present analysis. The concept of SVM, and the novel optimization technique and the software interface (SVM-HINT) developed within the thesis work are introduced in Chapter 6. The analysis strategy followed in the search for top-quark supersymmetric partners together with the background estimation methodology is described in Chapter 7. Finally, the results of the analysis and its interpretation are presented in Chapter 8.

Theoretical Framework

Contents

2.1 Standard Model	3
2.1.1 Elementary particles of the Standard Model	4
2.1.2 The Lagrange formalism and the gauge transformations	5
2.1.3 Quantum Chromodynamics	6
2.1.4 Electroweak theory	6
2.1.5 Shortcomings of the Standard Model	9
2.2 Beyond the Standard Model	11
2.2.1 Supersymmetry	11
2.2.2 The Minimal Supersymmetric extension of the Standard Model	13
2.2.3 Searches for Supersymmetry at colliders	16

The experimental observations of the sub-atomic particles unveiled a new window for the physicists. By 1970s, a theoretical framework, the Standard Model of high energy physics, explaining the observations and, more importantly, predicting new degrees of freedom has been formed. Despite its breathtaking success, the SM is being challenged by the particle physicists due to some conceptual gaps, and questions related to unaccounted experimental observations, and therefore, new theories that reach *beyond* the Standard Model are suggested.

In this chapter a very brief overview of the SM is given. The theories beyond the Standard Model with a special focus on Supersymmetric models, and their experimental status are discussed¹.

2.1 Standard Model

The Standard Model describes the elementary particles and their interactions (i.e. electromagnetic, weak, and strong interactions) within a field theory framework based on the $SU(3)_C \times$

¹The overviews are based on various textbooks and review papers [4–8].

$SU(2)_L \times U(1)_Y$ gauge group. The spin $1/2$ fermions form the matter, and bosons as the interaction mediators with integer spin complete the particle content of the SM.

The strong interactions are discussed under the Quantum Chromodynamics, whereas the electromagnetic and weak interactions are unified within a common electroweak theory.

2.1.1 Elementary particles of the Standard Model

The field interactions of the SM are mediated by massless gluon, photon, and massive Z^0 and $W^\pm$ bosons. The recently discovered Higgs boson [9, 10] is the quantization of the Higgs field which gives mass to the SM particles and Higgs boson itself.

	fermions			bosons	
quarks	mass: 2.2 MeV charge: +2/3 spin: 1/2 u	mass: 1.27 GeV charge: +2/3 spin: 1/2 c	mass: 173 GeV charge: +2/3 spin: 1/2 t	0 0 1 g	mass: 125 GeV 0 0 H
	mass: 4.7 MeV charge: -1/3 spin: 1/2 d	mass: 96 MeV charge: -1/3 spin: 1/2 s	mass: 4.18 GeV charge: -1/3 spin: 1/2 b	0 0 1 γ	
	mass: 0.51 MeV charge: -1 spin: 1/2 e	mass: 105 MeV charge: -1 spin: 1/2 μ	mass: 1.78 GeV charge: -1 spin: 1/2 τ	mass: 91.2 GeV 0 1 Z	
leptons	~0 0 1/2 ν_e	~0 0 1/2 ν_μ	~0 0 1/2 ν_τ	mass: 80.4 GeV ± 1 1 W	

Fig. 2.1: Elementary particles of the Standard Model presented with their quantum numbers and masses [6].

There are three lepton and quark families forming the matter in the SM. Each of these families has a heavier and lighter constituent. The minimum number of quark families is set by the CKM (Cabibbo – Kobayashi – Maskawa) matrix in order to include charge-parity violation into the quark mixing matrix that quantifies the decay probabilities between the quarks. Quarks of the SM are the main constituents of hadrons. The bound states they form can be successfully described by the Gell-Mann – Zweig quark model based on the $SU(3)$ symmetry. The latest discovered t quark is the heaviest elementary particle reaching the mass of a gold atom. A general overview of the SM particles with their masses is shown in Fig. 2.1.

Each lepton family consists of a heavier charged lepton, and a nearby neutral neutrino. The interactions within the SM are observed to conserve the lepton number of the same family, e.g. the number of electrons and electron neutrinos entering and exiting an interaction is conserved.

The restriction on the number of lepton generations arises from the measurements of ratio of the Z boson's decay width into invisible modes (neutrinos):

$$\frac{\Gamma_{\text{invs}}}{\Gamma_\ell} = \frac{N_\nu \Gamma(Z \rightarrow \bar{\nu}\nu)}{\Gamma_\ell} \sim 2N_\nu, \quad (2.1)$$

where the observed value of $\Gamma_{\text{invs}}/\Gamma_\ell$ is 5.942 ± 0.016 [6]. This observation put a minimum mass constraint of $m_{Z^0}/2$ on a possible fourth neutrino flavor.

2.1.2 The Lagrange formalism and the gauge transformations

The Lagrange function is a mathematical construction used to obtain the stationary points of a constrained problem. In the context of field theories, it is used to minimize the action, thus it is often employed to describe a dynamic system within certain constraints.

The evaluation of the system can be obtained by using the Euler-Lagrange equations given by

$$\partial_\mu \left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi_i)} \right) = \frac{\partial \mathcal{L}}{\partial \phi_i}, \quad (2.2)$$

where $\mathcal{L}$ is Lagrangian of the system, and ϕ_i are the fields. The solutions to Eq. 2.2 provide the equations of motion of the system of interest. The particles of the SM are realized as the quantization of these fields.

Imposing invariance under local gauge transformations, introduces a set of gauge potentials coupling to scalar and fermion matter fields.

An example can be given by requiring the $\mathcal{L}$ for fermions to be invariant under local transformations such as

$$\psi(x) \rightarrow e^{iq_\psi \alpha(x)} \psi(x), \quad (2.3)$$

where ψ is wave function of spin-1/2 particle, $\alpha(x)$ is a scalar phase and q_ψ is the electric charge, resulting in introduction of the vector potential A_μ with a coupling described by the electromagnetic field Lagrangian

$$\mathcal{L} = i\bar{\psi}\gamma_\mu D^\mu \psi - m\bar{\psi}\psi - \frac{1}{4}F^{\mu\nu}F_{\mu\nu}, \quad (2.4)$$

where D_μ is the gauge covariant derivative given by $D_\mu = \partial_\mu + iq_\psi A_\mu(x)$, and $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. Simultaneously, A_μ is transformed as

$$A_\mu(x) \rightarrow A_\mu(x) - \partial_\mu \alpha(x), \quad (2.5)$$

The Lagrangian given in Eq. 2.4 describes the interactions of charged fermions within the Quantum Electrodynamics (QED), and the discussed scalar phase transformations with an integer electric charge belongs to the unitary group $U(1)_{\text{em}}$. Solutions to the Euler-Lagrange equations of the Eq. 2.4, describe the nature of the QED interactions.

2.1.3 Quantum Chromodynamics

The color quantum numbers are introduced to solve the puzzle caused by the observation of hadron states like Δ^{++} (uuu). Without three color charges, such low lying fermion states would violate the Pauli's exclusion principle.

Three different color charges (red, blue, green) form an exact $SU(3)_C$ symmetry group. This local gauge symmetry dictates interactions between the color-charged quarks mediated by gluons, resulting in the theory of Quantum Chromodynamics (QCD). Gluons also carry color charges, thus enabling the self-interaction of the mediators.

The initial postulate of the QCD, known as the confinement phenomenon, dictates that only color singlet (colorless) particles can be observed in nature. The particles with color charges confine together to form colorless composite particles. Therefore, quarks and gluons can only be found as bound states. Another characteristic feature of the QCD can be observed when the quarks have high energies or really close to each other: They act as if they are free particles. This property is known as the asymptotic freedom [11, 12]. Therefore, a naive model of the QCD interactions can be built using a linear potential of the form $V(r) \sim r$.

Similar to the QED, a Lagrangian is constructed for the QCD with three color charges,

$$\mathcal{L} = i\bar{\psi}\gamma_\mu D^\mu\psi - m\bar{\psi}\psi - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} - (q_s\bar{\psi}\gamma^\mu\lambda\psi)A_\mu, \quad (2.6)$$

where λ represents the Gell-Mann matrices, and q_s is the corresponding color charge. Comparing to Eq. 2.4, the QCD Lagrangian has an additional term due to the self-interaction of gluons.

2.1.4 Electroweak theory

In order to explain the β decay of the neutron, Enrico Fermi formulated a single point interaction with an effective coupling G_F . However, this effective theory is only valid within the energy scale of $m_{W^\pm}$. A more generalized solution to the problem is given by Glashow-Salam-Weinberg model as a unified theory of the QED and weak interactions. Unification of electromagnetic and weak interactions is the most encouraging step towards the grand unified theories (GUT). Experimental observations indicate that the weak interactions has a chiral structure (unlike QED or QCD) such that only left handed particles and right handed anti particles interact via the weak interaction. Therefore, the mediators of weak interaction, Z^0 and $W^\pm$, only couple to the left handed-field doublets.

The simplest unitary group that includes these doublets is $SU(2)$. The gauge transformations that leave the weak interaction invariant in flavor space also introduce $U(1)_Y$ instead of the simple $U(1)_{em}$ group of QED, where hypercharge (Y) is defined as

$$\frac{Y}{2} = Q - T_3, \quad (2.7)$$

where $T_3 = \sigma_3/2$ and Q stands for the electromagnetic charge operator. Therefore, the electroweak theory (EWK) is a unified $SU(2)_L \times U(1)_Y$ gauge symmetric theory. Similar to the QED

and QCD, the covariant derivative can be given as

$$D_\mu = \partial_\mu - ig' \frac{1}{2} Y B_\mu - ig T W_\mu, \quad (2.8)$$

where the g and g' are the coupling constants, B_μ is the gauge field, T is a vector of Pauli matrices and W_μ is the three vector gauge field. The components of T (Pauli matrices) transform the isospin doublets, hence do not affect the right-handed fermions.

The Lagrangian of the electroweak interaction can be written as

$$\mathcal{L} = \bar{\psi}(i\gamma^\mu D_\mu - m)\psi - \frac{1}{4}W_{\mu\nu}W^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} \quad (2.9)$$

where $W_{\mu\nu} = \partial_\mu W_\nu - \partial_\nu W_\mu - g(W_\mu \times W_\nu)$, and $B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu$. The four gauge fields B_μ , $W_\mu^{1,2,3}$ introduced in Eq. 2.9 are not mass eigenstates of the EWK bosons. The mass eigenstates of the EWK bosons are obtained from following relations

$$\begin{pmatrix} \gamma \\ Z^0 \end{pmatrix} = \begin{pmatrix} \cos \theta_W & \sin \theta_W \\ -\sin \theta_W & \cos \theta_W \end{pmatrix} \begin{pmatrix} B \\ W^3 \end{pmatrix} \quad (2.10)$$

and

$$W^\pm = W^1 \pm iW^2, \quad (2.11)$$

the θ_W term is known as the Weinberg angle or the EWK mixing angle.

An unfortunate feature of the EWK Lagrangian is that no mass terms appear for the gauge fields. In fact, massive bosons are forbidden by the EWK gauge symmetry, i.e. these mass terms would violate the gauge invariance of Eq. 2.9. However, as shown in Fig. 2.1, unlike photons, Z^0 and $W^\pm$ bosons are massive particles and their masses are in the range of 80 GeV to 100 GeV. The problem is solved adequately by the spontaneous symmetry breaking mechanism through the Higgs field and Higgs particle.

Higgs mechanism

The $SU(2)_L \times U(1)_Y$ symmetry of the SM is spontaneously broken (i.e. the symmetry is not conserved only in the ground state of the potential) to $U(1)_{em}$ via the Higgs mechanism and mass is given to elementary particles in the process.

This is achieved by introducing a simplest possible potential given by

$$V = \mu^2 \Psi^\dagger \Psi + \lambda (\Psi^\dagger \Psi)^2, \quad (2.12)$$

where μ and λ are real numbers, and the Ψ is an isospin doublet of the form

$$\Psi = \begin{pmatrix} \psi^+ \\ \psi^0 \end{pmatrix}. \quad (2.13)$$

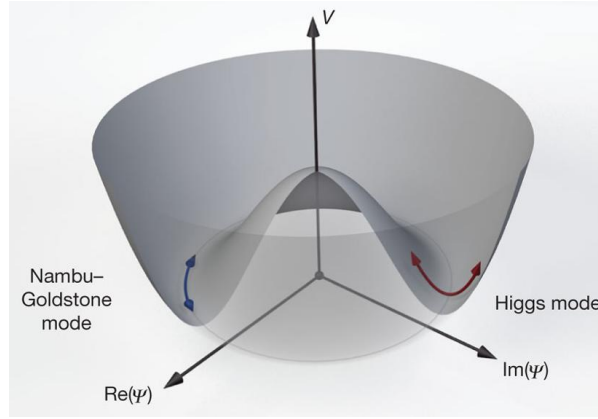


Fig. 2.2: The Higgs potential in the form of a bottom of a wine bottle in real and imaginary dimensions [13]. In the reference of the ground state, the potential takes an asymmetric form. Possible excitations of bosons in the ground state are indicated with arrows.

The potential has a form similar to the bottom of a wine bottle as shown in Fig. 2.2. In the reference frame of the particles at the ground state (where $V = 0$) system has a broken symmetry. For $\mu^2 < 0$, the Higgs field in the ground state can be written as

$$|\Psi\rangle = \sqrt{\frac{1}{2} ((\text{Re } \psi^+)^2 + (\text{Im } \psi^+)^2 + (\text{Re } \psi^0)^2 + (\text{Im } \psi^0)^2)} = \sqrt{\frac{-\mu^2}{2\lambda}}. \quad (2.14)$$

A gauge is chosen such that only real part of the ψ_0 takes a non-zero value, leading to

$$\text{Re } \psi^+ = \text{Im } \psi^+ = \text{Im } \psi^0 = 0, \text{Re } \psi^0 = \sqrt{\frac{-\mu^2}{2\lambda}} \equiv v, \quad (2.15)$$

where v is the vacuum expectation. As a result of this gauge choice, three massless Goldstone boson and one massive Higgs boson are formed as excitations around these minima. By imposing the gauge invariance, the Goldstone bosons disappear under local transformations (the gauge transformation *eats* the Goldstone bosons).

The coupling of the Higgs boson with the other EWK bosons can be written as

$$\mathcal{L}_{\text{HVV}} = gm_W H_{\text{SM}} (W_\mu^+ W^{\mu-} + \frac{1}{2} \sec^2 \theta_W Z_\mu Z^\mu), \quad (2.16)$$

and

$$\mathcal{L}_{\text{HHVV}} = \frac{g^2}{4} (W_\mu^+ W^{\mu-} + \frac{1}{2} \sec^2 \theta_W Z_\mu Z^\mu) H_{\text{SM}}^2, \quad (2.17)$$

the self coupling of the Higgs boson, in terms of its mass term $m_{H_{\text{SM}}} = \sqrt{-2\mu^2}$

$$\mathcal{L}_{\text{HH}} = -\frac{gm_W^2}{4m_W} H_{\text{SM}}^3 - \frac{g^2 m_{H_{\text{SM}}}^2}{32m_W^2} H_{\text{SM}}^4. \quad (2.18)$$

Finally, the mass of the fermions is included following the experimental input via Yukawa

coupling

$$\mathcal{L}_{\text{Yukawa}} = - \sum_{i=1} \frac{gm_{f_i}}{2\sqrt{2}m_W} \bar{f}_i f_i H_{\text{SM}}, \quad (2.19)$$

The spontaneous symmetry breaking is one of the success pillars of the SM. However, the arbitrariness included with the Higgs mechanism cannot be described within the theory, thus driving physicists to explore possible extensions to the SM.

2.1.5 Shortcomings of the Standard Model

The Standard Model is an enormously successful theory: its predictions have been tested via various particle collider experiments (e.g. SLAC, LEP, HERA, Tevatron, LHC and many more) where the controlled collisions of the stable elementary or composite particles are analyzed using advanced detectors, in addition to the nature observations. Nevertheless, the SM leaves questions arise from some experimental observances unanswered. Moreover, it fails to give intuitive solutions to some theoretical considerations:

Dark matter and energy

By observing the COMA galaxy cluster's movement, Zwicky introduced the concept of *dark matter* [14] that interacts with the SM fields either very weakly or not at all. The dark matter can be probed using the gravity: the rotation of the galaxies indicates that a large amount of material is filling the galactic interspace causing galaxy velocity of the spiral arms to be independent of their distance to the center. The fluctuations in the spectrum of the relic microwave background provide another hint. The SM is unable to offer a candidate for the dark matter.

A dark energy permeating the whole space is required to be introduced in order to explain the measurements of the cosmic microwave background [15, 16] and of the type-Ia supernovae at large red-shifts [17, 18], hence the observation of the accelerated expansion of universe.

Following these arguments, the matter and energy composition of the universe can be listed as [16]:

- ordinary SM matter: 4.9%,
- dark matter: 26.8%,
- dark energy: 68.3%.

Therefore, all known matter can only be accounted for 5% of Universe.

Neutrino mass

The solar and atmospheric neutrino data indicate that neutrinos are oscillating between the flavors [19, 20]. These oscillations, later verified using the nuclear reactor [21] and collider neutrinos [22], suggest that at least two of the neutrinos should be massive since the oscillation frequencies are proportional to the mass difference of the neutrino flavors. Even though, within

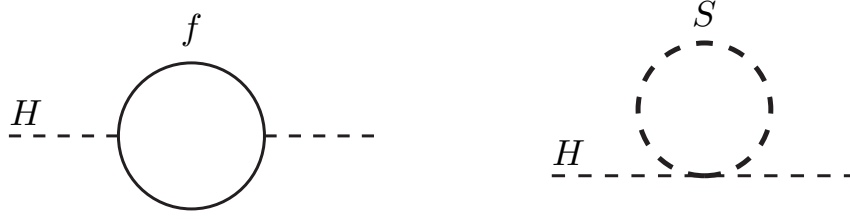


Fig. 2.3: Feynmann diagrams of the first order loop corrections due to a fermion (left) and a scalar boson (right) [8].

the SM, neutrinos are assumed to be massless, their mass can be introduced as an input to fermion Yukawa couplings. This approach would still be unable to explain the huge mass difference within the lepton families.

Existence of gravity

The gravitational interactions are not included by the SM. In fact, a renormalizable four-dimensional field theory that describes the gravitational interactions does not exist.

Theoretical considerations

There are theoretical arguments or inconsistencies in the SM, which hint to the necessity of expanding or even reconstructing the theory. Most of these considerations are related to the arbitrariness of the mass terms and selection of the symmetry groups which are mostly adopted in regard to experimental observations without satisfying theoretical explanations.

Hierarchy problem

All massive particles in the SM contribute to the radiative corrections to the Higgs boson bare mass, including the Higgs boson itself. The Feynman diagrams of the first order loop corrections due to a scalar (left) boson and a fermion (right) are shown in Fig. 2.3.

Considering the second term in the Higgs self-coupling Lagrangian in Eq. 2.19, the energy correction can be calculated using perturbation theory for the one-loop scalar-boson correction:

$$\delta m_{H_{SM}}^2 = \langle H_{SM} | \frac{g^2 m_{H_{SM}}^2}{32 m_W^2} H_{SM}^4 | H_{SM} \rangle = \frac{12 g^2 m_{H_{SM}}^2}{32 m_W^2} \frac{1}{16 \pi^2} \left(\Lambda^2 - m_{H_{SM}}^2 \ln \frac{\Lambda^2}{m_{H_{SM}}^2} + \mathcal{O}\left(\frac{1}{\Lambda^2}\right) \right) \quad (2.20)$$

The Higgs mass can be rewritten in terms of the quadratic term in the correction:

$$m_{H_{SM}}^2(\text{phys}) \simeq m_{H_{SM}}^2 + \frac{c_{H_{SM}}}{16 \pi^2} \Lambda^2 \quad (2.21)$$

here Λ is a cut-off value indicating the energy range of validity of the given theory. If the SM is the ultimate theory of nature, Λ must be around the reduced Planck scale (10^{18} GeV) where the quantum gravitational effects become important. This would require an extreme correction or a so called fine-tuning of 1 over 10^{30} in the observed 125 GeV boson Higgs mass.

Considering the SM as an effective theory (similar to Fermi's weak interaction model) and the absence of the fine-tuning as a validity condition, the Λ scale can be set to $\mathcal{O}(\text{TeV})$. This implies that the effective SM ceases to be valid around TeV scale, hinting for new physics at TeV scale.

The fermionic first order loop can be treated in a similar way as the scalar loop. In this case the contribution must be calculated using the Yukawa coupling Lagrangian given in Eq. 2.19.

$$m_{H_{\text{SM}}}^2(\text{phys}) \simeq m_{H_{\text{SM}}}^2 - c_{f_{\text{SM}}}^2 \left(\frac{1}{8\pi^2} \Lambda^2 - m_{f_{\text{SM}}}^2 \ln \frac{\Lambda^2}{m_{f_{\text{SM}}}^2} \right) \quad (2.22)$$

where the sign change is caused by the anti-symmetric nature of the fermion wave functions. The highest contribution to this term is expected from the heaviest SM fermion, i.e. the t quark.

2.2 Beyond the Standard Model

A large number of Beyond the Standard Model (BSM) theories in the form the SM extensions have been proposed to address the shortcomings of the SM. The most appealing proposal that offers a solution for each problem left unanswered by the SM is given in the form of a relativistic space-time supersymmetry, and known as supersymmetric extensions of the SM.

The present work focuses on a subset of these models, thus a relatively detailed overview is given in the following.

2.2.1 Supersymmetry

A symmetry between the fermionic and bosonic degrees of freedom intrigued the particle physicists starting from the first half of twentieth century. However, a theory providing a four-dimensional relativistic space-time supersymmetry has not been realized until the 1970s. This appealing Supersymmetric theory (SUSY) relies on linear transformations of the original fields. Each field of the SM is associated to a superfield introducing a supermultiplet. The quantum numbers of the supermultiplets differ from the SM partners only by spin $1/2$.

Implying an exact supersymmetry on the SM would require mass degenerate superpartners for each SM particle, which has not been observed experimentally. Therefore, a breaking of the supersymmetry in high energies is inferred. The origin of the SUSY breaking mechanism contains a lot of arbitrariness and is the weakest point of all supersymmetric extensions of the SM.

The motivations for examining SUSY is numerous; more importantly its existence may lead to impressive results that can be listed as:

Solution of the hierarchy problem

SUSY theories offer a solution for the SM hierarchy problem. Assuming the existence of a heavy complex scalar particle S with mass m_S , a similar contribution given in Eq. 2.21 can be written

for any scalar particle's loop contribution to the Higgs mass:

$$\Delta m_H^2 \simeq \frac{c_S}{16\pi^2} \Lambda^2, \quad (2.23)$$

where the terms logarithmically diverging in Λ are neglected. The comparison of Eq. 2.21 and Eq. 2.23 shows that $\lambda_S = 2|\lambda_f|^2$ exactly cancels the quadratical divergences. In SUSY models, each heavy fermion is associated to two scalars, hence the factor two has to be removed to cancel quadratic divergences:

$$c_S = |c_{f_{\text{SM}}}|^2. \quad (2.24)$$

Requiring Eq. 2.24, the remaining dominant logarithmic terms depends directly on the masses of the SM particle and its superpartner. For the fermionic case

$$\delta m_H^2 \simeq \frac{c_S}{16\pi^2} \left[m_{f_{\text{SM}}}^2 \ln \left(\frac{\Lambda}{m_{f_{\text{SM}}}} \right) - m_S^2 \ln \left(\frac{\Lambda}{m_S} \right) \right]. \quad (2.25)$$

Having $\Lambda \approx M_{\text{GUT}}$ and the $c_S \approx 1$, the lightest superpartners of the highest mass fermions must have a mass of the order of $\mathcal{O}(\text{TeV})$ to avoid a considerable fine tuning.

Unification of the gauge couplings

Precise measurements of the running gauge couplings performed at the Large Electron-Positron Collider (LEP) show that these do not unify when they are extrapolated from $Q = M_{Z^0}$ to the $Q = M_{\text{GUT}}$ scale using the renormalization group equations of the SM. Astonishingly, the supersymmetric evolution equations and provided superpartner masses around 1 TeV unify these three running coupling constant, hinting an aesthetically pleasing theory.

Cold dark matter candidate

In the SM, the baryon and lepton numbers are conserved due to the local gauge invariance principles. In SUSY theories, however, the baryon and lepton number conservations can be violated. However, such a violation has not been observed, and the lower limit on the proton lifetime of 10^{30} years [6] is a very strong argument in favor of baryon number conservation. In order to realize this conservation, a new quantum number, called *R-parity* is introduced:

$$R = (-1)^{3B+L+2S}, \quad (2.26)$$

where B is the baryon number, L is the lepton number, and S is the spin. By construction, all SM particles have positive R-parity, whereas all superpartners have a balancing negative R-parity. Imposing the conservation of R-parity prevents the protons decay, and implies that the SUSY particles can only be produced in pairs. Another important feature of such a conservation is that it provides a stable the lightest supersymmetric particle (LSP). The LSP is electrically and color neutral, such as neutral electroweak bosons and Higgs superpartners, thus it is one of the most prominent cold dark matter candidates.

Inclusion of gravity

Imposing a local symmetry on SUSY models introduces a spin-2 massless gauge field, the graviton, that mediates gravity together with its superpartner, the gravitino. The models where such a local invariance is imposed are referred to as *supergravity*. Like any quantum field theory of gravity, supergravity is also non-renormalizable. Nevertheless, the connection to gravity is appealing. The interested reader can find further information about supergravity in Ref. [8].

2.2.2 The Minimal Supersymmetric extension of the Standard Model

The Minimal Supersymmetric extension of the Standard Model (MSSM) is a direct extension of the SM, which realizes supersymmetry. It is the minimal supersymmetrical model that contains the smallest number of particle states and new interactions consistent with the phenomenology of particles. All SM particles are arranged with their superpartners in either chiral or gauge supermultiplets:

- Chiral (or scalar) supermultiplets: one massless spin-1/2 Weyl fermion with two chirality states together with one complex scalar field that can be decomposed in two real scalar fields.
- Gauge supermultiplets: one massless real spin-1 vector boson and one massless spin-1/2 Weyl fermion.

The supermultiplets of the MSSM are summarized in Table 2.1. The gauge supermultiplets consist of spin-1 vector bosons, whereas the Higgs particles have spin 0 and, therefore, must be arranged into chiral supermultiplets. Quarks and leptons are organized into chiral supermultiplets to allow the correct transformation of their left- and right-handed components under Lorentz transformations. Two Higgs doublets cancel out the contributions of the Higgs partners to the gauge anomalies.

The Lagrangian density of the MSSM can be constructed in analogy to the SM Lagrangians discussed earlier. Possible supersymmetric and gauge invariant interaction vertices can be derived from the SM vertices by replacing any two SM particles with the corresponding superpartners.

Additional terms are included to break supersymmetry. All possible terms that do not lead to quadratic divergences, which are consistent with $SU(3)_C \times SU(2)_L \times U(1)_Y$ are added to the Lagrangian. Such breaking is, therefore, referred to as *soft SUSY breaking*.

Similar to the SM, the EWK symmetry breaking in the MSSM is described by the Higgs mechanism. In the MSSM, the Higgs scalar fields are arranged into two complex $SU(2)_L$ doublets, consisting of eight degrees of freedom. After the electroweak symmetry is spontaneously broken, three degrees of freedom become the longitudinal modes of the Z^0 and $W^\pm$ bosons, which acquire mass. The remaining five degrees of freedom manifest themselves as five physical Higgs bosons: two charged Higgs bosons, $H^\pm$; one CP-odd neutral Higgs boson, A^0 ; two CP-even neutral Higgs bosons, H^0 and h^0 , with h^0 being the lightest by general convention.

Table 2.1: The field content of the MSSM and corresponding quantum numbers due to the symmetry groups [6]. Only one generation of quarks and leptons presented. The spin-0 fields are complex scalars, and the spin-1/2 fields are left-handed two-component Weyl fermions.

Super-multiplets	Boson fields	Fermionic partners	SU(3)	SU(2)	SU(1)
gluon/gluino/	g	$\tilde{g}$	8	1	0
gauge/	$W^\pm, W^0$	$\tilde{W}^\pm$	1	3	0
gaugino	B^0	$\tilde{B}^0$	1	1	0
slepton/	$(\tilde{\nu}, \tilde{e})_L$	$(\nu, e)_L$	1	2	-1
lepton	$\tilde{e}_R$	e_R	1	1	-2
squark/	$(\tilde{u}, \tilde{d})_L$	$(u, d)_L$	3	2	1/3
quark	$\tilde{u}_R$	u_R	3	2	4/3
	$\tilde{d}_R$	d_R	3	1	-2/3
Higgs/	(H_d^0, H_d^-)	$(\tilde{H}_d^0, \tilde{H}_d^-)$	1	2	-1
higgsino	(H_u^+, H_u^0)	$(\tilde{H}_u^+, \tilde{H}_u^0)$	1	2	1

In models with $m_{A^0} \gg m_Z$, the particles $H^\pm$, A^0 , and H^0 are much heavier than h^0 and nearly decoupled from the low-energy SM particles. Such models are resilient against the EWK precision tests.

The mass spectrum

All MSSM states with the same quantum numbers mix to form the physical mass eigenstates. From the lightest to the heaviest, four spin-1/2 *neutralinos*, $\tilde{\chi}_1^0, \tilde{\chi}_2^0, \tilde{\chi}_3^0$, and $\tilde{\chi}_4^0$, are formed by the mixing of the neutral bino, wino and the two neutral higgsinos. The charged winos and the charged higgsinos mix into four spin-1/2 *charginos*: $\tilde{\chi}_1^\pm$ and $\tilde{\chi}_2^\pm$. Neutralinos and charginos are collectively referred to as *gauginos*.

The 2×2 mixing matrix of the sfermions contains off-diagonal terms proportional to the corresponding Yukawa coupling. Hence, the third generation sfermions show a high mixing while the mixing of the two light generation sfermions are negligible in most of the cases. The resulting mass splitting pulls down the mass of the lightest top squark, allowing the top squark to reasonably be the lightest sfermion.

The gluinos are spin-1/2 particles forming a color octet and therefore cannot mix with any other particle of the MSSM.

Overall, a summary of the mixing of the MSSM sparticles is given as:

- neutralinos: $\tilde{B}^0, \tilde{W}^0, \tilde{H}_u^0, \tilde{H}_d^0 \rightarrow \tilde{\chi}_1^0, \tilde{\chi}_2^0, \tilde{\chi}_3^0, \tilde{\chi}_4^0$,
- charginos: $\tilde{H}^\pm, \tilde{W}^\pm \rightarrow \tilde{\chi}_1^\pm, \tilde{\chi}_2^\pm$,

- top-squark: $\tilde{t}_L, \tilde{t}_R \rightarrow \tilde{t}_1, \tilde{t}_2$,
- bottom-squark: $\tilde{b}_L, \tilde{b}_R \rightarrow \tilde{b}_1, \tilde{b}_2$,
- stau: $\tilde{\tau}_L, \tilde{\tau}_R \rightarrow \tilde{\tau}_1, \tilde{\tau}_2$.

The constrained MSSM

The soft-breaking of SUSY results in models with 105 free parameters. Therefore, interpreting experimental observations within these models is almost impossible. However, within the constrained MSSM (cMSSM), universality assumptions at the Grand Unified Theory (GUT) scale allow the reduction of the number of free parameters to four free parameters, and a sign constrain on an additional parameter:

- m_0 : running mass of all scalar particles at the GUT scale;
- $m_{1/2}$: running mass of all gauginos at the GUT scale;
- A_0 : trilinear coupling of all superpartners;
- $\tan(\beta)$: ratio of the VEV of the two higgs doublets;
- $\text{sign}(\mu)$: sign of the higgsino mass term.

Simplified SUSY models: T2tt

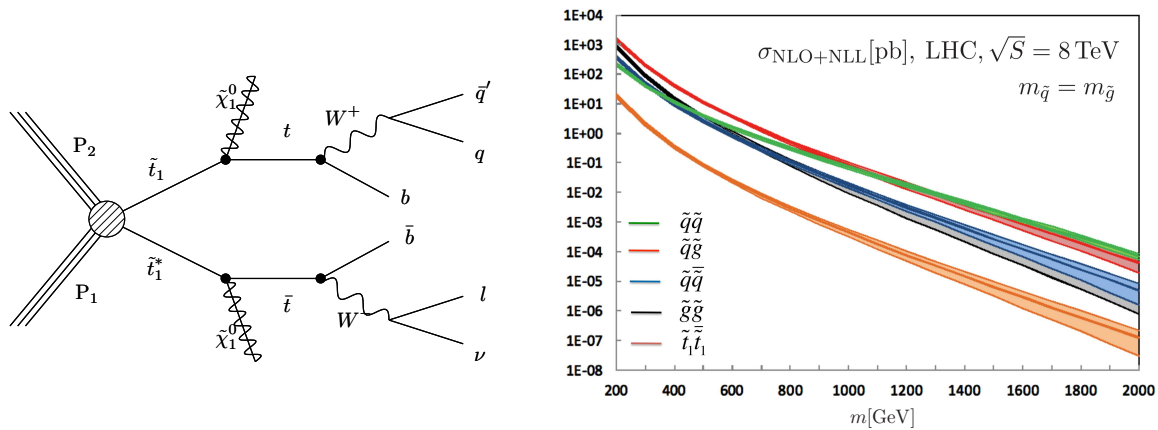


Fig. 2.4: Feynmann diagram of the T2tt simplified mode (left) with a single lepton final state and cross sections of the SUSY particle production at the LHC (right). The signature events in the detector can be seen as events with a single lepton, multiple jets with at least one of them tagged as a b -jet, and a higher missing energy due to R-parity conservation. The topology of the resulting in events heavily depend on the mass difference (ΔM) between the neutralino and the top-squark. The production cross-section of the events steeply falls as shown in the figure on right [23].

The experimental data should be interpreted with large number of free-parameters. Even though constraining these parameters with prior assumptions is useful for a more thoroughly interpretation of a complete theory, it heavily restricts the SUSY-parameter-phase spaces considered by the analyses. An effective theory approach, so-called simplified models, can be employed to loosen these restrictions by focusing on a limited set of SUSY particle production and decay modes, while allowing other parameters to vary freely. Therefore, an analysis interpreted within a simplified model can be used in the interpretation of a fully realized, complete framework (e.g. the phenomenological MSSM, or an example of such a reinterpretation can be found in Ref. [24]).

The highest fermionic loop divergence to the SM Higgs mass comes from the t -quark contributions. A scalar supersymmetric t -quark partner should reside in TeV or below TeV scale to cancel these divergent terms and reduce the fine-tuning to a considerable level. Therefore, the search for supersymmetric top-quark partners are quite appealing to probe new physics in the modern colliders. In order to enable a wider variety of model interpretations, effective simplified models containing t -quarks should be considered rather than in a full model.

In this thesis, the results are interpreted within the T2tt simplified model to search for supersymmetric top partner in the events with single lepton final state. A 100% branching ratio of $\tilde{t} \rightarrow t \tilde{\chi}^0$ is assumed within the T2tt model.

The SUSY process considered within the T2tt model is represented by the Feynman diagram shown in Fig. 2.4. The topology of the resulting in events heavily depends on the mass difference (ΔM) between the neutralino and the top-squark set within the model. For the regions ΔM close to t -quark mass, the decay presents a very similar topology to the t -quark pair production. At high top-squark mass region, the cross-section of the process steeply drops as seen in Fig. 2.4 (right), yet resulting in events with higher missing energy.

The decision of considering events with a single lepton final state (rather than dilepton or 0-lepton final states) is directly related to the t -quark decay chain, and particularly to the W -boson decay. The branching ratio of hadronic decay of W boson is 70%, whereas the leptonic is 30%. While requiring a lepton in the final state reduces the dominant hadronic SM background (particularly in the hadron collider experiments), by allowing a hadronically decaying channel restores the event rate loss due to the branching ratio of W -boson lepton decay.

2.2.3 Searches for Supersymmetry at colliders

As being the most appealing BSM theory, SUSY has been tested through various collider experiments. However, no direct evidence indicating the existence of SUSY models has been observed. Before the start of the LHC, the SUSY searches at the LEP and Tevatron heavily constrained the CMSSM, indicated that squarks and gluinos were most likely to have masses in the 500–800 GeV range. Neutralinos and sleptons were expected to be quite light, with the lightest neutralino and the lightest stau most likely to be found between 100 to 150 GeV [26].

With the discovery of 125 GeV SM compatible Higgs particle, further constraints on the SUSY models have been imposed. The discovered Higgs boson lies in the relatively small mass range favored by MSSM (the constrained Higgs sector of MSSM favors Higgs boson mass to be less than

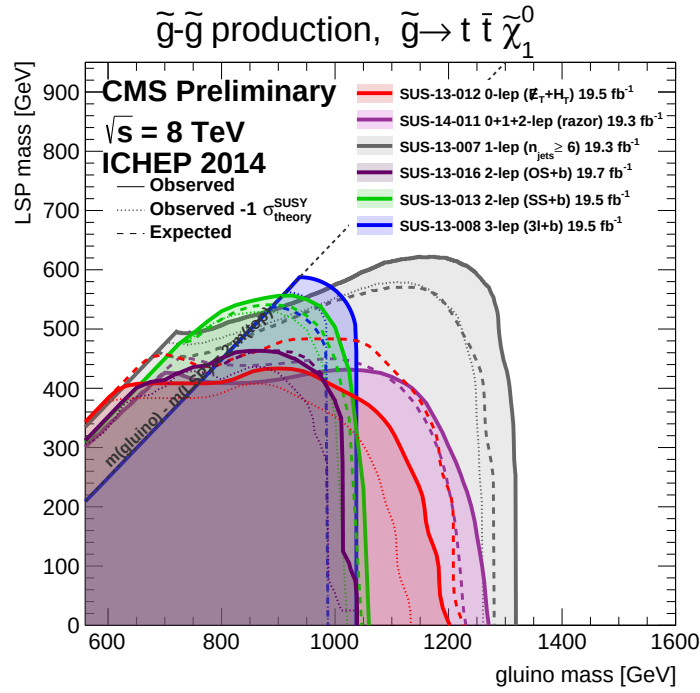


Fig. 2.5: Summary of the gluino pair production searches 8 TeV center of mass energy interpreted within T1tttt simplified model performed by the CMS collaboration [25].

140 GeV). The relatively high mass of the Higgs challenges the constraint model since here the Higgs boson mass is strongly bound by the Z^0 mass at tree level, but it can be lifted up by radiative corrections. However, these corrections cannot be made arbitrarily large without having to accept top-squark masses in the multi-TeV range, and hence a large fine tuning of the electroweak scale. Both CMS and ATLAS collaborations heavily constrain the natural CMSSM SUSY models with the 7 TeV center-of-mass energy.

The current searches in both collaborations are mainly interpreted within the simplified models. The results of the search for the gluino pair-production a center-of-mass energy of 8 TeV interpreted within T1tttt simplified model performed by the CMS collaboration is presented in Fig. 2.5. In Chapter 8, the results for the top-squark pair production is thoroughly discussed.

Experimental Setup

Contents

3.1 Large Hadron Collider	20
3.2 The Compact Muon Solenoid detector	23
3.2.1 Inner tracking detector	25
3.2.2 Electromagnetic calorimeter	27
3.2.3 Hadron calorimeter	28
3.2.4 Forward detector	29
3.2.5 Muon system	30
3.2.6 Trigger and data acquisition	30
3.2.7 Luminosity determination	32
3.2.8 Offline computing	32
3.3 Monte Carlo simulations	32

While the physicists' understanding of our universe expands and changes, for each secret uncovered, a bigger challenge is revealed. Therefore, the tools and techniques that are employed are required to be advanced and adapted as well. Starting from a table-top collider, the modern accelerators can easily enclose a mediocre-size city.

The most advanced particle collider to date, the Large Hadron Collider (LHC), provides six times higher center-of-mass energy and a thirty times larger event rate [27] than its predecessor Tevatron. These outstanding parameters enable the HEP community to further improve the precision of the SM measurements, and provide a capable research ground to search for the rare events proposed by new-physics phenomena discussed in Chapter 2. With this purpose, equally capable detectors were developed to collect and analyze the collision information.

In this chapter, the LHC and the Compact Muon Solenoid (CMS) detector, that are used in the present work, are described.

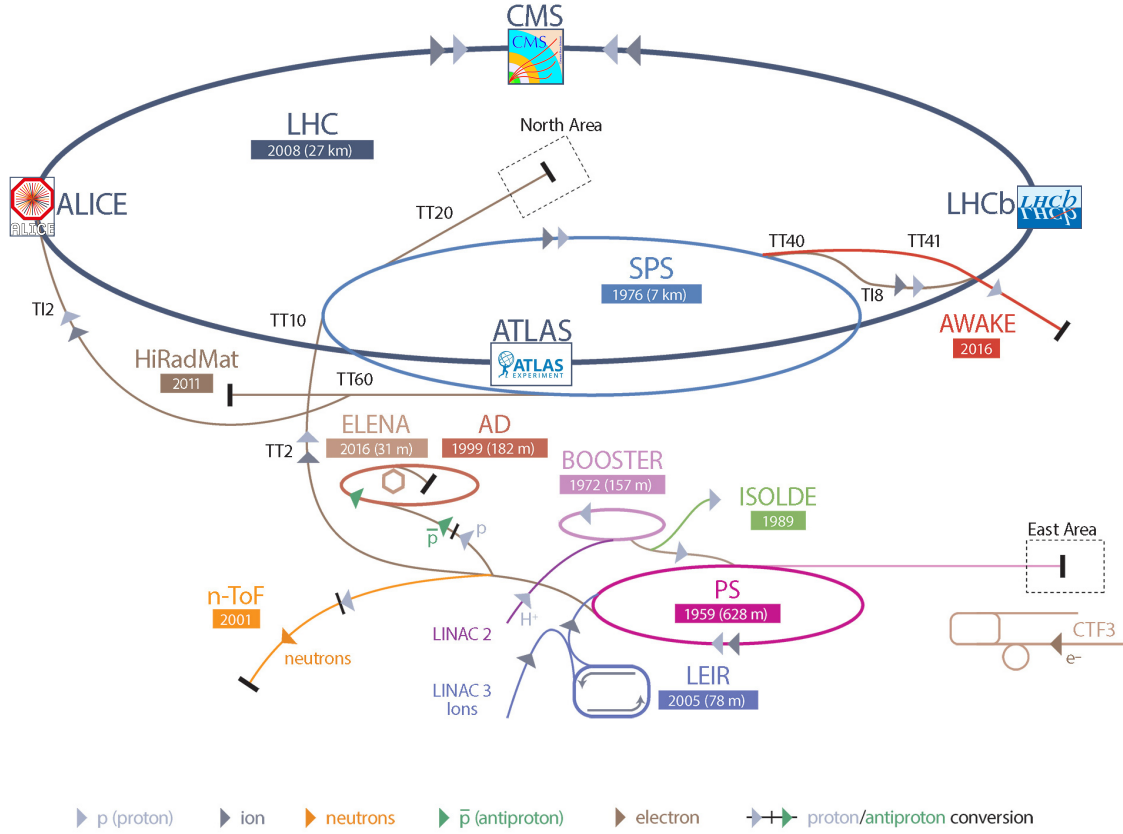


Fig. 3.1: Schematic view of the Large Hadron Collider with its pre-accelerators, detectors occupying the collision points. Starting from LINAC2, the proton beams follow Proton Synchrotron Booster (PSB) - Proton Synchrotron (PS) - Super Proton Synchrotron (SPS) path in the clock-wise direction. The figure is adapted from [29].

3.1 Large Hadron Collider

The European Organization of Nuclear Research (CERN) hosts the Large Hadron Collider since 2000. The LHC [28] is a hadron accelerator and collider complex that is placed in a 26.7 km circumference circular tunnel which crosses the French and Swiss borders, and is located in a depth of 45 m to 175 m from the surface. This colossal machine can accelerate proton beams up to 14 TeV. The particle density delivered by the LHC can be quantified with its *luminosity*. The instantaneous luminosity ($\mathcal{L}$) is estimated using a variety of relevant beam parameters

$$\mathcal{L} = \frac{N_b^2 n_b f_{\text{rev}} \gamma_r}{4\pi \epsilon_n \beta^*} F \quad (3.1)$$

where N_b is the number of particles per bunch, n_b the number of bunches per beam, f_{rev} the revolution frequency, γ_r the relativistic gamma factor, ϵ_n the normalized transverse beam emittance, β_n the beta function, and F accounts for the reduction due to the beam crossing angle. The total

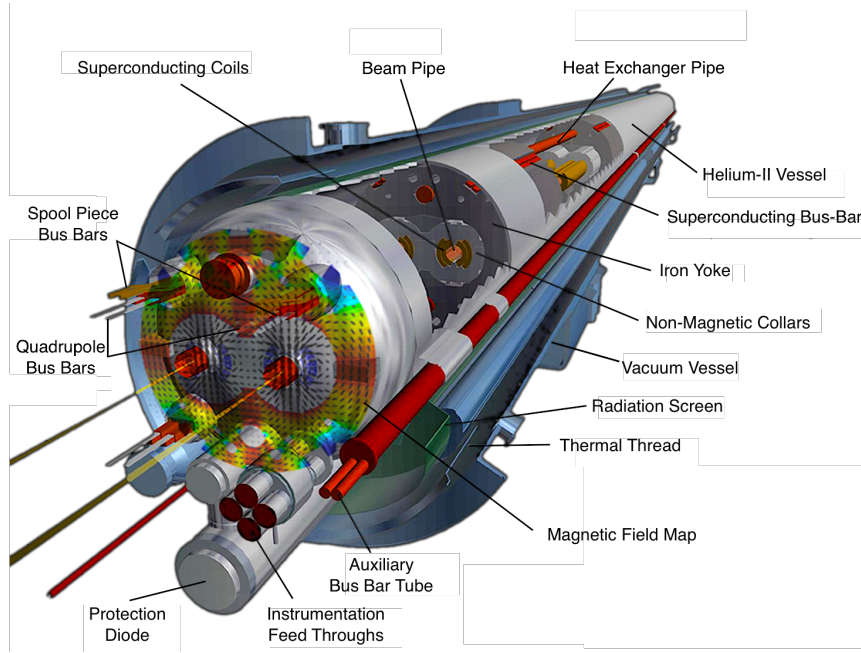


Fig. 3.2: Structural view of an LHC Dipole magnet. The beam pipes carrying beams in opposite directions share the same cold material and magnet for space and cost efficiency. The magnetic field lines are presented with small arrows. The figure is adapted from [30].

(integrated) luminosity can be defined as

$$L = \int \mathcal{L} dt. \quad (3.2)$$

The LHC is designed to collide protons with a luminosity of $10^{34} \text{ cm}^{-2}\text{s}^{-1}$. It is also capable of colliding lead beams (Pb) with the maximum luminosity of $10^{27} \text{ cm}^{-2}\text{s}^{-1}$ and energy of 2.8 TeV per nucleon in order to investigate rare quark-gluon interactions.

The corresponding event rates r at the LHC can be calculated as:

$$r = \mathcal{L} \times \sigma, \quad (3.3)$$

where $\mathcal{L}$ is the instantaneous luminosity and σ is the cross section for the process of interest.

The 14 TeV design energy corresponds to an inelastic proton-proton (pp) cross section of 70 mb, and with the design luminosity of $10^{34} \text{ cm}^{-2}\text{s}^{-1}$, $7 \times 10^7 \text{ Hz}$ interaction rate is expected.

At the energy and luminosity frontier of particle colliders, cutting-edge electronics are combined with state-of-the-art design and advanced-material choice in the construction of the LHC. There are more than 1500 magnets with various functionalities positioned inside the former LEP tunnel in order to reach the indicated specifications. A magnetic field of 8.33 Tesla is required to bend 7 TeV proton beams. The Niobium-titanium (NbTi) alloy superconductors are cooled down to 1.9K to achieve the desired magnetic field. The bending is mainly accomplished by 1232 dipole magnets, and more than 300 quadrupole magnets are employed for focusing the beam. To reduce

cost and to reach maximum performance within the space limitations, a two-in-one design is applied to almost all magnets enabling two opposite direction beam channels benefit from the same cold material and an opposite magnetic flux. The schematic of an LHC dipole magnet is shown in Fig. 3.2.

The proton beams at the LHC are generated by the Linear accelerator 2 (LINAC2) from a bottle of hydrogen gas, and follow Proton Synchrotron Booster (PSB) - Proton Synchrotron (PS) - Super Proton Synchrotron (SPS) path in order. The protons enter the LHC ring with 450 GeV energy, and can be accelerated up to 7 TeV within the LHC. Finally, after two beams are accelerated in opposite directions, they collide inside two general purpose high luminosity detectors, the Compact Muon Solenoid (CMS) [31] and A Toroidal LHC Apparatus (ATLAS) [32]; an ion detector, A Large Ion Collider Experiment (ALICE) [33]; and a fixed target detector for B physics; LHCb [34].

The present analysis uses 19.5 fb^{-1} data collected with the CMS detector in 2012, and the upgrade studies within the thesis are also done readout electronics of the CMS detector. Therefore, a detailed description of the CMS detector is presented in the following section.

3.2 The Compact Muon Solenoid detector

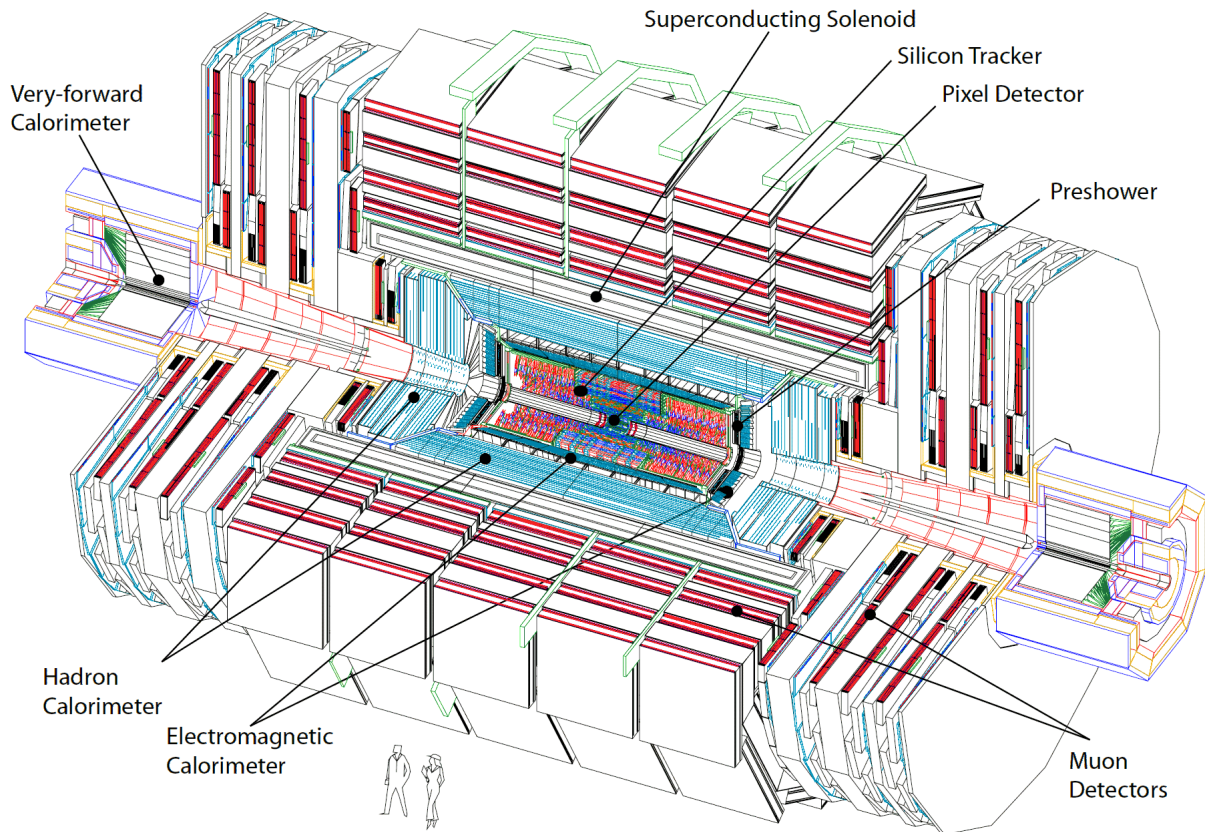


Fig. 3.3: An overview of the CMS Detector [31]

CMS [31, 35, 36] is one of two multi purpose detector located 100m underground, inside an experimental cavern at the Point 5 of the LHC near to the French village Cessy. The event rate delivered by the LHC requires CMS to have low-latency measurements and high-granularity readout which are achieved by fast electronics and a multi-layered, multi-channel detector structure.

The beams are squeezed $16.7\text{ }\mu\text{m}$, and they collide with a crossing angle $285\text{ }\mu\text{m}$ at the interaction point (IP) of CMS. The resulting charged particles are bent with a homogeneous magnetic field of 3.8 T that is produced by a superconducting solenoid coil. Even though only one fifth of the CMS detector's volume is enclosed by the solenoid, the rest of the volume is mostly filled with the gigantic muon system mounted in a surrounding iron return yoke. Outside of the coil, the magnetic field points to the opposite direction with a magnitude of 2 T.

The physics program of CMS comprises making precision measurements for the SM, searches for physics beyond SM, exploring TeV scale physics, understanding the origin of electroweak symmetry breaking and the search for dark matter, and detecting high energy ion collisions.

Such an ambitious physics program can only be achieved with a capable detector, thus requiring CMS to measure particle interactions with high precision and speed. These requirements can

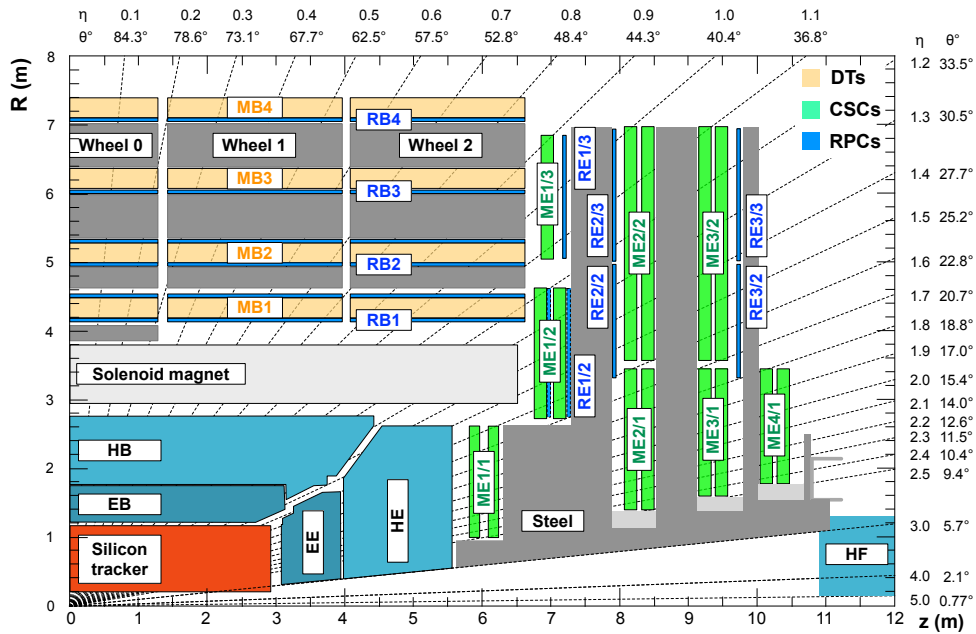


Fig. 3.4: Multi-layered structure of the CMS detector visible from a longitudinal quadrant slice [37]. Starting from the interaction point, particles interact with the tracking system, the electromagnetic and hadronic calorimeters, and the muon system in the transverse plane.

be elaborated as follows:

- Efficient online triggering and good momentum resolution and identification of charged particles.
- Good muon identification and p_T resolution in a wide spatial range, e.g. to achieve dimuon mass resolution,
- High electromagnetic energy resolution, and good photon and electron identification, e.g. to achieve good diphoton and dielectron mass resolution,
- Wide HCAL coverage and hadronic shower reconstruction, e.g. to obtain good dijet mass and missing-transverse-energy resolution.

The CMS detector satisfies these requirements with a multi-layered structure which consists of different detector subsystems. The type, size and location of the detector subsystems are dictated by the physics requirements and radiation tolerance. Table 3.1 gives an overview of the main detector subsystems required for the detection and identification of the SM particles. For the CMS detector, a radius of ~ 1 m is sufficient for the tracking system to measure the secondary-decay vertices, and charged-particle trajectories within an acceptable accuracy inside the 4 T magnetic field. The size of the calorimeters is determined by the characteristic distance to contain the

Table 3.1: The main subsystems required for the detection and identification of the SM particles with characteristic signatures. Even though these systems provide a rather good description of the physics objects on their own, a combined approach is followed in the CMS collaboration for the reconstruction of these particles (cf. Chapter 5)

Particle	Signature	Main subsystem
quarks	jets	ECAL, HCAL (and tracking system)
e, γ	electromagnetic shower	ECAL and tracking system
neutrinos	missing transverse energy	ECAL and HCAL
μ	ionization	Muon absorber and detectors, and tracking system
c, b	secondary decay vertices	vertex and tracker

full shower within the calorimeters. About 20 radiation lengths (X_0) are needed for the electromagnetic calorimeter (ECAL), while a depth of 9 nuclear absorption lengths (λ_I) is required for the hadronic calorimeter (HCAL). Due to increased radiation fluence close to the beam axis, the tracking system and ECAL are centralized perpendicular to the beam direction.

The CMS detector is almost symmetrical around the beam axis, thus the CMS coordinate system is defined in cylindrical coordinates, with the z-axis along the counter-clock-wise beam direction. The azimuthal angle ϕ starts from the direction pointing the center of the LHC circle. Instead of the longitudinal angle θ , a more convenient pseudorapidity (η) coordinate is used. η is defined as,

$$\eta = -\ln \left[\tan \left(\frac{\theta}{2} \right) \right]. \quad (3.4)$$

The pseudorapidity is a special case of *rapidity*, which is a relativistic velocity measure that is Lorentz invariant under longitudinal boosts. Rapidity can be seen as a hyperbolic transformation of the velocity. In the limit where particle masses are negligible, rapidity simplifies to η .

This choice of coordinate is convenient for the CMS detector since particles are expected to be produced with an almost uniform distribution in η axis [27].

After describing selection of the coordinate system, starting from the closest to the IP, all relevant detector subsystems are described in following.

3.2.1 Inner tracking detector

With more than 200 m^2 of silicon surface (with 1440 pixel and 15148 strip detector modules), the CMS tracking detector [38] is the largest silicon tracking system ever built. It consists of a pixel detector and a silicon strip detector, which are enclosed by the corresponding endcaps. The tracking system covers $|\eta| < 2.5$. It provides state-of-the-art measurements of the trajectories of charged particles and secondary-vertex reconstruction.

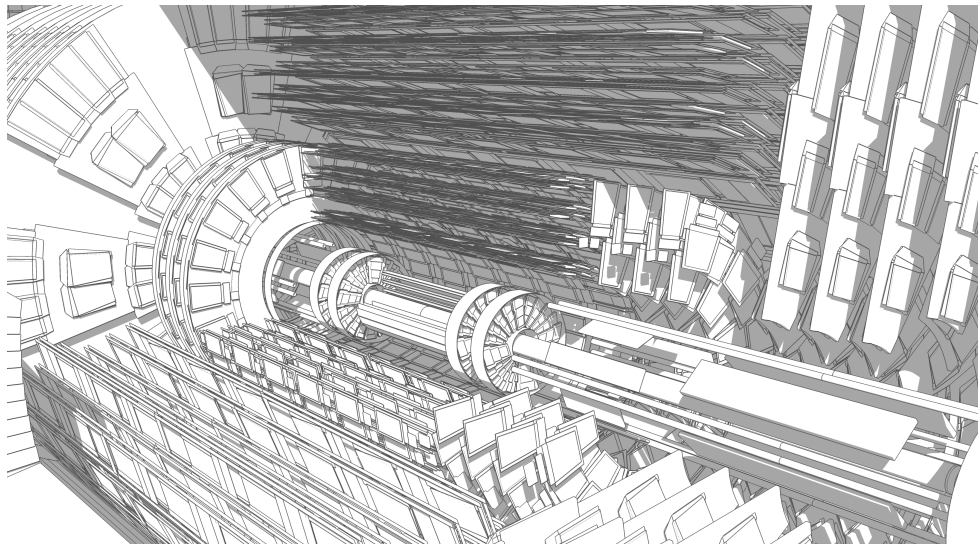


Fig. 3.5: The tracking system of the CMS detector. The pixel detector is positioned at the center of the tracking system and surrounded by the strip detectors [31].

Pixel tracking detector

The pixel tracking detector is the closest subsystem to the interaction point. Its main purpose is to measure the origins of tracks with high precision. This is especially important for tagging jets originated from b quarks, separating prompt electrons from converted photons, and handling pile-up events by reconstructing distinct vertices. Due to high radiation exposure and high flux of the charged particles around the beam axis, radiation hardness of both the sensors and the front-end readout electronics is crucial.

The pixel detector consists of a barrel part of three concentric cylindric layers and two endcap disks on each side. The barrel cylinders are 53 cm long and have radii of 4.4 cm, 7.3 cm, and 10.2 cm. The endcaps are rings with an inner radius of 6 cm and an outer radius of 15 cm. They are positioned along the beam axis at ± 34.5 cm and ± 46.5 cm with respect to the center of the coordinate axis. The barrel consists of 768 pixel modules in total, the endcaps of 672 modules; each pixel has a surface of $100 \times 150 \mu\text{m}^2$. In order to prevent uncovered regions, the blades carrying the pixel modules in the endcaps are arranged in a windmill-like structure as illustrated in Fig. 3.5. Due to the close distance to the interaction region, the pixel tracker is exposed to the highest particle flux in the detector of about 107 Sv at $r = 10$ cm. The expected radiation dose under design conditions for an integrated luminosity of 500 fb^{-1} is estimated to be 840 kGy at $r = 4$ cm and 190 kGy at $r = 11$ cm.

Silicon strip detector

The silicon strip detector encapsulates the pixel detector, and it consists of several subsystems itself: Two cylindric barrels; the tracker inner barrel (TIB), the tracker outer barrel (TOB), and two endcaps on each side; the tracker inner disks (TID) and the tracker endcap (TEC). The TIB has the length of 130 cm, while the TOB covers 220 cm. In the transverse axis, the first TEC disks

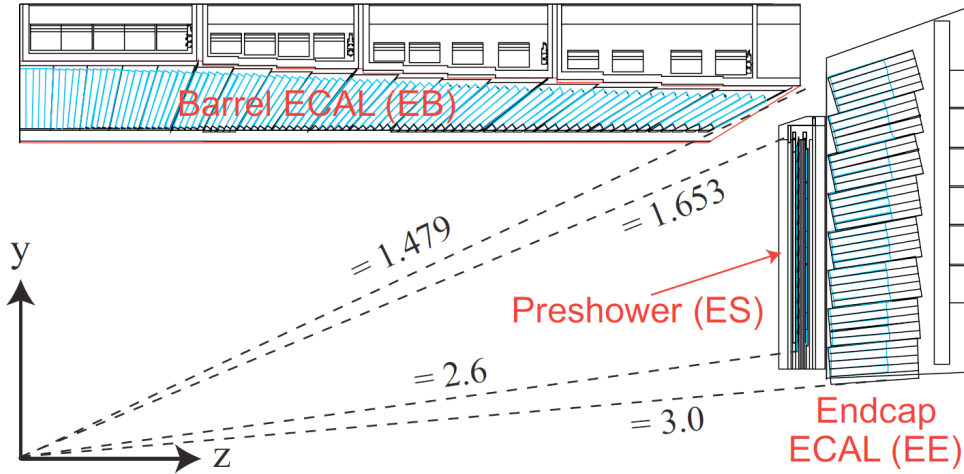


Fig. 3.6: Cross section of an ECAL detector quadrant [31]. The blue rectangles (separated by the thicker black lines) show the PbWO_4 crystals (supercrystals).

have a distance of $z = \pm 120$ cm from the detector center, the outer-most disks are at $z = \pm 280$ cm. The TID is positioned in the gap between the TIB and the TEC. The whole silicon strip tracker comprises 15 400 modules, recording tracks at an operation temperature of -20°C .

The TIB consists of four layers of silicon sensors, of which the inner two are made of so-called stereo modules, i.e. two layers of complimentary modules, providing a two dimensional measurement in ϕ and in z . Similar to the TIB, the first two of the six TOB layers are also made of stereo modules.

The endcaps on each side comprise nine disks (TEC) plus three smaller disks called TID. Stereo modules are used for the first two rings of the TID and the first, second, and fifth ring of the TEC.

3.2.2 Electromagnetic calorimeter

The Electromagnetic Calorimeter (ECAL) [39] is composed of 61200 lead tungstate (PbWO_4) crystals placed in the barrel (EB) and by 7324 crystals in each of the two endcaps (EE). The EB covers a range of $|\eta| < 1.479$, while the EE cover $1.479 < |\eta| < 3.0$. A preshower detector (ES) is installed in front of EE as shown in Fig. 3.6. The high granularity of the ECAL is possible due to the short Moliere radius and fast response time of the crystals.

The same size crystals are grouped together in 5×5 configuration to create ECAL supercrystals used for object reconstruction. This formation is extrapolated to the CMS-HCAL system as well.

Two different photo-detector technologies are used in the Electromagnetic calorimeter, Avalanche photodiodes (APDs) in the barrel and vacuum phototriodes (VPTs) in the endcaps. These detectors are chosen, and positioned in the detector with respect to their radiation hardness.

The crystals in the EB are distributed into 260ϕ and $2 \times 85 \eta$ segments. They are aligned with the η projection by a three degree offset to prevent uncovered inter-regions. Each of them has a $0.0174 \times 0.0174 \eta - \phi$ coverage facing the IP.

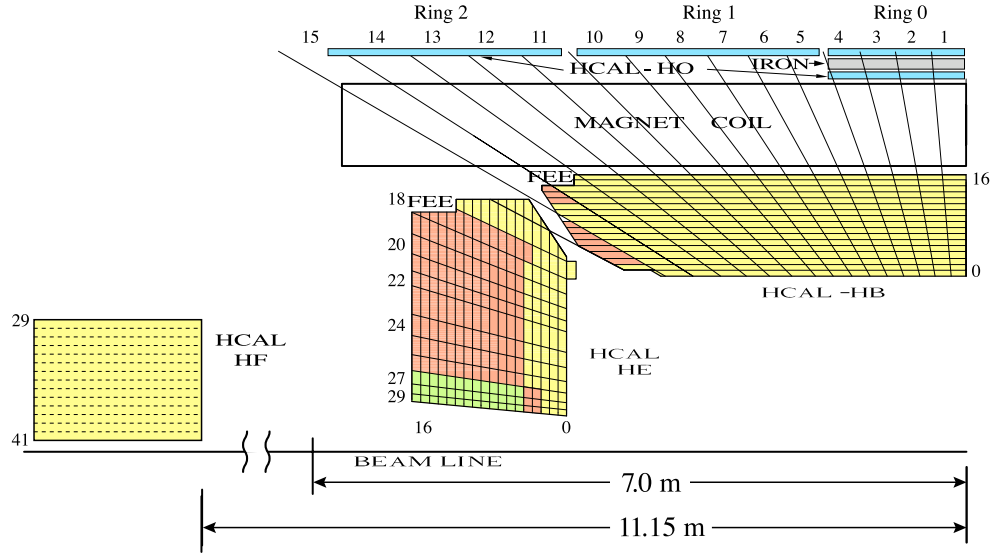


Fig. 3.7: Structure of an HCAL quadrant. From low to high η the barrel, endcap and forward calorimeters are shown. The HO is the only calorimeter system outside of the solenoid [31].

The EE system is constructed similar to the EB. The crystals forming a single supercrystal are aligned in the same direction unlike the EB crystals as seen in Fig. 3.6.

The preshower detector is a sampling calorimeter placed in front of the EE. and covers $1.653 < |\eta| < 2.6$ region. It has two layers: Lead radiators detect the starting point of electromagnetic showers, and strip silicon detectors measure energy.

3.2.3 Hadron calorimeter

The CMS Hadron Calorimeter (HCAL) [40] is radially placed between the ECAL and the magnet, and covers an azimuthal range of $|\eta| < 5.2$.

The structural design of the HCAL can be divided into four parts: barrel (HB), endcap (HB), outer (HO), and forward calorimeter (HF). The HCAL is employed to measure hadron jets, and to estimate missing transverse energy due to neutrinos and possibly new physics processes. The HO has a supporting role which is to ensure the measurement of hadrons that partially escape from the detection of the HE and the HB. HF is to take measurements at the forward region of CMS at $3.0 < |\eta| < 5.0$. A schematic r - z view of a quadrant of the HCAL is shown in Fig. 3.7. The locations of the four HCAL detector subsystems along with the front-end electronic (FEE) modules, are indicated.

Barrel calorimeter

The HB is a sampling calorimeter made of plastic-scintillator tiles inserted into a brass absorber ($\lambda_I = 16.42$ cm, $X_0 = 1.49$ cm). It covers the pseudorapidity range $|\eta| < 1.3$, and consists of 16 layers of absorber plates and 17 layers of scintillator tiles. The plastic-scintillator cell-size

is $0.087 \times 0.087 (\Delta\eta \times \Delta\phi)$. The arrangement of the scintillator layers prevents projective dead areas.

The scintillation light is collected by wavelength-shifting fibers embedded into the tiles, and then channeled into photodetectors via clear fibers. Hybrid photodiodes (HPD) are used in the HCAL as photodetectors since they provide a high gain, and can operate under intense axial magnetic fields contrary to phototubes.

Endcap calorimeter

The HE is placed in the region $1.3 < |\eta| < 3.0$ and is made of 17 layers of absorber and 18 layers of scintillators. It is constructed similarly to the HB: only the plastic-scintillator cell size differs from the one of HB for $|\eta| > 1.6$, where it is about $\sim 0.17 \times 0.17 (\Delta\eta \times \Delta\phi)$. The geometry of the absorber layers is driven by the need to minimize the cracks between HB and HE.

The longitudinal segmentation of the readout increases in higher pseudorapidity ranges (at the transition region between HB and HE as well as at higher $|\eta|$) as shown in Fig. 3.7.

Outer calorimeter

As the only calorimeter subsystem placed outside the solenoid, the HO detects the tail of high-energy and/or late developing showers, thus functioning as a *tail catcher*. Since the central region is less deep in terms of nuclear absorption length ($\sim 7 \lambda_I$) seen from the nominal IP, for $|\eta| < 0.35$ the HO consists of two layers of plastic scintillator tiles with an additional 19.5-cm-thick iron absorber. Therefore, it compensates the relatively low absorption material budget in the central region, and increases it above $10 \lambda_I$.

Forward calorimeter

The forward calorimeter (HF) extends the coverage of the HCAL in pseudorapidity up to $|\eta| = 5.0$. It is composed of two steel cylinders (one each side) with the front face located 11.2 m away from the IP. The beam pipe passes through a cylindrical hole of radius 12.5 cm. The HF uses a Cherenkov-based radiation-hard technology in order to cope with the high radiation fluency in the forward region. The sensitive part of the detector is composed of quartz fibers bundled to form $0.175 \times 0.175 (\Delta\eta \times \Delta\phi)$ towers. Photomultiplier tubes (PMT) are used to convert the collected light into an electrical signal.

3.2.4 Forward detector

Detection of particles at the very forward region of the CMS detector is maintained by the CASTOR (Centrauro And Strange Object Research) detector. CASTOR is composed of quartz tungsten plates which ensures continuous operation of the detector even under high radiation fluence. The CASTOR detector covers $5.2 < |\eta| < 6.6$.

3.2.5 Muon system

Muons are minimal ionizing over a large energy range typical for the LHC and penetrate through the whole CMS detector. The muon system [41] is the outermost detector not only names the CMS detector, but also gives 80% of its volume. The iron yoke surrounding the muon system returns the magnetic flux from the solenoid, resulting a 2 T magnetic field. The muon system is composed of four subsystems:

Drift tube system

The CMS barrel muon system is composed of 250 drift chambers placed in 4 concentric cylindrical layers. The drift tube system measures electromagnetic cascades coming with muons and can be used as tracking detectors.

Cathode strip chamber system

There are 468 cathode strip chambers (CSC) aligned in CMS Endcap Muon system that covers an η range of 0.9 to 2.4. Built as multiwire proportional chambers, CSCs measure the radial position of muons.

Resistive plate chamber system

The Resistive Plate Chambers provide a good spatial and timing resolution for muons $|\eta| < 1.6$ in a short time period (much less than < 25 ns). The fast measuring time enables RPC for a similarly fast dedicated muon trigger. It is designed as a parallel plate gaseous detector.

Optical alignment system

Due to constructional tolerances, magnetic field distortions and time dependent deformations, the muon chambers and the central tracker in the CMS may be misaligned. However, a precise measurement of muons requires a flawless alignment of the order of $\sim 100 \mu\text{m}$. To achieve this precision, each subsystem is positioned using an optical alignment system, and the measurements are calibrated using the gathered information.

3.2.6 Trigger and data acquisition

The CMS detector has a data output rate of 40.078 MHz per bunch crossing (24.95 ns) [27, 31], which is impossible to store with existing technologies. Therefore, a two-step triggering system is required to reduce data to a manageable size. The first step is Level-1 Trigger (L1) and the second level is the High Level Trigger (HLT).

Level-1 trigger and readout electronics

The aim of the CMS Level-1 Trigger [27, 31] system is to achieve fast data reduction and transfer data to higher trigger system with high reliability. This is accomplished by mostly subsystem specific systems. The Level-1 Trigger is composed of fast electronics which allow to make individual decision as low as sub ns.

Two kinds of electronics technologies are employed with this purpose, ASICs (Application Specific Integrated Circuits) and FPGAs (Field Programmable Gate Arrays). ASICs are mainly radiation hard, and are designed to accomplish a hardwired specific task, hence they are more reliable and faster comparing to other solutions. However, since they are designed (and produced) with specific purposes, ASICs are expensive to replace, upgrade or debug. FPGAs, multi purpose digital integrated circuits, are in this regard are more flexible, but prone to radiation damage due to logic block density and necessity of internal storage units. Currently, the CMS trigger and data acquisition systems use both FPGA and ASIC systems together. While ASICs are usually installed behind detectors itself, the FPGAs are placed in control rooms close to detectors.

As mentioned earlier, the L1 trigger is specialized on local decision instances for each sub-detector. Such instances are built locally for each readout channel and later combined to regional trigger primitives.

Calorimeter Trigger: The Calorimeter Trigger uses the information obtained from the ECAL and HCAL, and resolves the transverse energy, missing transverse energy, jets and jet counts. It also measures the timing of the events (which should be in the same bunch crossing), and synchronizes data gathered from different calorimeters.

Muon Trigger: The main purpose of this trigger is to match and connect different segments of the muon system and to calculate tracks of the muons. It also provides a good momentum and timing resolution of particles.

Global Trigger: The global trigger has a five-level structure: Input, logic, decision, distribution and read-out. The global trigger decides whether an event will be accepted or not using the information from all sub-detector trigger systems. After the decision, the *L1Accept* signal is distributed, and accepted events are sent to the HLT with an output rate of approximately 30 kHz (with a maximum of 100 kHz).

High-level trigger and data acquisition

Main purpose of the HLT and Data Acquisition is to further reduce data by filtering events. The event filtering has two main stages: to reconstruct physics object and to mark events with interesting features. The HLT requires to read events at the pace of the L1-Trigger output and, thus, requires massive parallelism with a huge amount of computer power. The whole process of the HLT is maintained by a computer farm with more than 9000 processor cores working at high frequencies. The output of the HLT is 200 Hz, which corresponds to ~ 350 MB/s, and with about 100 days of running, CMS is expected to collect more than 3 PB of data per year.

3.2.7 Luminosity determination

The instantaneous and total (integrated) luminosity is one of the most important parameters that are required to be estimated. In the LHC, this measurement is obtained using multiple specific and non-specific detectors. Particularly in the CMS detector, the most precise measurement ($\sigma = 2.5\%$ syst. + 0.5% stat.) is obtained using the pixel tracking and HF subsystems. In order to achieve this precision, The Pixel Cluster Counting (PCC) [42] method exploiting the pixel tracking information is employed. The studies and cross-checks are done using the luminosity information gathered from the HF subsystem.

3.2.8 Offline computing

Just like the detector, the CMS Computing Model has a multi-layered structure. Various data formats are used in the CMS computing model. Type of data tiers and formats used in the CMS computing model can be listed as:

- *DAQ RAW* contains information collected from ASICs and FPGAs with L1 trigger results.
- *RAW* is composed of reduced data coming from computer farms and contains L1 and HLT selection information. These data are reconstructed by the event reconstruction algorithms.
- *RECO (Reconstructed Data)* contains information of the reconstructed objects, hits and clusters.
- *AOD (Analysis Object Data)* is the reduced version of RECO data. It only includes information required for the physics analysis.
- *GEN* is generated Monte Carlo simulated events without detector simulation.

Computing at CMS has four hierarchic tier levels. The raw data are stored and distributed at Tier-0 positioned at the CERN site. Also, first reconstruction steps are applied at this tier. Tier-1 operates the reconstruction, skimming and calibration steps, and also provides a second secure mirror of the raw data. Finally, Tier-2 and Tier-3 provide local services and global grid distribution of the reconstructed data. Moreover, they operate the overall Monte Carlo sample generation for the experiment.

3.3 Monte Carlo simulations

Monte Carlo (MC) generators are widely employed in HEP collider experiments to provide a modeling of the underlying physics processes for finite event yields. MC samples are generated on event-by-event basis: the events are sampled with the estimated cross-section for given model of interest. The generator is fed with the parton density information to simulate the hard parton-level process and to calculate the corresponding matrix elements (ME). Afterwards, parton shower libraries are used to model ISR and FSR on these processes. An infrared cutoff on the parton showers is required to ensure the validity of the perturbation theory. Hadrons originating

from the non-perturbative region are simulated using hadronization models. The following MC generators are used in this thesis:

PYTHIA: PYTHIA (6.4) [43] is a general-purpose MC event generator that simulates processes for SM and BSM. For the ME estimation, the analytical leading-order calculations are combined with phenomenological higher-order approximations. The hadronization is based on the Lund String model which is described in Ref. [44]. In the present analysis, the PYTHIA event generator is used with a set of parameters referred to as Z^{2*} tune [45–47].

MADGRAPH: MADGRAPH (5.1) [48] is a generator for matrix elements on tree level. Parton showering (PS) and hadronization steps are not implemented in MADGRAPH, therefore it is interfaced with PYTHIA. This interfacing may lead to a double counting due to radiation generated by ME and parton showering. An ME-PS matching scale is introduced to prevent such cases.

POWHEG: Starting from hardest processes, POWHEG [49–51] provides NLO ME calculations. Similar to MADGRAPH, POWHEG is lacking PS and hadronization, therefore it is also interfaced with PS and hadronization capable generators.

TAUOLA: The τ leptons with polarization properties can be simulated with the dedicated TAUOLA [52] package.

The MC simulated events are independent of the detector topology, and unable to provide a reference for the recorded data. Therefore, the particles' interaction with the detector is required to be calculated. With the GEANT4 software framework, the detector topology as well as the material budget, the magnetic and electric fields, the intercommunication network structure, and possible electronics noise and cross-talks can be simulated to provide interactions with the generated particles. The CMS collaboration follows two different approaches for the detector interaction simulation of the generated particles. In addition to an overall GEANT4 CMS detector simulation, a faster parametrized software implementation referred to as FastSim [53] is provided. The FastSim simulations are validated with benchmark studies and dedicated test-beams, and compared with the GEANT4 CMS detector simulation.

For a generic collider experiment a similar simulation environment to the FastSim is provided by the DELPHES3 [54] software framework. The software is aimed to be employed in the phenomenological studies, where an advanced detector simulation is not required. The simulation framework includes magnetic field, generic electromagnetic and hadron calorimeters, a generic muon system, and a flexible reconstruction sequence.

The CMS HCAL Front-end Readout Control System Upgrade

Contents

4.1	μTCA	36
4.1.1	CMS HCAL μ TCA	40
4.2	The control system structure	42
4.2.1	CCM server, TCDS, TTC and AMC13	42
4.2.2	ngFEC-FC7	45
4.2.3	Front-end crate	53
4.3	System test	55
4.3.1	ngFEC test	55
4.3.2	Control system test	56
4.4	Outlook	58

The CMS detector front-end electronics operates under a strong magnetic field and is exposed to a high level of radiation fluence due to high-energy-hadron collisions. A direct intervention on the CMS on-detector electronics system during the operation is not possible without opening the detector, which requires an interruption of the LHC machine operation and data taking for several months. Therefore, a reliable remote control and diagnosis system is crucial for the operation of the CMS apparatus. The front-end control system is responsible for remotely distributing the LHC clock and other fast control signals to the digitizer cards, and providing access to the I²C control buses for initialization of the digitizer registers, and the readout of temperature and voltage sensors of the front-end detector readout electronics.

As a consequence of the CMS detector upgrades, the CMS HCAL readout electronics are required to be capable of handling higher data-transfer rates due to the increased segmentation of the detector and the addition of timing information. Moreover, the CMS back-end electronic

infrastructure, located in a radiation protected area, was relying heavily on the outdated VME standard [55]. The VME standard is unable to cope with the increased data-transfer rates without external communication lines and lacks some basic functionalities, i.e. hot-swapping, cooling management, modern connectivity options etc. Therefore, the CMS Hadronic Calorimeter (HCAL) electronics infrastructure is upgraded [56] to μ TCA [57] and improving (and in most cases redesigning) the accompanying modules, software and firmware features, reliability, and performance. To accompany this upgrade, a completely new module, the next generation Front-End Controller (ngFEC), is introduced. It simplifies the former front-end controller by transmitting fast and slow control signals over the existing optical fibers and eliminating the additional twisted pair cable connections to the front-end crates. It reduces the initialization time despite the additional channels by driving all control buses simultaneously, provides redundant paths to adjacent front-end crates, and the ability to reprogram FPGAs on the front-end modules in situ.

In this chapter, an introduction to μ TCA is given, which is then followed by the description of the software and hardware modules deployed in the readout-electronics control structure. A special emphasis is put on the ngFEC firmware development and overall system test since these were carried out by the author as main contribution within the work of this thesis to the CMS HCAL electronics upgrade, and discussed in detail in Sections 4.2 and 4.3.

4.1 μ TCA

Micro Telecommunications Computing Architecture (μ TCA) [57] is an open standard, developed for building next-generation high-performance telecommunication and computer systems in a small form factor. It is based on and complimentary to the Advanced TCA architecture. With respect to ATCA, μ TCA is optimized for lower cost, and is more flexible in terms of availability requirements.

Advanced Mezzanine Cards (AMCs) are the core of μ TCA's processing and IO functionality. Power management, cooling and diagnostic control of AMCs are handled by the μ TCA Control Hub (MCH). μ TCA crates have a dual-star topology as shown in Fig. 4.1, which allows MCHs (positioned in the central slots) to communicate with each AMC in the crate. To accomplish this task, a specific module called MicroTCA Carrier Management Controller (MCMC) is placed on the MCH, which communicates with module management controllers (MMCs) on the AMCs and with enhanced MMCs (EMMCs) on power modules (PMs) and cooling units (CUs). This communication is established via Intelligent Platform Management Interface (IPMI). IPMI can use an I²C based intelligent platform management bus (IPMB), or *external* communication channels like ethernet. The MCH can also act like a switch enabling the AMCs to communicate with each other as well. In addition to these baseline specifications, the μ TCA standard has five different versions differentiating with respect to the requirements of the usage scenarios:

- MicroTCA (MTCA.0): This defines the baseline standard of μ TCA, mainly used in indoor applications. All other versions of μ TCA standard are based on MTCA.0.

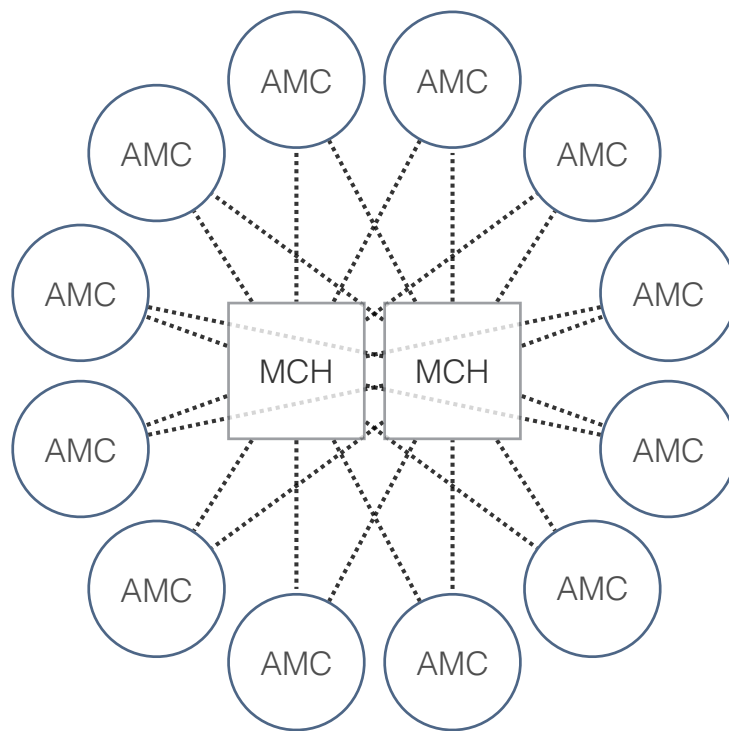


Fig. 4.1: Dual-star topology of the μ TCA crate: circles correspond to 12 AMC slots in the crate and the squares shows the central hub connections. This topology of μ TCA crates enables to have a redundant management of the crate. However, due to precise timing requirements and the necessity of a module that has a fast connection to each of the AMCs in the crate to create a collection network for the data path, the second MCH slot is occupied with the AMC13 module in the CMS HCAL μ TCA infrastructure.

- Air-Cooled Rugged MicroTCA (MTCA.1): It is the ruggedized version of the MTCA.0 for exterior and interior usage scenarios where the effect of vibration is important.
- Hardened Air-Cooled MicroTCA (MTCA.2): Mainly used by applications on vehicles (i.e. on the transportation market), it is the reinforced-ruggedized implementation for harsher usage scenarios.
- Hardened Conduction Cooled MicroTCA (MTCA.3): This version of the standard is the most robust implementation that is designed for extreme conditions of aerospace and military applications.
- MicroTCA Enhancements for Rear I/O and Precision Timing (MTCA.4): This standard is mainly developed by the HEP community to provide an additional input output area and corresponding rear transition module (MicroRTM).

The decision to adopt the μ TCA standard can easily be justified by comparing μ TCA standard over its predecessors. It provides advancements over legacy standards, such as: better power and cooling management, small form factor, hot swapping, and faster backplane communications.

Due to the high granularity of the CMS apparatus and DAQ system, the all-in-one-approach in the MTCA.4 standard is not suitable for the CMS electronics infrastructure. Therefore, the CMS back-end electronics infrastructure is operated with the MTCA.0 standard. System elements of μ TCA given by the MTCA.0 standard can be listed as:

AMCs

MTCA.0 allows use of AMCs without further modification. μ TCA supports six different sizes of AMCs, from the smallest single compact-size module ($73.8 \times 13.88 \times 181.5$ mm) to the biggest double full-size module ($148.8 \times 28.95 \times 181.5$ mm)

μ TCA carrier

The carrier provides the requirements for the AMC standard including power delivery, interconnections, and IPMI management for twelve AMC boards (unlike eight boards supported by ATCA standard). These requirements are mainly handled by an MCH board. Therefore, the MCH boards become the single-point-of-failures, and in order to prevent this, a redundant MCH can be added to the system.

Backplane and connectors

The μ TCA backplane connects twelve AMC connectors with the MCH and provides high speed (over 12 Gb/s) serial communication.

Mechanical infrastructure/ subrack

The μ TCA subrack holds the AMCs, MCH, PMs, CUs, and a backplane that connects these modules. Following the HEP convention, in this thesis, the term *subrack* is referred to as *crate*.

Cooling and thermal subsystem

The process capability of a μ TCA crate is expected to be really high. As a consequence, it has a high-power consumption density. A cooling system consisting of various filters and fans is necessary.

Control and management infrastructure

Each AMC connector enables MCMC module of MCH to obtain *presence detect*, *enable*, *geographic address* signals for the given AMC slot. There are also two IPMB, IPMB-0 and IPMB-L, connections to AMCs, PMs and CUs. Three clock signals (TCLKA, TCLKB, FCLKA) are distributed to each AMC in order to provide an overall crate synchronization. This topology allows a *shelf manager* software to control and communicate with the modules via MCHs through an ethernet connection. A shelf manager can monitor the voltage and temperature sensors, and control the fan levels of

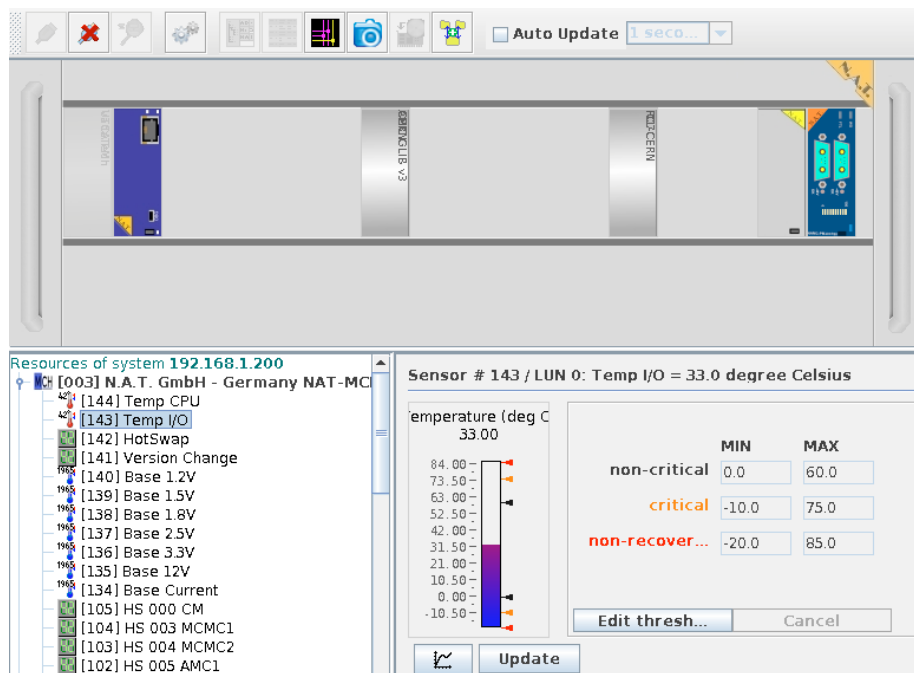


Fig. 4.2: The Natview shelf manager provides monitoring and control functionality of the μ TCA crate. The screenshot shows the graphical user interface of the Natview. The modules and corresponding sensors are detected and *warning* alarm levels of each sensor can be set through the interface. Cooling levels can be adjusted to response these alarm levels in a μ TCA crate.

CUs. Figure 4.2 shows the shelf manager software, Natview, screenshot of the DESY-CMS- μ TCA test-stand crate.

Interconnect fabric infrastructure

In an μ TCA crate, an interconnect fabric provides the main connectivity among the AMCs. This interconnect consists of a central switch and a number of high speed serial lanes to each AMC slot. Lines on μ TCA are differential high speed SerDes interconnects, with typical bandwidth capability of at least 3.125 Gb/s in each direction. Although some μ TCA implementations using inexpensive connectors and backplanes run these lanes at slower rates such as 1 Gb/s, many μ TCA crates permit these bit rates to increase to beyond 12 Gb/s per Lane.

μ TCA is compatible with various formats depending on the AMCs communication protocol and implemented on the MCHs fabric. Supported formats are specified in the AMC standard, including Peripheral Component Interconnect (PCI) Express/Advanced Switching (AMC.1), Ethernet (AMC.2), Storage Interfaces (AMC.3), and RapidIO (AMC.4). Future additions, as well as custom or proprietary backplane protocols, are also possible the supported protocols.

The μ TCA crate controlled by a single MCH forms a a star network, where MCH is the central hub. With two MCHs, the formation becomes a dual star topology as shown in the 4.1. Within this

formation, μ TCA provides seven fabric channels (Fabric A-G) to each AMC. These fabric channels consist of two ports (lanes).

Power infrastructure

One important function of the μ TCA Carrier is to supply and control the power to the AMC. The AMC standard specifies a 12 V main payload power provided for each AMC. Power from the input supply is converted to 12 V to provide radial payload power to each AMC by the μ TCA PMs. The power subsystem also supplies 3.3 V management power for AMCs. The power control logic on the PM performs sequencing, protection, and isolation functions. The power subsystem is controlled via the carrier manager which performs power budgeting to ensure adequate power is available prior to enabling power channels.

The power subsystem consists of one or more PM. Each PM is responsible for converting the input supply that arrives at an input power connector on its face plate (either AC or DC) to the individual branches of 12 V payload power and 3.3 V Management Power needed to run the AMCs, MCHs, and CUs. The μ TCA standard can manage up to four PMs if extra input or redundancy is required.

PMs also include the supervisory functions necessary to manage the power subsystem. They have circuitry to detect the presence of AMCs, MCHs, and CUs, and to energize individual power branches. PMs also monitor the power quality of each branch and protect against overload. If a redundant power module is configured, it will automatically take over the power channel responsibilities of a failed Primary Power Module.

Test infrastructure

An optional JTAG Switch Module (JSM) supports serial scan testing of complete μ TCA crates, as well as their individual elements. This testing is most often carried out during manufacturing tests in a factory setting, but μ TCA can also support the use of these JTAG test capabilities in active systems in the field. JSM functions may be located in special slots on the backplane, or integrated into other modules.

4.1.1 CMS HCAL μ TCA

The ngFEC crate benefits highly from the advancements introduced with the μ TCA standard. The hot swapping feature ensures a continuous operation of the crate while replacing AMCs. The cooling system can be automatically adjusted in case of a sudden temperature change with our custom-developed crate control system, and the crate safely shuts down itself in case of an electrical failure. Overall the system is less susceptible to external interruptions, since most of the communication between the modules is established via backplane connections. Each crate is monitored and controlled by a central WinCC server reinforced with a custom interface. WinCC is a Supervisory Control and Data Acquisition (SCADA) system developed by Siemens. It aims

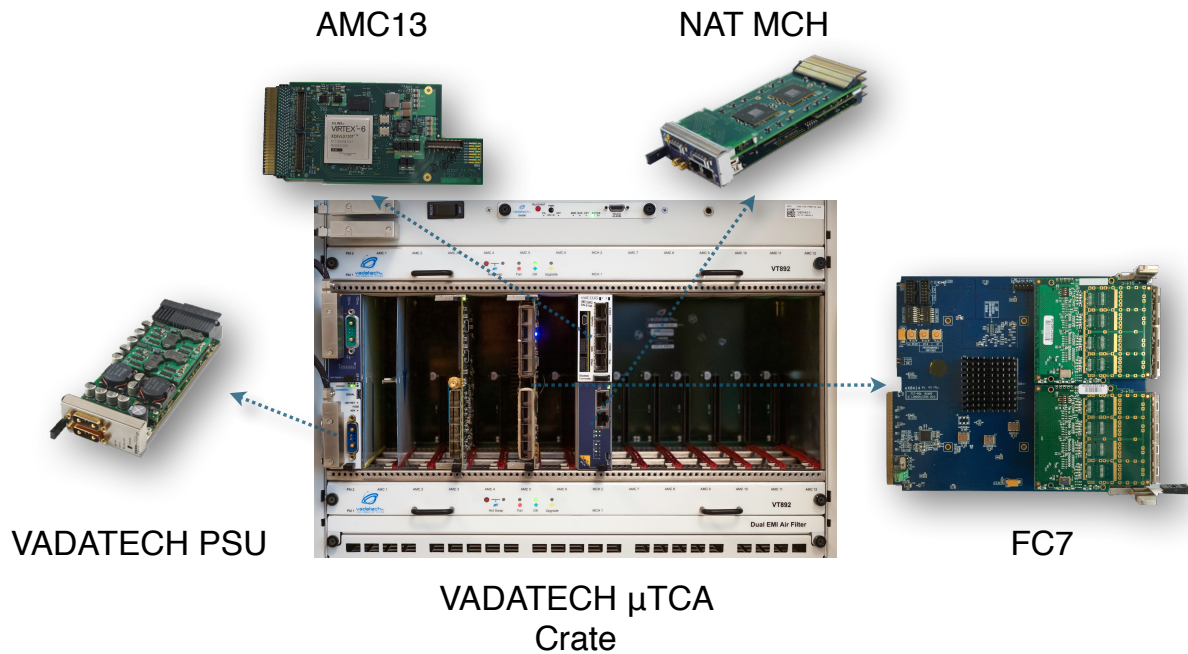


Fig. 4.3: The control system components connected to the ngFEC crate. The typical system would consist of an AMC13, an MCH, two redundant power supplies and twelve ngFEC-FC7 boards.

to provide a control architecture for monitoring automated processes and is a hardware-control-system layer that is purely software, which usually is deployed for directly controlling complex industrial hardware systems. The diagnostic data received from the back-end modules are stored in the WinCC database and constantly monitored. Warning and error signals are generated in case of an unexpected event for immediate action depending on the diagnostic data stored by the WinCC server.

An ngFEC μ TCA crate can hold twelve Advanced Mezzanine Cards (AMC), and it supports a dual star topology where two central hubs are connected to all other modules as shown in Fig. 4.1 through the backplane. The central slots are occupied by a MicroTCA Control Hub (MCH) and an AMC13 board. While the MCH board provides the crate management and the communication with the CCM Server by providing gigabit ethernet connectivity to each AMC, the AMC13 recovers the LHC clock and fast commands, and distributes them through the backplane of the ngFEC μ TCA crate.

4.2 The control system structure

The next generation Front-End Controller (ngFEC) crate connects and organizes the control path between the front-end readout modules and the CMS control systems. It also distributes the LHC Clock (40.0788 MHz) to the front-end readout system. A typical ngFEC μ TCA crate is equipped with one AMC13 board [58], one μ TCA carrier hub (MCH), two redundant power supplies and up to twelve FMC Carrier 7 (FC7) boards [59]. Figure 4.3 shows positions of these modules in the ngFEC μ TCA crate.

The slow-control commands formed by the Crate Control Module server (CCM Server) are sent to the ngFEC-FC7s via ethernet connection through the MCH, while the fast commands are received from the Trigger (Timing) Control Distribution System (TCDS) and distributed to the ngFEC-FC7s through the backplane by the AMC13 board.

The CERN-developed Gigabit Transmission (GBT) [60] protocol is used to send fast and slow control commands to the front-end modules. The responses are stored in the dual-port block RAMs in the Kintex[®]-7 (XC7K420T) [61] FPGA positioned on the ngFEC-FC7 board, and then read, analyzed and transferred to the databases or higher-level control softwares by the CCM server. The general layout of the CMS ngFEC system, starting from the Trigger Timing and Control (TTC) and AMC13 to the front-end modules, is shown in Fig. 4.4.

The control-system modules are positioned both in radiation safe and in radiation hard (closer to the detector) environments. Whereas the back-end systems are hosted by high-level server computers (the CCM Server) and FPGAs (ngFEC-FC7, TTC), front-end (near detector) modules should consist of only radiation hard or tolerant FPGAs and ASICs. All the modules and communication paths that is shown in Fig. 4.4 are described in the following sections.

4.2.1 CCM server, TCDS, TTC and AMC13

CCM server

The CCM Server is the low-level control software which is directly connected to higher-level central DAQ and diagnostic databases, and the ngFEC crate. The CCM Server sets the configuration parameters and enables the front-end modules. It also reads the debugging and the sensor information (temperature, voltage etc.) from the front-end modules and drives the manual fast controls like reset signals. It has a client-server software structure, which enables multiple end-users to access simultaneously to the slow and manual fast control system. The server is connected to the ngFEC μ TCA crates directly. The CCM server, developed with C++11, is a part of the CMS xDAQ software package. An HTML and command line interface is provided for the CCM Server clients.

The CCM Server communicates with the WinCC server via the Distributed Information Management (DIM) protocol and the RBX manager via JSON-RPC. Both the WinCC server and the RBX manager use an Oracle database to hold setting parameters for the front-end modules and store the diagnostic information. In order to control the silicon photo multiplier (SiPM) control board and Peltier cooling system in the front-end crate, it gets the parameters from the WinCC

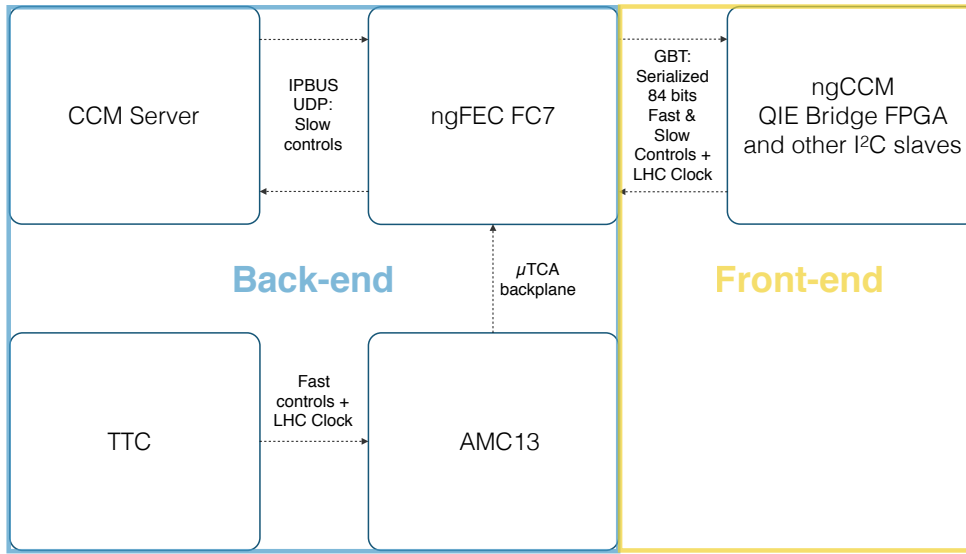


Fig. 4.4: The general layout of the CMS HCAL ngFEC system: The CCM Server sends slow and manual fast controls as UDP packets to the ngFEC-FC7 via IPBUS protocol. Timing trigger control (TTC) signals distributed by Trigger (timing) Control Distribution System (TCDS) are recovered from the fiber by AMC13 and transferred to the μ TCA backplane. Single bidirectional fiber connection is established between ngCCM and ngFEC to transfer fast and slow controls as well as the LHC clock together.

server and forms slow-control commands, and sends them to the ngFEC-FC7 board through the IPBUS protocol [62]. IPBUS is a reliable IP-based protocol aiming for controlling hardware modules easily. It communicates with a 32-bit wishbone bus of the integrated circuit and uses the user datagram protocol (UDP) as transfer protocol between software and hardware. Wishbone bus is a parallel communication bus used in intercommunication of the modules inside of an integrated circuit. The handshake between master and slave relies on a strobe signal driven by master and acknowledgment or error response from the slave.

For readout modules and pulser calibration cards, the CCM server receives the parameters from the RBX manager and follows the same path to write these settings to the front-end modules. As an exception, the Peltier-cooling system is adjusted automatically by the CCM Server without having constant data stream between the WinCC server due to timing constraints. More detailed information about the CCM Server - ngFEC-FC7 communication is given in the ngFEC-FC7 firmware discussion in Section 4.2.2.

TCDS and TTC

The TTC signal carries the global LHC clock, the CMS trigger primitives, readout electronics reset signals, and the first bunch of a beam orbit indicator. Due to the high granularity of the CMS apparatus, these signals are very important to have a synchronous data readout which then can be reconstructed into physics objects.

The upgrade of the detector requires higher segmentation in the readout paths, thus the TTC

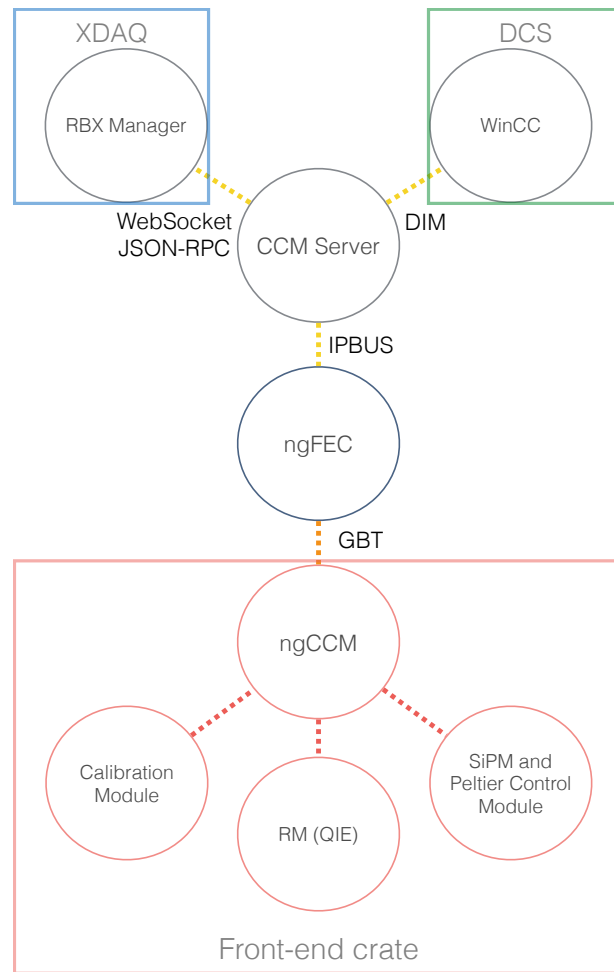


Fig. 4.5: The figure shows the general structure of the HCAL readout-electronics control system. Circles indicate the slow control elements and the dashed lines shows the modules' interconnections. Each protocol that is used in the communication is stated next to connection. The CCM Server connects high level control softwares to the HCAL specific electronics modules.

paths. Moreover, a DAQ overload prevention mechanism becomes necessary to cope with the increased data rate. Due to the performance and the granularity constraints on the previous TTC distribution system, a new distribution system called trigger (timing) control distribution system (TCDS) is developed. It allows more partitions, hence, more sub-detectors and higher granularity. New trigger throttling system (TTS) is introduced to the TCDS in order to ensure that the trigger logic is not overloading.

The system relies on the FC7 AMC board enhanced with custom FMC boards and firmware to distribute TTC signals with fiber cables, replacing the outdated VME standard with the μ TCA standard. The TCDS delivers the TTC signals to the AMC13 board in the ngFEC crate.

AMC13

As its name indicate, the AMC13(XG) board is an AMC board with a small form factor, which can be plugged - unlike other AMCs - into the redundant MCH slot in a μ TCA crate, which, in this particular case, can be called the thirteenth AMC slot in the μ TCA crate. AMC13 has three printed circuit boards (PCB), each handling the different functionalities: The first PCB carries four SFP+ connectors, part of the clock chain, Kintex[®]-7 FPGA, and two 256 MB DDR3 memory. The second PCB houses the rest of the clock chain, a flash memory, MMC module and the Spartan 6 FPGA. The last PCB handles the connectors for the front panel; JTAG and MMC console. The first two PCBs are connected to the backplane of the μ TCA crate.

The AMC13 recovers the fast control signals and 40.0788 MHz LHC clock from the TCDS system via SFP+ connection. Incoming signal carrying encoded TTC commands has the clock frequency of $4 \times$ LHC clock frequency (approximately 160 MHz). The carrier clock is extracted from this signal and divided by four to the LHC clock frequency in the first PCB and then routed to the backplane from the second PCB. The TTC signal, on the other hand, is first routed to Spartan 6 FPGA on the PCB 1 from the Kintex[®]-7 FPGA on the PCB 2. The Spartan 6 FPGA phase aligns the signal with the recovered LHC clock and transmits the data over backplane as a DDR stream. The AMC13 board uses FCLKA pin to distribute clocks to μ TCA backplane, fast controls are streamed from the third lane (Fabric B) of the backplane.

4.2.2 ngFEC-FC7

The required functionality of the ngFEC-FC7 board can be listed as:

- The ngFEC-FC7 should be able to communicate with both front-end modules and the CCM server reliably.
- The fast and slow control signals should be sent to the front-end modules through the existing single-mode bidirectional fiber cable.
- It should distribute the LHC clock to the front-end modules with sub-nanosecond precision.
- In order to diagnose possible interruptions in the $\text{TTC} \rightarrow \text{AMC13} \rightarrow \text{ngFEC-FC7} \rightarrow \text{front-end modules}$ transmission path, it should provide debugging registers related to TTC signals and front-end communication.
- It should support the redundancy such that it is able to control a second front-end crate through interconnections between front-end crates over a single fiber connection.
- Each ngFEC-FC7 should be able to communicate with 6 front-end crates simultaneously through bidirectional fiber connections.

The ngFEC system is using the I²C protocol to communicate with the front-end sensors and the memory elements (debugging, id registers, etc.). I²C is a two-wire, resource efficient, multi master and slave serial bus developed by Philips Semiconductors. Since it only requires two wires

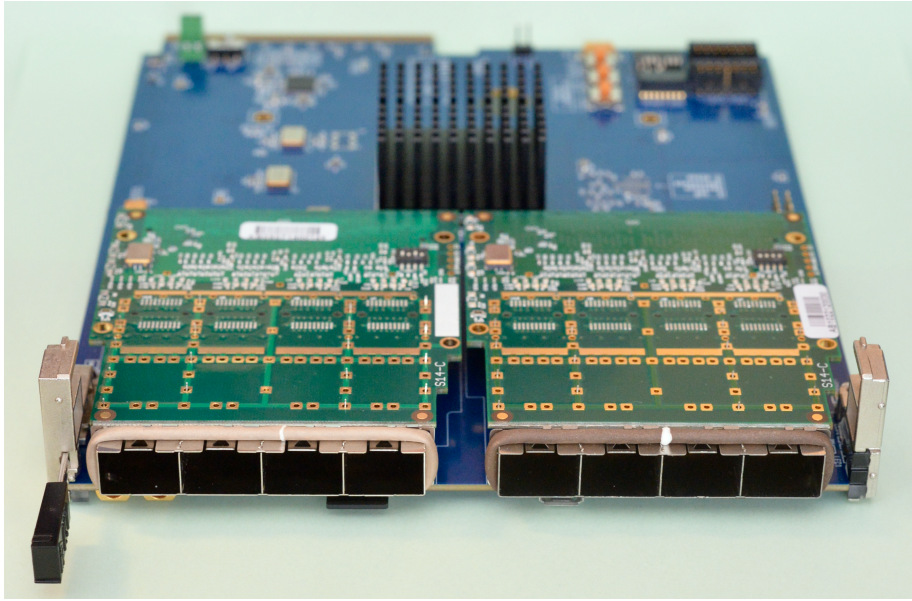


Fig. 4.6: Figure shows the front view of the ngFEC-FC7 board without the front-panel. Two FM-S14 SFP+ transceiver boards are mounted to the base FC7 board. Six SFP+ connectors positioned in the middle are used for the ngFEC - ngCCM communication.

(SDA - 1-bit data and SCL - clock), I²C is used to communicate with the high number of front-end peripherals, sensors and memory elements.

The ngFEC μ TCA crate is capable of controlling over ten thousand I²C buses on the front-end crate through six SFP+ connectors located on the ngFEC-FC7 board. To achieve this, an I²C master is initialized for each front-end primary I²C bus. Although there is only one ethernet connection between the ngFEC crate and the CCM Server, since the IPBUS protocol is almost thousand times faster than I²C protocol, the commands received from the CCM server can be buffered and later executed with I²C transfer speed. In the following, the general functionality of the ngFEC-FC7 board and the custom firmware are discussed.

Hardware

The FC7 board is a custom FMC carrier advanced mezzanine card (AMC) compatible with μ TCA which provides high-speed input/output paths required for data acquisition and detector control applications in high luminosity-energy physics experiments. It is built around a powerful Kintex[®]-7 FPGA. The FC7 board carries two low-pin (68 single-ended or 34 differential pins) FPGA mezzanine card (FMC) connectors. These two slots are occupied by the FM-S14 SFP+ transceiver boards in the ngFEC-FC7, each driving 4 SFP+ connections with transfer speed of up to 10 Gb/sec. Six of these SFP+ connectors are connected to the ngCCM boards in the front-end. The single-mode fiber cables are used to establish these connections. A rear view of the ngFEC-FC7 board with the FM-S14 boards is given in Fig. 4.6.

One of the most important tasks of the ngFEC-FC7 board is the distribution of the LHC clock to the front-end modules, hence the clocking path carries utmost importance. The differential

clock received from the backplane of the μ TCA crate is fanned out to two identical outputs (with less than 20 ps skew) by a SY58608U Fanout Buffer, and then enters to a SN65LVDT250 4×4 crosspoint switch. The GBT logic requires two different clocks synchronized to each other but with different frequencies (LHC clock frequency and $3 \times$ LHC clock frequency), and none or constant phase shifts with respect to the LHC clock recovered from the backplane. Thus, the clock exiting from the cross switch enters the CDCE62005 clock generator and jitter cleaner. Final clocks are fanned out once again in a SN65LVDT250 4×4 crosspoint switch to provide clocks for both halves of the Kintex[®]-7 FPGA.

The FC7 board also includes 4 GBit DDR3-RAM to extend the storage options of the Kintex[®]-7 FPGA.

Firmware

Since the FC7 board is powered by a Kintex[®]-7 FPGA, Xilinx ISE software is used to develop the firmware of the ngFEC-FC7. The firmware is designed with two different low level hardware description languages. Very High Speed Integrated Circuit Description Language (VHDL) is used in the development of the main modules, accompanied by VeriLog in some cases.

The ngFEC-FC7 firmware consists of two main parts: system logic and ngFEC logic. The system logic provides the base functionality to the board: it enables the ethernet communication with the server computers and provides I²C master - slave elements for various sensors positioned in the board. After power on, it establishes communication with the MMC module. The ngFEC logic, on the other hand, is the part where the CCM server communication handshake, command buffering, and front-end communication are handled. The implementation of the ngFEC logic will be discussed in more detail. A general diagram of the main modules that are included in these two parts is shown in Fig. 4.7.

TTC - ngFEC-FC7 communication: The clock and serialized fast control information that are received from the TCDS by AMC13 are directly sent to the μ TCA-crate backplane, where it is de-serialized by the TTC decoder logic. The AMC13 board uses the FCLK pins to distribute the clock within the μ TCA backplane. The fast-control signals are streamed from the third differential pins (p and n) of the backplane with double data rate (DDR). The phase difference between the LHC clock and the fast-control bit stream is adjusted dynamically. There are multiple counters added here to provide debugging information.

CCM Server - ngFEC-FC7 communication: The ethernet connection is established between the server computers and the AMC boards through the MCH and the μ TCA backplane in the ngFEC crate. The recovered IPBUS - UDP packets are translated into 32-bit wishbone words with 32-bit address registers in the system part of the firmware. With an internal master-slave communication, these wishbone words are distributed to the ngFEC logic. A custom handshake protocol is developed to process the information delivered by the CCM Server. Once these wishbone words carrying the CCM Server commands reach the ngFEC logic, they are separated into four different

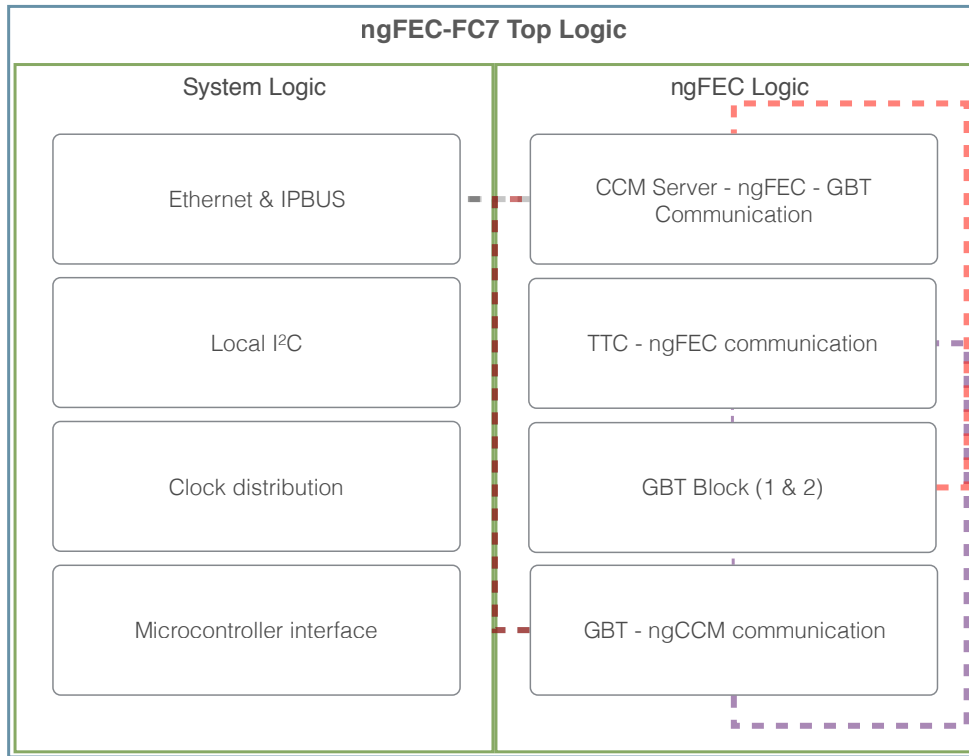


Fig. 4.7: There are two main parts in the ngFEC-FC7 firmware: the ngFEC logic handles the detector control system (the LHC clock, fast and slow controls), and the system logic provides the basic functionality. The main communication paths for the ngFEC system are shown with the dashed lines.

categories: slow controls, manual fast controls, FC7 diagnostic and debugging requests, and the JTAG communications. Each packet that is received from the CCM Server by the ngFEC logic is separated into these four categories for each SFP+ connection and I²C master instance. Table 4.1 shows the address map for the incoming wishbone packets to the ngFEC logic.

Slow control commands consist of front-end slow control commands and front-end sensor diagnostic information requests. The word *slow* in the naming is due to the fact that I²C protocol that is deployed to communicate with the front-end modules and I²C has a relatively slow transfer speed of 100 Kbit/sec. Because of the different transfer speed rates between the I²C and IPBUS protocols, the slow control commands are buffered and then translated to I²C via the I²C master instances in the ngFEC logic, and executed. For each I²C slave on the front-end, there are 4×32 bit registers reserved on the ngFEC-FC7 for indicating status, control and input-output size of the transaction. The control bits are set by the CCM server to send "write" or "process" commands to the ngFEC-FC7. The ngFEC-FC7 changes the status bits upon the completion of the writing of the pipelined commands to the buffer and transaction with the front-end slaves. This execution sequence enables the CCM Server to make efficient use of computing time and network bandwidth. Buffering slow-control commands also allows simultaneous communication with the I²C slaves, since the transactions can start simultaneously once they are buffered.

Table 4.1: The ngFEC logic separates the incoming instructions with an address map. The address is given in little endian formation: The first 8 bits separate the ngFEC - CCM Server communication logic from the rest of the communication. The addresses corresponding to the slow command buffers are divided into two sub-buffers as received commands and their responses from the front-end modules.

Wishbone address bits	Supported values in hexadecimal	Mapping definition
[32 - 25]	[6E]	Offset address of the CCM Server - ngFEC-FC7 communication logic
[24 - 21]	[8 - 1]	GBT link (SFP+ connection)
[20 - 17]	[F - 0]	I ² C master
[16 - 0]	[1388 - 0000]	The memory offset for the slow command buffers and status registers

ngFEC-FC7 - ngCCM communication: The communication between the ngFEC-FC7 board and front-end modules is established via GBT/GTX protocol. The GBT protocol (Appendix A) provides communication speeds up to 10 GBit/s in an irradiated environment. The GBT implementation in the ngFEC-FC7 - ngCCM communication operates at 4.8 GBit/s (120×40.0788 Mbps). The 120-bit GBT frames are sent and received at speed of the LHC clock frequency. These GBT frames consist of an 4-bit header word, 84 bits data and 32 bits Reed Solomon forward error correction (RS FEC) bits. In order to ensure a reliable communication between front-end and back-end systems, this scheme allows the GBT logic to correct 16 sequential error bits which can occur under extreme irradiation conditions at the CMS HCAL detector. In addition to the RS FEC, a one-to-many pseudo-random-binary sequence (PRBS) error rate test is implemented to detect the faulty patterns that cannot be corrected by the RS forward error correction algorithm. After establishing a reliable communication between the ngFEC-FC7 and ngCCM, important diagnostic information, clock and control commands are exchanged. The diagnostic information from the front-end sensors are read via I²C protocol through the fiber connections which have a data transfer rate of 4.8 Gbit/sec whereas the I²C transfer speed is around 100 kbit/sec. Therefore, slow control commands as well as the ngFEC-FC7 specific commands received from the CCMServer are stored in dual-port RAM blocks in the Xilinx® Kintex®-7 FPGA on the FC7 and then processed by the ngFEC-FC7 logic.

The fast controls received from TTC through the AMC13 are directly sent to next generation Crate Control Module (ngCCM) [63]. The clock and the serialized fast control signals are then recovered by the ngFEC-FC7 board.

The commands stored in the buffers are read and executed by following the instruction scheme given in Table 4.2 for the first most significant 8 bits. These instructions include both I²C master related commands and local logic instructions (wait and timeout cycles etc.). Data stored in the

Table 4.2: The buffered information also contain the instructions for the ngFEC logic: according to the first 8-bit of the buffered command, ngFEC logic executes the corresponding instruction. The table shows bit-mapping of these instructions following the little endian ordering. The most significant two bits give the execution cases, once the execution case is satisfied ngFEC logic starts following the instructions. The instructions set by the bits [29-27] are exclusive to each other. Table 4.3 shows the content of the data fields for each instruction.

Wishbone data bits	Instruction
32	Skip this command if the reply status byte is set
31	Skip this command if the reply status byte is not set
30	Clear reply status
29	Write I ² C master control register
28	Read I ² C master status register
27	Wait for the given wishbone clock cycles
26	Empty
25	Write output to the BRAM buffer
[24 - 1]	Data field

Table 4.3: The table shows the data field corresponding the instruction

Instruction	Wishbone data bits	Content
Write I ² C master control register	[16-9]	value for the control register
	[8-1]	Data to be transmitted to front-end
Read I ² C master status register	[24-17]	Bit mask for the reply status byte
	[16-1]	Timeout in wishbone clock cycles
Wait for the given wishbone clock cycles	[24-1]	Number of wait cycles

Table 4.4: The buffered information also contain the instructions for the ngFEC logic: according to the first 8-bit of the buffered command, the ngFEC logic executes corresponding instruction. The table shows bit-mapping of these instructions following the little endian ordering. The most significant two bits give the execution cases, once the execution case is satisfied ngFEC logic starts following the instructions. The instructions are set by the bits [29-27] which are exclusive to each other. The table 4.3 shows the content of the data fields for each instruction.

Wishbone data bits	Instruction
32	Skip this command if the reply status byte is set
31	Skip this command if the reply status byte is not set
30	Clear reply status
29	Write I ² C master control register
28	Read I ² C master status register
27	Wait for the given wishbone clock cycles
26	Empty
25	Write output to the BRAM buffer
[24 - 1]	Data field

last 24 most significant bits carry information regarding the instructions given in the first 8 bits. If the "write to buffer" instruction is enabled, the responses received from I²C masters are stored following the scheme given in Table 4.4. The ngFEC crate communicates with the ngCCM via GBT protocol through fiber connections. One bit (SDA) is reserved in the 84-bit GBT word for each I²C slave on the front-end crate. The I²C master on the ngFEC-FC7 controls the SDA bits and all the masters are sharing the same I²C clock. Therefore, by updating the SDA bit with the given I²C clock, even though the data transfer rate is 4.8 Gbit/s, I²C communication can be maintained simultaneously and independently for each front-end I²C slave.

The ngFEC-FC7 is the connection point of the front-end modules, the CCM server and TTC signals provided by TCDS. One of the main challenges for connecting these three different systems is the different communication speed between each module's connection to ngFEC-FC7. Moreover, even within one interconnection, different communication protocols are used, therefore, slower communication protocols are carried with faster ones. The ngFEC-FC7 logic flow for a single I²C master-slave communication is displayed in Fig. 4.8.

The design multiplicity of the control links to be driven from each ngFEC-FC7 board is six connections. The ngFEC-FC7 board can support up to seven control links. The main constraint in the

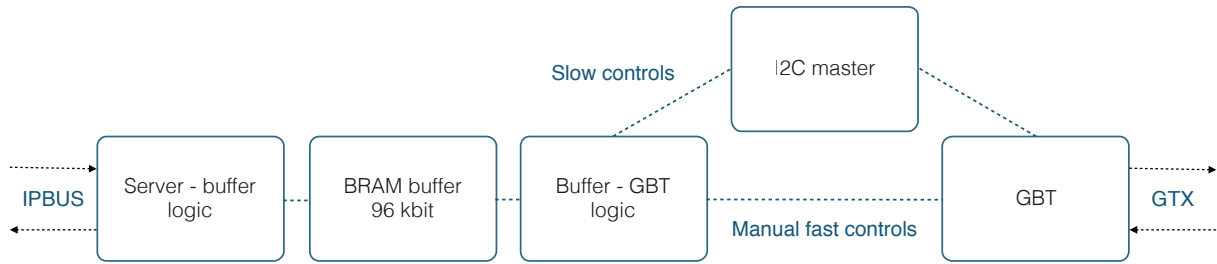


Fig. 4.8: The ngFEC-FC7 communication logic with a single I²C slave

number of control links is the clocking resources in the Kintex[®]-7 FPGA. This constraint is related to the fact that in order to have minimum skew, the clocks that are connected to the high speed logic elements should be distributed by global buffers or horizontal buffers. Phase-locked loops (PLL) are control structures that are essential to these high speed logic elements. A basic PLL consists of an oscillator which can synchronize with an input clock. It can adjust delay or frequency of the output clock with respect to the input clock. Structurally, the Kintex[®]-7 FPGA is partitioned into sixteen sub-regions (8 vertical slices divided into two horizontal sub-regions), and vertically divided into two main halves (top and bottom). Each half of the FPGA has a limited number of global buffers, and is independent in terms of clocking scheme. There are four independent PLLs (two of the advanced PLLs called MMCMs allowing dynamic phase locking) in each vertical octant. One high-speed GTX common module (which also provides quad PLL) is positioned in each of the eight vertical slices. GTX is a low-latency, high-speed serial transceiver that supports multiple communication modes within Xilinx FPGAs. Four GTX channels can be connected to one GTX common module. A GBT bank instance requires a GTX common instance and GBT links use GTX channel instances. All of these modules, hereby, constrain the layout of the GBT logic as they must be supplied by high-speed low-skew clocks. The GBT implementation in the ngFEC-FC7 firmware uses two GBT-banks each connected to three GBT links. The GBT logic operates with two different clock frequencies, the LHC clock recovered from the backplane is used as the frame-clock and the multiplied clock as transmit reference clock. The transmitted clock is three times LHC clock frequency and it is carried via the GBT/GTX serialized stream. It is multiplied and synchronized to LHC clock in the CDCE62005 chip with minimal jitter. The transmitted clock should have the same frequency with the same phase for all links, therefore, the same clock is used for each GBT bank and only buffered twice using the horizontal buffers. The latter can be used by exploiting the fact that all GTX channel connections connected to a GTX bank lie in the same sub-clocking-region in the Kintex[®]-7 architecture. This clocking scheme allows to save internal FPGA PLLs and global buffers for synchronizing, distributing and dividing the received clocks back to the LHC clock frequency. The received clocks are dynamically phase adjusted since the time that it takes to be transmitted back from front-end crates can be different for each link. Therefore, each received clock requires a separate PLL. Since all PLLs are distributed evenly on the Kintex[®]-7 architecture, an additional global buffer is required for collecting the clock from different sub-regions. Three SFPs (the second, third and fourth connectors) are connected to the

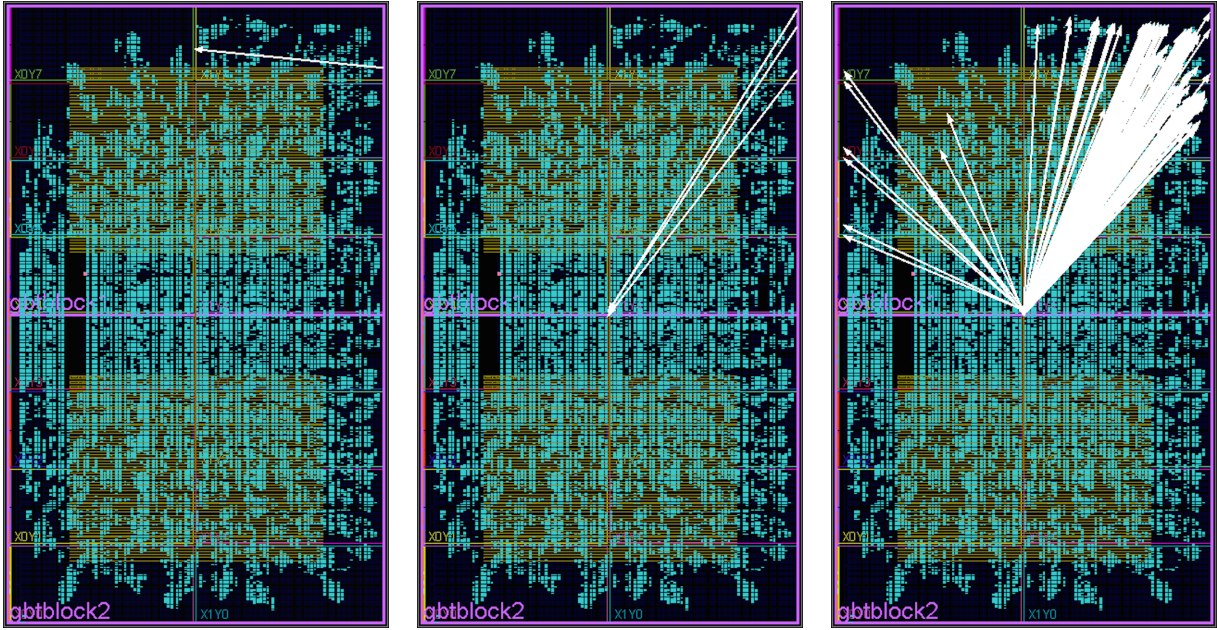


Fig. 4.9: Area utilization of the ngFEC-FC7 firmware on Kintex[®]-7 FPGA: the white lines indicate the clock source and destinations. The figure on the left shows the distribution of the LHC clock as GBT transmit clock on the ngFEC-FC7 Kintex[®]-7 FPGA through a horizontal buffer for a single GBT bank. The recovered clocks from the front-end modules follow a different path: First they are buffered in the center of the FPGA (center figure) then they are distributed to phase locked loops (PLLs) and GBT logic blocks (right figure).

same clocking-sub-region, therefore, by omitting the first connector (which would require a separate GBT bank), three links can be efficiently driven from the top FM-S14 board. For the bottom connector, since all pins are connected to the same clocking-sub-region, all four connections can be used with the minimum possible clocking resource usage without introducing a second GBT bank.

4.2.3 Front-end crate

The front-end (FE) crates are located close to the CMS detector, and must be able to operate in the radiation-harsh environment. Therefore, they consist of radiation-hard or tolerant modules. As a consequence, FPGAs used in the front-end modules are flash-based FPGAs rather than SRAM based FPGAs, which tend to have smaller memory cell size, thus more susceptible to radiation damage. Different partitions of the CMS HCAL have different form factor and management requirements.

Readout modules

The FE crate hosts front-end readout modules of the CMS HCAL. A CMS HCAL FE readout module contains Charge Integrator and Encoder (QIE10) [64] application specific integrated circuits (ASICs) which sum and digitize charges collected by fast sources like Hybrid Photo Diodes (HPD),

Photo Multiplier Tubes (PMT) etc. The FE readout board also carries a ProASIC3L [65] bridge FPGA which provides access to other components on the board (QIE, GBTx chips etc). An I²C slave is implemented on the bridge FPGA, routing the slow control commands to other components by switching I²C lines with a custom scheme between the CCM Server and the bridge FPGA. Therefore, with one I²C link to the FE readout board, the CCM Server effectively controls and monitors all the components on the board.

ngCCM

The front-end crate is controlled by the next generation Crate Control Module (ngCCM) [63], carrying a Versatile TransReceiver (VTRx) [66], which is a CERN developed, radiation hard, SFP+ based transceiver. A duplex single-mode fiber connects the ngCCM to the ngFEC. Via this connection, ngCCM recovers the LHC clock and control signals, and distributes them to the front-end crate. Triple modular redundancy (TMR) is enabled in the ngCCM firmwares (except for the continually changing high-speed logic elements) to make the firmware resilient against radiation damage. To protect the board from radiation induced latch-ups, an automatic detection and power cycle system is implemented. The CMS HCAL sub-detectors are exposed to different radiation fluences. They also require different form-factors and material/shape allowances for the front-end electronics. Therefore, two different versions of the ngCCM boards are designed to meet these requirements, namely HCAL forward (HF), and HCAL barrel and endcap (HB and HE)

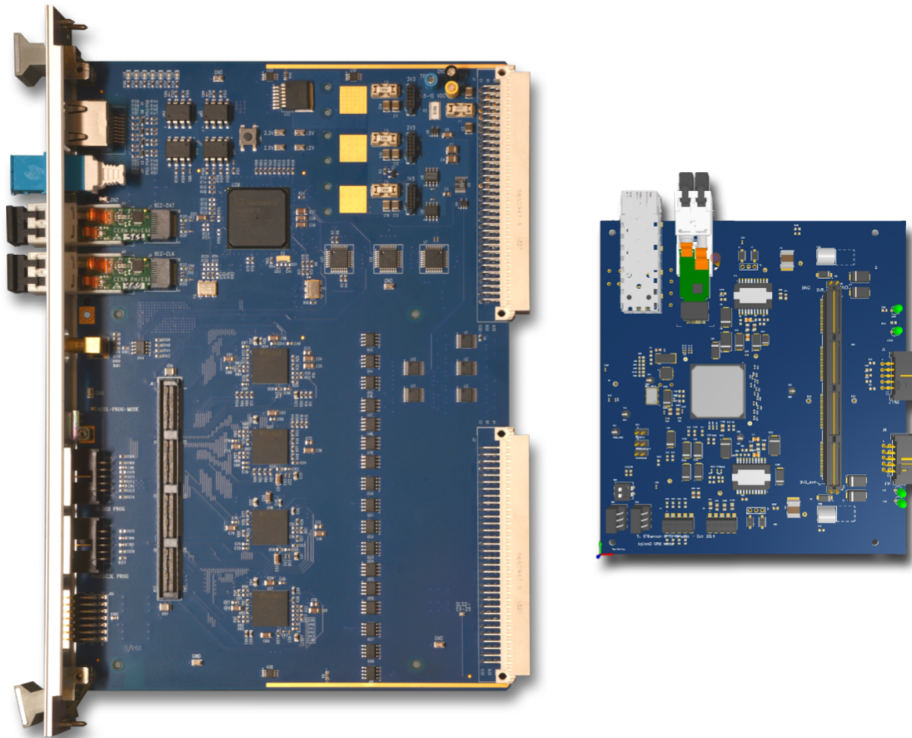


Fig. 4.10: The HF ngCCM motherboard (left) and the HF ngCCM serializer-deserializer mezzanine board (right) [63].

ngCCMs. The HF ngCCM board consists of two parts: ngCCM motherboard and ngCCM serializer deserializer (SerDes) mezzanine. The SerDes board emulates the GBTx chip, and hence handles the GBT communication. It recovers the LHC clock, fast and slow control signals via the GBT protocol implemented on a IGLOO2 FPGA [67]. The motherboard distributes the fast control signals and the LHC clock and routes the slow control commands to the readout modules. These signals are also multiplexed by the ngCCM motherboard to be transferred to the neighboring ngCCM board in a case of failure of the direct communication between the ngFEC and the neighboring ngCCM board, hence providing redundancy to the control path. It also has an internal I²C slave that sends temperature and status information to the CCM server. In total, there are sixteen I²C slaves directly reachable from the HF front-end crate backplane (thirteen different boards), two internal ngCCM I²C slaves, and a redundant path. All of these applications are handled by four ProASIC3L FPGAs positioned on the motherboard.

Due to space limitations, the HB/HE ngCCM circuitry is distributed to four smaller PCBs. The central two of these PCBs are identical control boards, the other boards are very similar to each other and handle the clocking scheme. One of the identical control boards handle full redundancy is connected to the adjacent ngCCM via an optical fiber cable. Similar to HF ngCCM, in case of a failure, the communication with ngFEC is routed from the neighbor ngCCM. Unlike HF ngCCM, a complete control path is provided with the redundant connection instead of a single I²C line. This is especially important since full detector disassembly is required to access the HB/HE ngCCM boards.

4.3 System test

Since the main work within the context of this thesis work was done on the development of the ngFEC and the test of the control system, the test procedure is discussed in two sub-sections: ngFEC test and the overall system test.

4.3.1 ngFEC test

The testing methodology followed to check the ngFEC system is given in the following:

Hardware

The hardware of the ngFEC system is tested for long continuous operations at the DESY test stand. The temperature and voltage are recorded by various control software communicating with the crate via MCH. FC7 boards with the ngFEC firmware are tested for two weeks before being sent to the production systems. During these tests, voltage and temperature levels are recorded and checked using the ChipScope Pro software over the JTAG connection. An example 120 minutes test period is shown in Fig. 4.11. The FM-S14 boards are tested using a KC-705 evaluation board with the provided BERT firmware, and looping back one of the fiber cables to another SFP+ module. Loop-back tests were also performed to cross-check the PRBS error counters and the

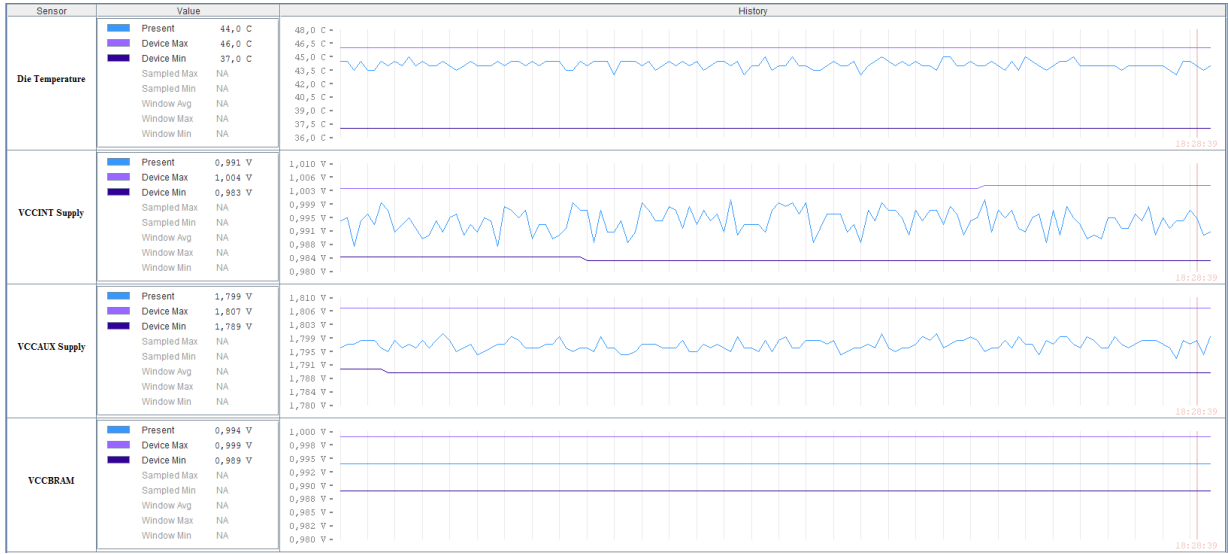


Fig. 4.11: Voltage and temperature sensors' readings of the ngFEC-FC7's XADC chip. The values are read from the JTAG connection of the ngFEC-FC7 and they are within the expected limits.

FM-S14 boards on the ngFEC-FC7. The AMC13 boards are checked with similar loopback tests as well.

Firmware

In order to check the functionality of the basic communication with the CCM server, ChipScope debugging cores are used in the ngFEC-FC7 firmware. After ensuring a reliable communication between the ngFEC and ngCCM server, the ngFEC buffers and the communication protocol are tested thoroughly. The LHC clock received from the TCDS system is measured to match the expected frequency of 40.788 MHz. This is done by an internal clock rate measurement logic of the ngFEC firmware that uses an independent 100 MHz μ TCA-backplane clock as a measurement reference. One of the most important responsibilities of the ngFEC is to distribute the received LHC clock with a constant phase shift, which then can be corrected by the central systems. The following methodology is used to test this functionality: The LHC clock is routed to a debug pin on the FC7 board and this debug pin is connected to a probe of a high-speed oscilloscope. The same clock is routed directly from another AMC slot of the backplane. The ngFEC-FC7 is repetitively restarted to observe possible phase shifts. The tests performed at the DESY test-stand shows no indication of a visible phase shift between restart cycles.

4.3.2 Control system test

There are four main communication paths that need to be tested in the control system overall: TCDS - AMC13, AMC13 - ngFEC-FC7, CCM Server - ngFEC-FC7, and last but not least ngFEC - ngCCM.

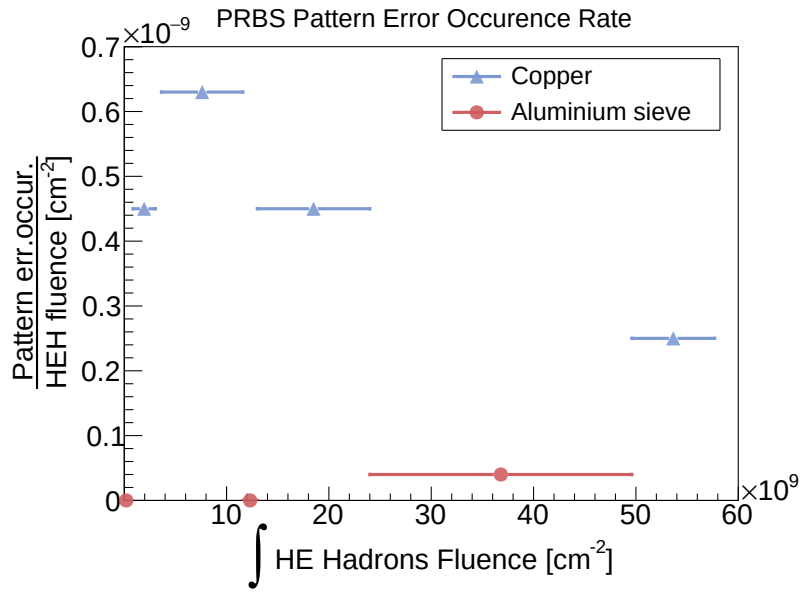


Fig. 4.12: PRBS pattern error occurrence rate in the ngCCM $\rightarrow$ ngFEC path. The errors are observed in bursts. These error bursts are expected to be due to single point failures in the ngCCM board, therefore, the occurrence rates are quoted to indicate these instances. The connection is automatically recovered after these error bursts.

The first two paths can be tested via the internal AMC13 and ngFEC TTC debugging counters. Both the ngFEC and the AMC13 firmwares are equipped with counters that checks the double data rate TTC stream for possible single and double errors. Both firmwares have debugging registers to check periods of per orbit signals indicating the first bunch of an orbit. Moreover, the ngFEC-FC7 has a three-bin histogram indicating the timing of the QIE reset signal (early, on-time, late). All of these counter registers are recorded during the system test to ensure the reliability of the connection. The AMC13 firmware can also produce these signals locally to perform the tests independent of the central TCDS system.

Since the front-end crate is not accessible during the operation, ensuring reliability of the ngFEC - ngCCM communication is extremely crucial. All these tests can be performed at a test-stand. However, the most important test scenario is the front-end module's behavior under radiation. Therefore, a radiation test was performed at the CERN CHARM facility in October 2015. The PRBS error rate count was measured on both ends (ngCCM and ngFEC) via scripts built on the CCM server with different radiation levels. The scripts read the system status once in every two minutes. Each two minute period with an increase in the error counters was recorded as an error occurrence. No errors have been observed in the ngFEC $\rightarrow$ ngCCM path and maximum of 5 error occurrences per hour with an average six hundred pattern errors were observed in the ngCCM $\rightarrow$ ngFEC path under $5.8 \times 10^{10} \text{ cm}^{-2}$ high energy hadrons integrated fluence that is equivalent to the radiation exposure expected at 13 TeV with 3000 fb^{-1} integrated luminosity. The rate of these negligible error burst are shown in Fig. 4.12.

4.4 Outlook

The CMS HCAL detector readout control system, armed with improved modules, firmwares and softwares, is more reliable, capable and faster than the legacy control system of the CMS HCAL. The ngFEC-FC7 firmware with the new CCM server can simultaneously control up to sixteen I²C slaves and send fast controls for each SFP+ link. The system has been successfully tested under high radiation fluence. The assembly is expected to be finalized by the end of 2016 for the HF and 2017 for HE/HB.

Object Reconstruction and Variable Definitions

Contents

5.1	Object reconstruction	60
5.1.1	Particle-flow algorithm	60
5.1.2	Primary vertex	60
5.1.3	Muons	61
5.1.4	Electrons	63
5.1.5	Taus	64
5.1.6	Jets	65
5.1.7	Missing transverse energy $\cancel{E}_T$	69
5.2	Event and object selection	71
5.2.1	Event cleaning	73
5.2.2	Preselection	74
5.3	Variable definitions	76
5.3.1	Low-level variables	76
5.3.2	High-level variables	78

Raw information obtained from the CMS detector are in the form of detector hits and deposited charge information which cannot be used directly in a physics analysis. In this chapter, the relevant physics objects that are reconstructed from the raw information are described. These objects are used to construct discriminator variables which are later used to classify the signal and background events. In Section 5.2, methodologies for various object reconstructions are presented. Selection of the objects for forming the variables is summarized in Section 5.3, which is followed by the description of these variables in Section 5.3.

5.1 Object reconstruction

In this section the object-reconstruction methodologies used in the present analysis are described.

5.1.1 Particle-flow algorithm

With its large tracker system and high granular ECAL, the CMS detector is suitable to employ the particle-flow algorithm [68] which aims to effectively use the information gathered from the detector subsystems in the object reconstruction.

In the following sections, the CMS reconstruction and identification for each physics object used in this analysis is separately provided. This section focuses on the PF algorithm which is widely employed in the reconstruction of these objects. The CMS PF algorithm uses three main *elements*:

- Tracks reconstructed from hits in the tracker,
- Energy deposits in the calorimeters,
- Tracks reconstructed from hits in the muon system.

The elements that are possibly related to a single object are *linked* together to form a *building block*. The link algorithms vary according to element type and the related object e.g. The tracks in the muon system is fit to the inner tracks, if the fit converges to an acceptable χ^2 p -value, and in case of electrons, tracks are linked to the tangential ECAL clusters due to bremsstrahlung.

The objects are reconstructed in the order of highest expected reconstruction performance to lowest, thus gradually eliminating unrelated blocks for the more ambiguous object reconstructions. Therefore, due to the clean environment of the muon system, the first objects that are reconstructed are muons, which are then followed by electrons, taus, photons, and hadrons.

5.1.2 Primary vertex

The primary vertex reconstruction aims for the precise determination of the position of the collision of interest. In order to increase the collision rate, hadron beams are strongly focused in the transverse plane at the CMS interaction point [69]. Even with the increased density, the collision occurrence follows Poisson statistics, and only a few dozen of collisions occur in addition to the collision of interest. However, due to spatial resolution restrictions, eliminating the impact of additional - *pile-up* collisions can be challenging.

The primary vertex reconstruction starts with a track cluster selection process. Various track quality criteria including a set of constraints on the transverse impact parameter, the number of tracker hits, and the normalized-track χ^2 , are required. The selected tracks are clustered with respect to their transverse distance from the beam axis (with a maximum of 1 cm distance). The primary vertices are reconstructed using these track clusters with the adaptive vertex fitter [71]. A weight is assigned to each track proportional to the track's association to the vertex candidate.

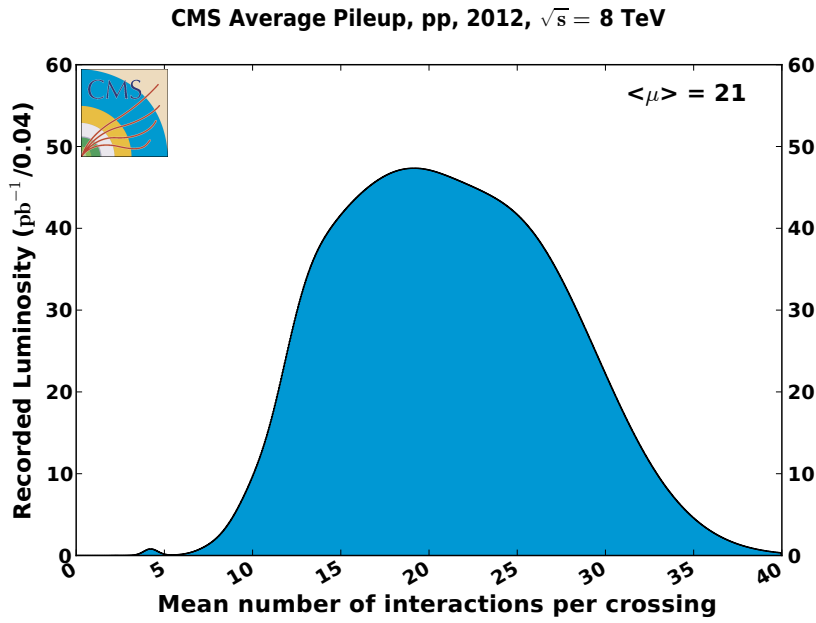


Fig. 5.1: The recorded luminosity of data with respect to the mean number of pile-up per bunch crossing in 2012 [70].

Similar to other adaptive algorithms, instead of trimming the fluctuating tracks, a lower weight is assigned. The algorithm iterates until it reaches a predefined vertex-candidate-position stability.

The algorithm returns the number of degrees of freedom in order to quantify the goodness of the fit:

$$N_{\text{dof}} = 2 \sum_{\text{tracks}} w_i - 3, \quad (5.1)$$

In this thesis, a *good primary vertex* is required to have the N_{dof} to be greater than 4.

Pile-up

The number of pile-up interactions is distributed following Poisson statistics for a given instantaneous luminosity. Over a range of integrated luminosity, pile-up distributions due to different instantaneous luminosities form a sum of Poisson distributions. Estimation of this cumulative distribution is not possible, thus the MC samples are generated with a prior distribution which may differ from the actual pile-up distribution. The MC samples are corrected a posteriori by applying weights.

5.1.3 Muons

The Compact *Muon* Solenoid detector with its dedicated muon system, has a state of the art muon reconstruction and identification capability. Relatively high energetic muons with their long-lifetime reach to the muon system where the composition of particles is less diverse.

The following algorithms are used for the reconstruction of muons [72] in CMS:

- **Standalone muon:** only tracks from the muon system are used in this case. Standalone muons are important in applications that need a short reconstruction time, i.e. trigger.
- **Tracker muon:** The tracks from the inner tracker system, which are reconstructed with the Kalman filter, are extrapolated to the muon system. Matching tracks are used for forming the tracker muons.
- **Global muon:** In this reconstruction scheme the muon system hits are matched to the tracker system. In addition to the tracker information, the matching energy deposition in the calorimeter system is also taken into account in the global muon reconstruction.

The transverse momentum resolution is dominated by the silicon tracker, and only for very high-momentum muons ($p_T \gtrsim 200$ GeV) the global fit improves the momentum resolution as shown in Fig. 5.2. The global muons that satisfy following criteria are labeled as the *good muons*

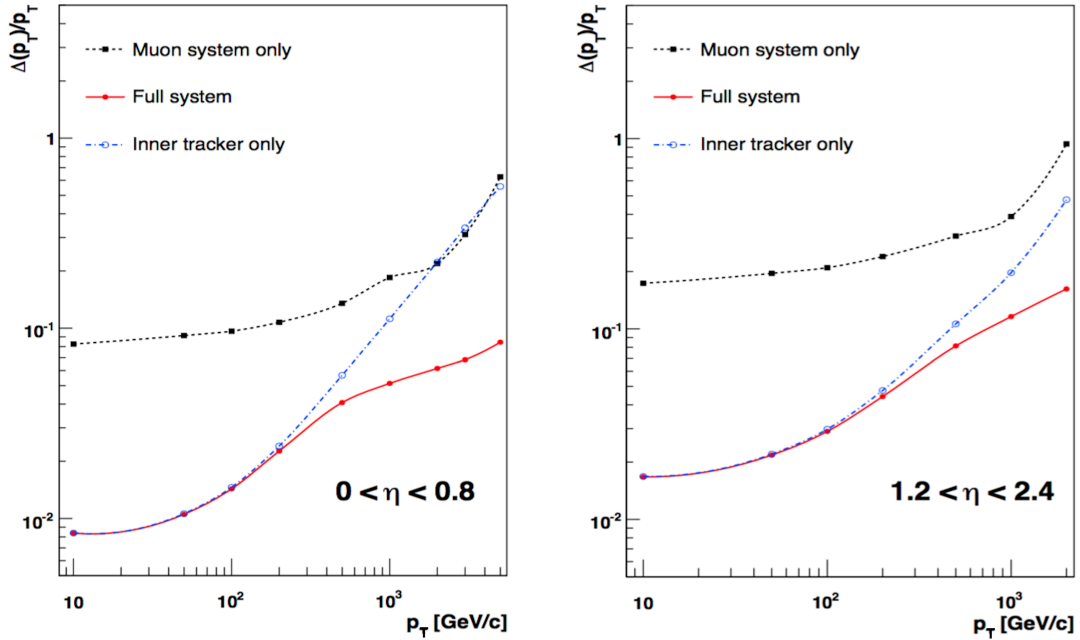


Fig. 5.2: The muon p_T resolution as a function of p_T using the muon system only (black dashed curve), the inner tracker only (blue dash-dotted curve), and both (red solid curve). Left: $|\eta| < 0.8$. Right: $1.2 < |\eta| < 2.4$ [31].

and used in the present analysis:

- The χ^2 normalized to the number of degrees of freedom in the fit ($\chi^2/\text{n.d.f.}$) is required to be smaller than ten and the global fit must include at least one muon chamber hit. These requirements are employed to suppress punch-through hadrons and muons from hadrons decaying in flight.
- Readout from two muon segments is required to prevent accidental track-muon system matches.

- To further suppress muons from hadrons decaying in flight, the inner track has to be matched to at least one hit in the pixel detector. The goodness of the p_T measurement is ensured by requiring more than five tracker layers with hits.
- To reject cosmic-ray muons, the reconstructed muon trajectory has to pass the primary vertex closer than 0.2 mm in the transverse plane. Requiring the track to pass closer than 1 mm to the primary vertex in the z -direction is discarding muons coming from pile-up interactions.
- The muon is required to be isolated from the other particles produced in the event. The absolute isolation is defined as the sum of the energies of all charged particles, photons, and neutral hadrons reconstructed by the particle-flow algorithm within a cone of $\Delta R \equiv \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} < 0.3$ around the muon track. The expected average contribution coming from the pile-up interactions is subtracted. The relative isolation Iso_{rel} , defined as the ratio between the absolute isolation and the p_T of the muon, is required to be smaller than 15%.
- The spatial acceptance is restricted to the tracker, ECAL and muon system coverage (i.e. $|\eta| < 2.4$) and the transverse momentum is required to be higher than 10 GeV.

5.1.4 Electrons

The electron reconstruction relies on the excellent silicon tracker and ECAL performances. Due to their low masses, electrons radiate higher bremsstrahlung compared to other leptons, hence leaving a distinct tangential energy deposition signature in the ECAL.

The electron identification and reconstruction [73] are initialized by forming ECAL clusters. The clusters are reconstructed using a supercluster algorithm. The strategy is to collect energy coming from showering due to bremsstrahlung of electrons in a cone of 0.3 rad around a seed crystal. In order to select trajectory seeds, hits on the innermost tracker layers are matched with superclusters. To ensure these particles are coming from the interaction point, a matching process uses the energy-weighted average impact point of the electron and photons coming from the bremsstrahlung of electrons. This method is especially efficient for high p_T electrons.

The default track reconstruction algorithm in CMS is the Kalman filter. Even though the filter does not require a certain prior distribution, it performs better with Gaussian yields. Bremsstrahlung of electrons result in an energy loss which can be described with a Gaussian function, leaving the actual yield as a superposition of Gaussian distribution functions. Therefore, to increase the reconstruction efficiency, a generalized Kalman-like filter is applied. The Gaussian Sum Filter (GSF) is optimized for superimposed Gaussian yields. Thus to construct trajectories of electrons, all electron seeds are used and electron track parameters are estimated using a GSF fit.

The final step is the estimation of the true fraction of electron energy by taking the difference between the momentum of the outermost track and innermost track segment. All information gathered in these steps is put in together in the reconstruction of *GSF electrons*.

GSF electrons passing the following criteria are referred to as *good electrons* in this thesis:

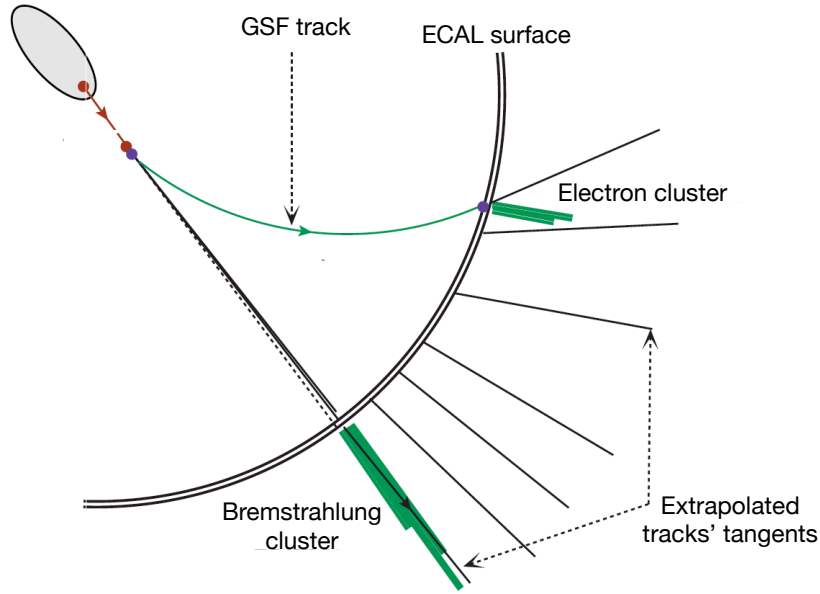


Fig. 5.3: Due to higher bremsstrahlung, an electron leaves distinct tangential tracks during the interaction with the detector. The extrapolated tangential tracks and ECAL energy clusters are combined together to reconstruct electrons.

- The ratio of the energy deposition in the HCAL over the ECAL energy deposit is required to be less than 0.12 for the barrel and 0.1 for the endcaps region. Moreover, the difference in the supercluster energy deposit and the GSF fit momentum is restricted to $|1/E_{SC} - 1/|\vec{p}_{GSF}|| < 0.05$.
- A conversion vertex fit finder is employed to filter the electrons coming from photon conversions. The relevance to the primary vertex is ensured by requiring the reconstructed track to be within 0.2 mm in transverse plane and 0.1 mm in the longitudinal plane.
- A relative isolation $\text{Iso}_{\text{rel}} < 0.15$ similar to the muon Iso_{rel} is required.
- The spatial acceptance is restricted to the tracker and the ECAL (i.e. $|\eta| < 2.4$), and the transverse momentum is required to be higher than 10 GeV.

5.1.5 τ aus

Compared to other leptons, the reconstruction of τ leptons with their high mass and short lifetime is not only challenging, but also unreliable. In the present analysis, hadronically decaying τ leptons are reconstructed to reject dileptonic events which escape the single-lepton requirement.

The τ lepton identification algorithms [74] use decay-mode-identification techniques to distinguish τ leptons from jets.

Even though τ leptons decay dominantly to hadrons, the leptonically decaying τ leptons are also visible in CMS as decay products of secondary lower mass leptons (muons and electrons) which survive after the corresponding lepton selection. These events are accepted in the event selection due to their low influence in the present analysis.

5.1.6 Jets

Color charged particles cannot be observed individually in a particle detector due to *confinement* phenomena. During *hadronization*, they are fragmented into colorless bound states (i.e. baryons and mesons) or elementary particles. Therefore, a parton or a gluon leads to a cluster of particles in the detector which is called a *jet*. Following energy and momentum conservation, energy content and directional properties of these particles are given by the initial *hadronized* particle. Hence, collecting these secondary particles and reconstructing the initial colored object's feature, though challenging, is necessary for understanding the underlying physics process. Various techniques are constructed to tackle this problem. Jet reconstruction techniques should have two distinct features in order to avoid mis-reconstruction of jets:

- *Infrared safety*: If the clustering algorithm does not require an energy threshold, any object even with very low energy can be mistakenly assigned as a center of a jet. For instance, two particle clusters that are close to each other can be merged into a single jet due to a soft object in the middle of the two clusters. Algorithms that are robust against the negative impact of soft objects are described as infrared safe.
- *Collinear safety*: Collinear splitting of a jet can cause an algorithm to misclassify a single hard jet as two softer jets. Collinear-safe algorithms can distinguish such cases and jets with multiple splitting can be reconstructed correctly.

Jet reconstruction algorithms can be discussed under two titles:

Cone algorithms

A jet can be imagined as a spray of particles in the detector bounded to a cone shape. Within a presumed cone, a colored object can be reconstructed by taking the central axis of the cone as its direction and the sum of energies within the cone as its energy.

A subset of cone algorithms chooses one particle as a midpoint and reconstruct cones with a predefined cone radius. However, these algorithms, without thresholds, are not infrared and collinear safe (IRC safe). Iterative cone algorithms start from the hardest object as a seed, construct clusters within a predefined cone and eliminate them in the calculation of softer objects. Comparing to the midpoint approach, iterative algorithms are robust against soft objects, but they are not collinear safe. By removing the concept of seed, the cone algorithms can be IRC safe. One of the practically useful implementation of this type of algorithms is the Seedless Infrared-safe Cone (SIScone) algorithm which iteratively considers a subset S of all particles such that a cone encapsulating only the subset S can be constructed.

Sequential algorithms

Instead of a geometrical shape assumption, sequential jet reconstruction algorithms cluster two of the closest objects together and iteratively reconstruct a closed area of objects until they reach a termination criterion [75]. The Cambridge/Aachen (C/A) algorithm uses a predefined value for the distance as the truncation criterion which is defined as:

$$d_{i,j} = \frac{\Delta_{i,j}^2}{R^2} \quad (5.2)$$

where $\Delta_{i,j}^2 = (\Delta\phi)^2 + (\Delta y)^2$. However, Eq. 5.2 is a purely geometrical distance, disregarding momentum information of the individual objects. The k_T algorithm updates Eq. 5.2 and considers a distance weighted to the transverse momentum of softer objects:

$$d_{i,j} = \min(\Delta(k_{T,i}^{2p}, k_{T,j}^{2p}) \frac{\Delta_{i,j}^2}{R^2}, \quad (5.3)$$

and an additional *particle-beam* distance constraint is included in Eq. 5.3, given by:

$$d_{i,B} = k_{T,i}^{2p}, \quad (5.4)$$

where p is 1 for the k_T algorithm. The k_T algorithm preferably merges softer objects, and it is IRC safe. However, the particle clusters that are formed by the k_T algorithm tend to be geometrically irregular as shown in Fig. 5.4. The clusters with jagged boundaries complicate certain detector corrections, since irregular boundaries may correspond to regions that have different corrections from the given radius R . More regularly formed jet topologies can be achieved by setting the p parameter to -1 as in the anti- k_T algorithm, thus clusters are centralized around harder objects with smoother boundaries as shown in Fig. 5.4. Therefore, modern sequential jet reconstruction algorithms can be summarized by Eq. 5.3 and Eq. 5.4, where

- $p = 0$ for A/C algorithm,
- $p = 1$ for k_T ,
- $p = -1$ for anti- k_T .

In this study, the anti- k_T algorithm is used for jet reconstruction.

According to information obtained from sub-detectors to reconstruct jets, three different types are defined in CMS:

- **Calorimeter (Calo) Jets:** The calorimeter towers, consisting of the ECAL crystals and HCAL cells are solely used for reconstructing Calo jets. In order to reduce pile-up event contribution, a requirement on E_T of 0.3 GeV is applied on the tower energy deposits.
- **Jet-plus-tracks (JPT) jets:** The Calo-jet algorithm is extended using charged-particle track information for JPT jets. The tracking subsystem information is useful for correcting energy reconstruction of charged objects and for inclusion of member charged particles that do not reach the calorimeters.

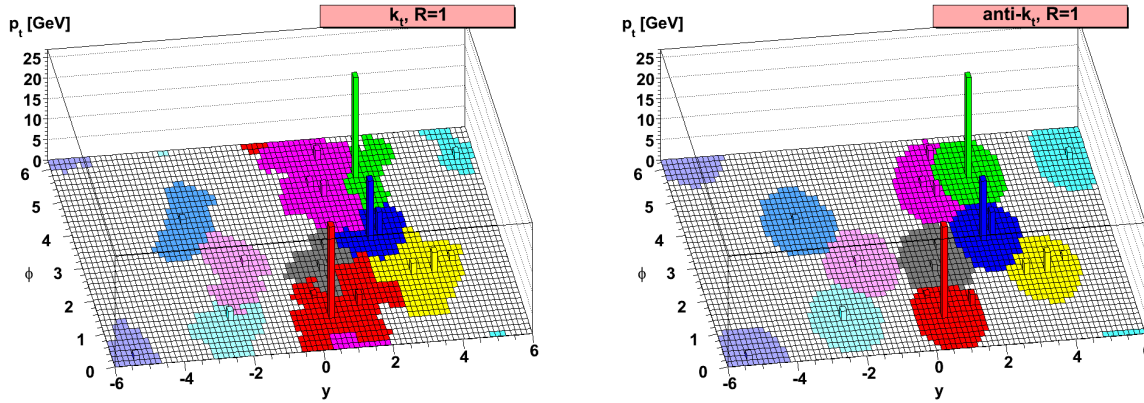


Fig. 5.4: The topologies of sequential k_T and anti- k_T jet algorithms [75]. The anti- k_T algorithm reconstructs a cone with smoother boundaries comparing to the k_T algorithm, and the cones are centralized around harder objects.

- **Particle flow (PF) Jets:** As the most sophisticated jet reconstruction algorithm, the PF jet feeds the PF objects which are described in Section 5.1.1 to the clustering algorithms.

The CMS Collaboration has centrally defined quality criteria for the selection of correctly reconstructed jets [76]. Jets in the present analysis are required to pass the *loose* identification criteria: The anti- k_T jets with a radius $R = 0.5$ should have a neutral hadron energy fraction below 99%, a neutral electromagnetic energy fraction below 99%, and at least two constituents. Furthermore, for jets in the high-pseudorapidity region $|\eta| < 2.4$, where tracking information is available, the *loose* jets should have a charged-hadron energy fraction and charged particle multiplicity greater than zero, and a charged-electromagnetic energy fraction below 99%. The spatial acceptance is restricted to $|\eta| < 4.7$ and the minimal transverse momentum is requested to be at least 10 GeV.

In order to prevent a jet related to a muon or an electron to be identified as separate entity, jets within the vicinity of good muons and electrons are rejected.

Jets satisfying the requirements described above are labeled as *good jets*.

Jet energy correction

Due to the non-uniform and non-linear energy response of the calorimeters, the reconstructed energy of jets differs from the corresponding particle energies. Moreover, electronics noise and pile-up further increase the measurement uncertainty. In order to correct the jet energy measurements, a calibration of the jet energy is required. The jet energy is corrected following a multi-step procedure referred to as Jet Energy Calibration (JEC) [77].

The following items are corrected in the present analysis:

- Removal of energy discrepancies due to PU interactions and instrumental noise (C_{PU}).
- Calibration of non-uniformity of the detector response in η plane (C_{rel}).
- Calibration of non-uniformity of the detector response with respect to the p_T (C_{abs}).

- Correcting the mismatches between data and simulation (C_{res}). It is applied only on data.

All these scale factors are composed in the JEC, which can thus be formulated as:

$$\mathcal{C} = C_{\text{PUI}}(p_T^{\text{raw}}, A_j, \rho) \cdot C_{\text{rel}}(\eta) \cdot C_{\text{abs}}(p'_T) \cdot C_{\text{res}}(p''_T, \eta) \quad (5.5)$$

where p'_T and p''_T are the jet p_T with the previous corrections applied, $p'_T = C_{\text{PUI}} \cdot C_{\text{rel}} \cdot p_T^{\text{raw}}$ and $p''_T = C_{\text{abs}} \cdot p'_T$.

Fig. 5.5 shows the JEC as a function of the jet η for $p_T = 50$ GeV on the left and as a function of the jet p_T for $\eta = 2.0$ on the right for the different jet types described in Section 5.1.6. The calorimeter jets require the largest corrections since they are formed from the energy depositions in the calorimeters directly, whereas the particle-flow jets (PF jet) need only small corrections since they profit from the additional information obtained from the tracking detector and the granularity of the ECAL exploited by the particle-flow algorithm. Moreover, the PF jets corrections are not only the smallest, but also the most uniform in both p_T and η .

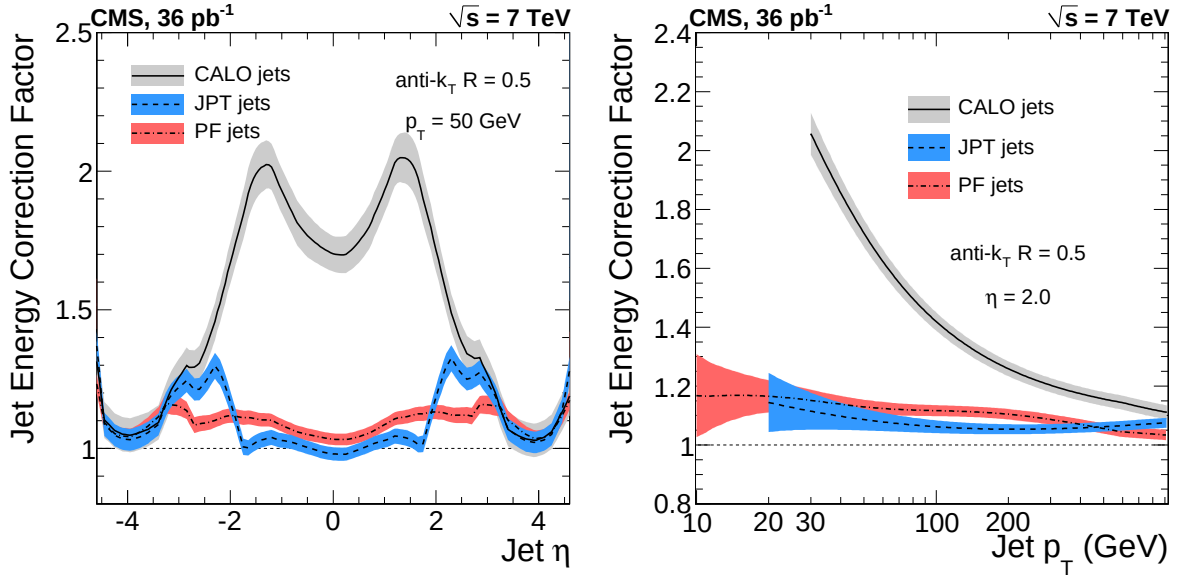


Fig. 5.5: Total jet-energy-correction factors for Calo-jets (solid line), JPT-jets (dashed line), and PF-jets (dotted-dash line) reconstructed with the anti- k_T algorithm ($R = 0.5$). The left plot gives the total JEC factors as a function of jet η for $p_T = 50$ GeV, and right plot gives the JEC factors as a function of jet p_T for $\eta = 2.0$ [77]. The corresponding uncertainties are indicated with the bands.

b-jet identification

Due to hadronization, identification of particles carrying color charge is nearly impossible in HEP detectors, yet it provides a unique opportunity to classify many rare SM and new physics processes. Only jets originating from b quarks, and to a limited extend c-quarks, can be distinguished and separated from other lighter partons due to b- and c-quarks longer lifetime and

higher mass. b-quarks (hence b-jets) can occur in many interesting new physics processes and top quark decay channels.

There are various algorithms that can be employed to identify the b-jets, using various reconstructed objects such as tracks, vertices and leptons. Features of these objects are trained with multivariate analysis techniques to obtain a high-level discriminator. This identification process is often referred to as *tagging*.

The CMS collaboration employs various algorithms to identify b-jets. These tagging algorithms can be divided into two classes with respect to the main features that are used in the identification.

The *impact parameter* is the vertical distance of the decay point (*secondary vertex*) to the interaction point (*primary vertex*). The distance of the tracking detector hits to interaction point affects the performance of the identification algorithm, hence two of these 8 hits should be present in the pixel detector. The secondary vertex is observed as an additional vertex after the primary collision vertex. b-quarks have a longer decay-length from the IP compared to other partons; and decays into harder colorless particles, creating a secondary vertex.

Identification algorithms use either one of feature sets related to the *impact parameter* and *secondary vertex*, or combine both. A more detailed description, comparison, and performance of these algorithms can be found in Ref. [78]. In this thesis, the combined secondary vertex (CSV) algorithm is employed.

The CSV algorithm follows a combined approach to the identification problem, using secondary vertices and taking advantage of the track impact parameter. It uses a multivariate analysis approach to classify b-jets and provides a discriminator as shown in Figure 5.6. With respect to the misidentification probability of lighter parton jets, various cut points on the tagging discriminator are decided by CMS. These points are named as 'loose', 'medium' and 'tight' with the corresponding misidentification probabilities of 10%, 1%, 0.1%. The following features are used in the training of the CSV algorithm: statistical significance (with respect to charm quark threshold) of the impact parameter, distance between primary and secondary vertex, invariant mass of particles associated to the secondary vertex, track multiplicity, energy fraction of particles associated with the secondary vertex compared to the energy of the jet as well as η of the tracks.

5.1.7 Missing transverse energy $\cancel{E}_T$

One of the main goals of this thesis is to discover or investigate the discovery potential of SUSY models providing *dark matter* candidates like neutralinos as LSPs, with the CMS detector. These particles, as well as very low mass, weakly interacting neutrinos, escape the detection, and hence give rise to an imbalance in the vectorial sum of the reconstructed object momenta. Therefore, having a grasp of such an observable is important for understanding underlying physics, and discriminating background from signal processes. However, the estimation of all momentum components is not possible due to undetermined longitudinal momentum components of partons. The sum of transverse momentum components is expected to be null. As a consequence, only the transverse component of energy as the transverse magnitude of momentum is considered and

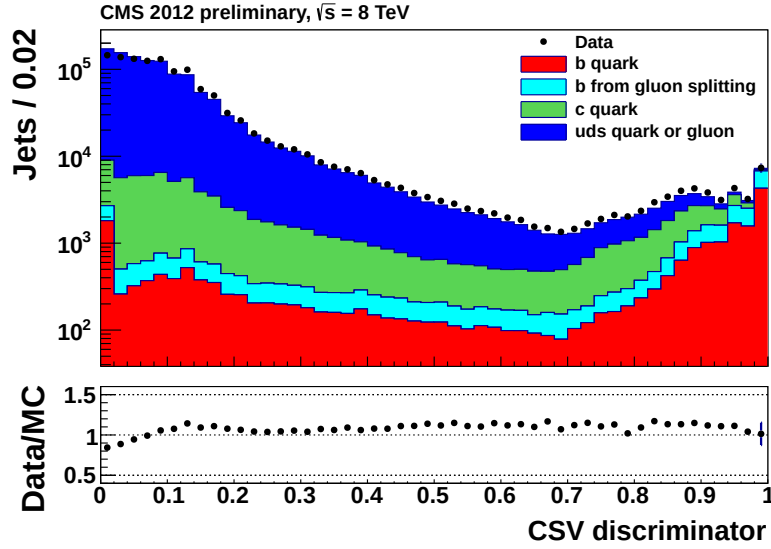


Fig. 5.6: The CSV discriminator for 8 TeV CMS b-jet identification [78]. As the discriminator approaches one, the purity of jets originated from b quarks increases.

given as:

$$\vec{\cancel{p}}_T = - \sum_{\text{all objects}} \vec{p}_T \quad (5.6)$$

$$\cancel{E}_T = \left| \sum_{\text{all objects}} \vec{p}_T \right|. \quad (5.7)$$

CMS introduces three different methods to calculate $\cancel{E}_T$:

- *Calo- $\cancel{E}_T$* : The missing transverse energy is estimated using solely the calorimeter measurements.
- *TC- $\cancel{E}_T$* : The Calo- $\cancel{E}_T$ is improved by corrections obtained from the tracking detector.
- *PF- $\cancel{E}_T$* : The PF objects are used in the calculation of $\cancel{E}_T$.

Estimation of this observable is not straightforward due to detector inefficiencies. The jet energy corrections described in Section 5.1.6 are propagated to the estimations and referred to as *type-I corrections* to the $\cancel{E}_T$. Due to rotational symmetry $\cancel{E}_T$ should be independent of ϕ of the detector axis. However, calorimeter efficiencies, as well as the beamspot displacement, and anisotropic detector responses modulate $\cancel{E}_T$ in the ϕ axis. This effect is corrected by moving the origin of the coordinate system in the transverse plane. In this thesis, the ϕ and type-I corrected PF- $\cancel{E}_T$ is used as the missing transverse energy.

5.2 Event and object selection

The present analysis considers the data sample recorded by the CMS detector during the 2012 LHC run at $\sqrt{s} = 8$ TeV with a total integrated luminosity of $\mathcal{L} = 19.5 \text{ fb}^{-1}$, fully validated by the CMS-DQM experts.

Trigger

As discussed in Chapter 3, it is not possible and in most cases not necessary to store and analyze all collision information. Therefore, a CMS data analysis starts with the selection of a suitable trigger path among a list of predefined paths. In the present analysis, events with a single isolated lepton above a certain p_T threshold are required. The L1 electron-gamma trigger energy threshold is set to 20 GeV, the objects collected within this trigger path is then fed to HLT.

Looser criteria are required for muons due to the less crowded multiplicity in the dedicated muon system, hence lower L1 and HLT thresholds for muon triggers than the ones used for electrons keeping a similar output rate. Specifically, the L1 used in this analysis fires on muon tracks with $p_T > 16$ GeV and the following HLT selects muon candidates passing isolation requirements and $p_T > 24 \text{ GeV}^1$.

Trigger efficiency measurements

The efficiencies of the single-lepton triggers are measured using the tag-and-probe method described in Ref. [79, 80]. The tag is required to pass the full offline analysis selection, have $p_T > 30 \text{ GeV}$, $|\eta| < 2.1$, and be matched to the single-lepton trigger. The probe is also required to be a good lepton and have $|\eta| < 2.1$, but the p_T requirement is relaxed to 20 GeV in order to measure the p_T turn-on curve. The tag-probe pair is required to have opposite-sign, same flavor, and an invariant mass in the range around the Z mass, namely 76–106 GeV.

The measured trigger efficiencies are displayed in Fig. 5.7. These trigger efficiencies are applied to the MC simulated events.

Lepton identification and isolation efficiency measurements

Similar to the trigger efficiency measurements, a tag-and-probe method is employed in the estimation of the identification and isolation efficiencies [79, 80]. The details of the methodology as well as the corresponding efficiencies are detailed in Appendix B.

Simulated samples

In the following, the relevant background processes are listed including information about the MC generator used for the particular process. Parton showering and hadronization are simulated with PYTHIA using the Z2* tune.

¹In terms of the CMS trigger path names, HLT_IsoMu24(_eta2p1) and HLT_Ele27_WP80 were used for muons and electrons, respectively.

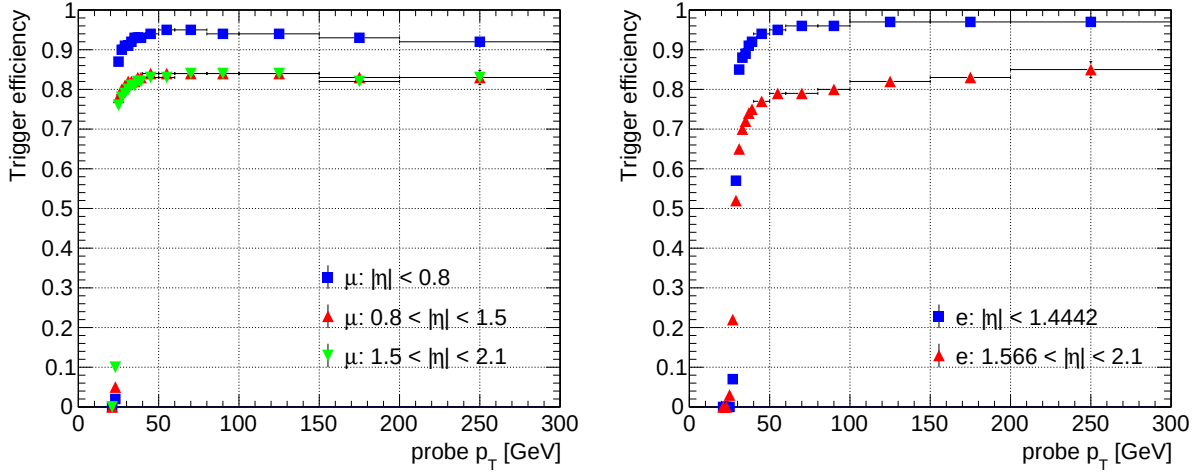


Fig. 5.7: Efficiency for the single-muon trigger on the left and single-electron trigger on the right as functions of probe-lepton p_T [79, 80], separated into several probe-lepton $|\eta|$ ranges.

$t\bar{t}$ is simulated with POWHEG, and is the main background processes. Additional samples generated with MADGRAPH are used to investigate systematic variations due to the different generation procedure, the particular choices of the top mass, matching scale, and hadronization scale.

single top is possible via s , t , and tW channel. The latter is the dominant process since both the top and $W^\pm$ can decay leptonically.

W+jets is binned in jet multiplicities.

$t\bar{t}V$ includes $t\bar{t}Z^0/\gamma^*$, $t\bar{t}W$, $t\bar{t}W^+W^-$.

di-boson consists of Z^0Z^0 , $Z^0W^\pm$, and W^+W^- ,

tri-boson consists of very low cross-section processes: $Z^0Z^0Z^0$, $Z^0Z^0\gamma^*$, $Z^0Z^0W^\pm$, $Z^0W^+W^-$, and $W^+W^-W^\pm$.

Drell-Yan is simulated with MADGRAPH, inclusive in the number of jets, but divided into two samples: $10 < M_{ll} \leq 50$ GeV and $M_{ll} > 50$ GeV.

QCD multijets is simulated with MADGRAPH, refers to events composed entirely of jets, without top quarks or W/Z bosons from the hard scattering. It is the process with the highest cross section at any hadron collider, but negligible after the preselection described in Section 5.2.2.

These samples are rearranged and grouped in a way that is more consistent with the background estimation methodology (cf. Chapter 7):

$t\bar{t} \rightarrow l\bar{l}$ dileptonic decaying $t\bar{t}$;

1l top semileptonic decaying $t\bar{t}$ and single top;

Rare di-boson, tri-boson, and $t\bar{t}V$.

Simulated signal samples

The production of a top-squark pair decaying as $\tilde{t} \rightarrow t \tilde{\chi}^0$ process is considered in the simulated signal MC samples within the T2tt framework which is described in Chapter 2. The events are generated with MADGRAPH assuming unpolarized quarks, and partons are decayed, showered and hadronized through PYTHIA with the Z2* tune. The CMS detector is simulated with the FastSim software as discussed in Chapter 3. Each signal sample is normalized to the reference cross-section for SUSY top-squark pair production calculated as a function of the $\tilde{t}$ mass at NLO approximation using PROSPINO. The parameters of the model are $m_{\tilde{t}}$ and $m_{\tilde{\chi}^0}$, whereas the mass splitting $m_{\tilde{\chi}^\pm} - m_{\tilde{\chi}^0}$ has been fixed to 5 GeV.

Top p_T reweighting

The shape of the p_T spectrum of the top quarks and antiquarks in data is softer due to the discrepancy between NLO and NNLO corrections [81–84]. Based on the observations, event scale factors have been derived and applied to the simulated $t\bar{t}$ samples used in this analysis [85].

5.2.1 Event cleaning

The noisy detector cells, and other kinds of detector and reconstruction failures can have a negative impact on the event reconstruction. These erroneously reconstructed events present themselves with a spurious momentum imbalance. Even though selecting only good lumi-sections filters out most of these erroneous measurements, analyses considering events with high $\cancel{E}_T$ requirements can still be affected.

The following procedure is followed in the present analysis to remove such events [86]:

- **Scraping veto:** at least 25% of tracks in the event are required to be high purity tracks (with the quality flag set using the number of hits on the track, its χ^2 , etc.).
- **Tracking system failure veto:** the track reconstruction algorithms may fail for events with a large number of clusters. This results with events that have calorimeter deposits without any track signatures. A very effective filter was set up to clean away these events: the scalar sum of the p_T of the tracks belonging to the good vertices, divided by the scalar sum of the p_T of all jets in the event, is found to show a clear distinction between misreconstructed and good events. A threshold at 10% on this ratio is imposed.
- **HB-HE noise filter:** The HB-HE noise is due to instrumentation issues associated with the hybrid photodiodes (HPD) and readout boxes of the HCAL, such as particles hitting the HPDs directly. It extends to the TeV scale energies and the rate for this type of noise is of

the order of several Hz at an energy threshold of 100 GeV. This filter is applied to real data events only.

- **ECAL noise filter:** EB and EE have single noisy crystals which are masked at reconstruction. Even though the multiplicity of these crystals is very low, a significant amount of energy may be lost leading to high mismeasured $\cancel{E}_T$. Using the distance between the masked cells and jets together with the energy surrounding the ECAL masked cells, it is possible to reject high $\cancel{E}_T$ events from this source. This filter is applied to the data events only.
- **ECAL- and HCAL-laser filter:** ECAL laser miscalibrations are removed selecting electrons with $p_T > 20$ GeV and at least one crystal in the supercluster with laser correction > 2 . Events containing HCAL calibration laser firing in the collision bunch-crossing were observed and rejected. These filters are applied to real data events only.
- **Beam halo filter:** secondary particles are produced in showers which are initiated by collisions of the beam with residual gas inside the LHC vacuum chamber or by interactions of the beam halo with limiting apertures. A beam halo identification algorithm has been developed exploiting timing information and hit topology in CSC, ECAL, and HCAL subsystems.
- **Anomalous ρ veto:** ρ is defined as the median of the ratio of p_T over the area of the k_t jets clustered with $R = 0.6$ in the event and it is an estimate of the pile-up energy per unit area in the η - ϕ plane [87]. It is used to subtract the pile-up contribution from the lepton isolation. Events with $\rho > 40$ GeV are rejected.
- **Calo and PF $\vec{\cancel{E}}_T$ directions consistency veto:** $\Delta\phi(\text{Calo } \vec{\cancel{E}}_T, \text{PF } \vec{\cancel{E}}_T) < 1.5$, to remove noise from the HO and from the tracker TOB-TEC boundaries.

5.2.2 Preselection

The following preselection criteria are applied on the physics objects:

Lepton selection

Signal events that are considered in this analysis are expected to be produced centrally in the CMS detector.

Driven with the trigger acceptance and corresponding efficiency, the events with one good electron (muon) with $p_T > 30(25)$ GeV and $|\eta| < 1.4442(2.1)$ are selected. The of the endcap electrons is motivated by the observation of an excess of events featuring large transverse mass, which is not understood.

Electrons and muons passing the described selection will be referred to as *selected electrons* and *selected muons* in the following. After this selection, the main background processes are vector boson and QCD multijet production.

Jet selection

The jets are required to be identified as coming from the primary interaction by the pile-up-jet identification algorithm. Jets satisfying the mentioned criteria are labeled as *selected jets*. Only events with at least three selected jets with $p_T > 30$ GeV and $|\eta| < 2.4$ are selected.

A further requiring at least one b-tagged jet among the selected jets largely reduces the QCD and the vector boson background contribution, and makes the $t\bar{t}$ pair-production the main background.

Isolated track veto

The isolated-track veto is designed to suppress the $t\bar{t} \rightarrow l\bar{l}$ -background by eliminating leptons that cannot be reconstructed as a lepton, but leave an isolated track as a signature.

In this thesis, the relative isolation of a track is defined as the ratio of the scalar sum of the p_T of charged PF candidates with $d_Z < 0.1$ cm from the primary vertex within a distance $\Delta R < 0.3$ from the track over the p_T of the considered track itself. The tracks satisfying the following conditions are called isolated track:

- $\Delta R > 0.4$ between the isolated track candidate and the selected lepton,
- $p_T > 5(10)$ GeV,
- relative track isolation $< 0.2(0.1)$.

Events with at least one isolated track having opposite sign with respect to the selected lepton are rejected.

Tau veto

After applying the isolated track veto, the majority of the remaining $t\bar{t} \rightarrow l\bar{l}$ events have at least one τ lepton. Most of these τ leptons decay hadronically. In order to further suppress the $t\bar{t} \rightarrow l\bar{l}$ background, events containing τ candidates with opposite charge and $\Delta R > 0.4$ with respect to the selected lepton are rejected.

$\cancel{E}_T$ requirement

A $\cancel{E}_T$ threshold of 80 GeV is required to further exclude QCD and the Drell-Yan backgrounds.

5.3 Variable definitions

Understanding and extracting the information from the collected data is the biggest mission of an experimental physicist. As a consequence, extracting features - or as called in experimental physics, variables - using physics intuition from theories and/or hypothesis is extremely important. With this purpose, after reconstructing the physics objects, in this section the variables that are used in the present analysis are discussed.

Table 5.1: List of variables in high end low-level-variable sets

low-level-variable set	high-level-variable set
$\cancel{E}_T$	H_T -ratio
H_T	m_{T2}^W
$p_{T,l}$	Centrality
η_l	Y
$p_{T,jet(1,2,3)}$	$\Delta\phi(W, l)$
$p_{T,b\ jet1}$	$m(l, b)$
n_{jet}	$\Delta\phi_{\min}(j_{1,2}, \cancel{E}_T)$
$n_{b\ jet}$	$\Delta r_{\min}(l, b)$
$\eta_{b\ jet1}$	Hadronic top χ^2

Physics variables are categorized with respect to their mathematical complexity as high-level and low-level variable sets. After the preselection, the main background is $t\bar{t}$. The $t\bar{t}$ events with missing lepton shows different characteristics comparing to $1l$ top background. Therefore, a different set of variables is required to efficiently eliminate these processes individually.

5.3.1 Low-level variables

The low-level-variable set consists of basic properties of the reconstructed physics objects, e.g. the transverse momentum (p_T) of these objects, and their linear combinations.

Unlike the decay chain of the background events, the signal events are expected to have one extra non-detectable particle $\tilde{\chi}^0$ in their decay chain, which brings an additional contribution to $\cancel{E}_T$. Therefore, $\cancel{E}_T$ is expected to have a strong classification power as it is demonstrated in Fig. 5.8. H_T is the scalar sum of transverse momentums of all jets, and can be considered as a measure for the hadronic activity. As discussed in Chapter 3, pseudorapidity (η) is used instead of the polar angle in beam direction, therefore, η_l represents the pseudo-rapidity of the lepton. Finally due to their involvement in the top quark related decay processes, the transverse momentum of the leading b -jet, and the multiplicity of these b -quark jets $n_{b\ jet}$ are also used. A selection of these variables is presented in Fig. 5.8.

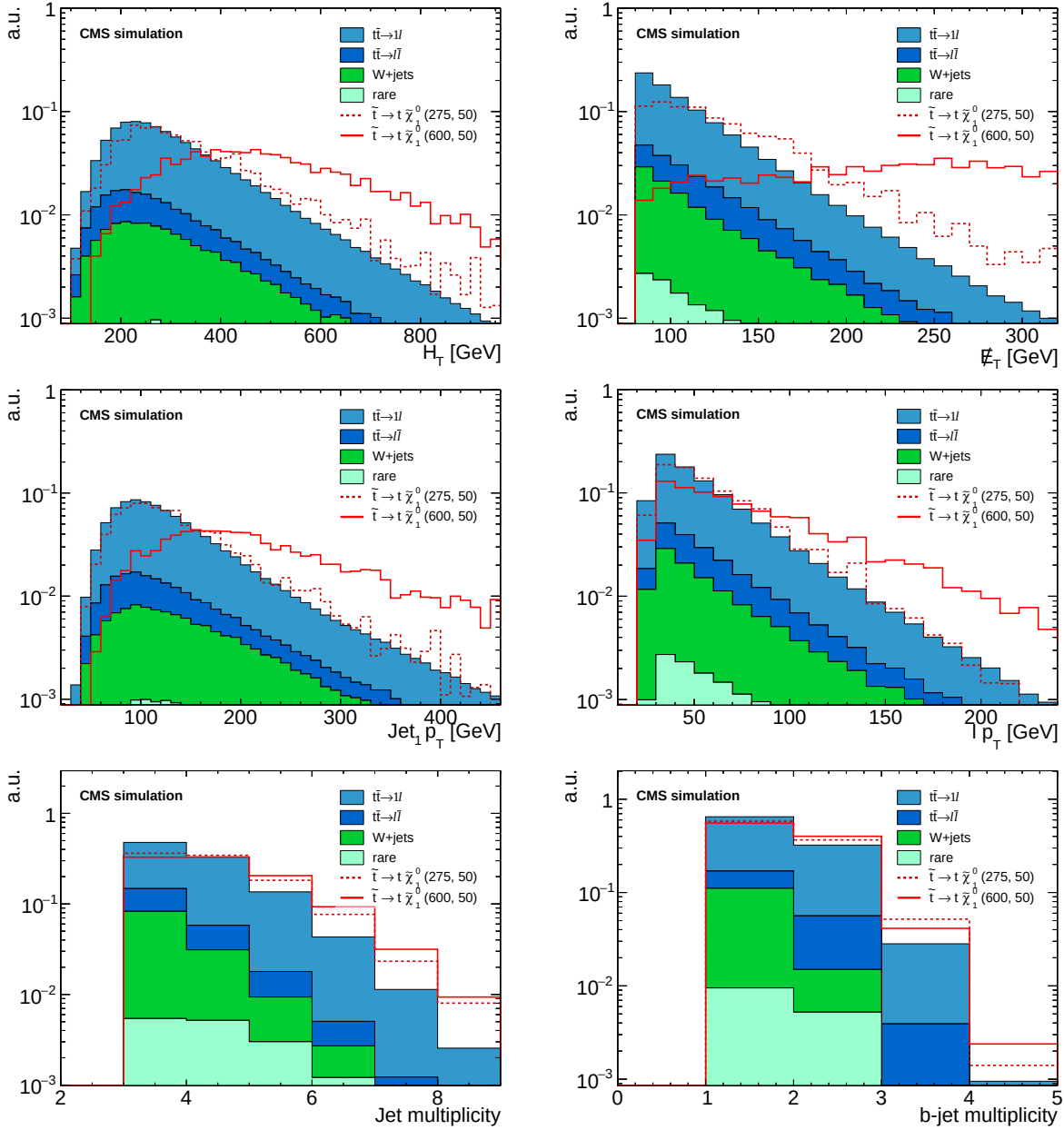


Fig. 5.8: A sub-set of low-level variables that are used in the present analysis. The y-axis shows the number of events (normalized to unity), and the x-axis shows the variable value for given bin. Two signal samples generated with different $\Delta M \doteq m_{\tilde{t}} - m_{\tilde{\chi}_0^0}$ are presented with the SM background simulations. Closer to $\Delta M \simeq m_t$, the signal distributions show similar characteristics to the background simulation. The simulated distributions from different background processes are stacked, while the signal distributions are individually presented.

5.3.2 High-level variables

The high-level-variable set includes complex variables constructed from the low-level variables using physical intuition to improve the classification performance.

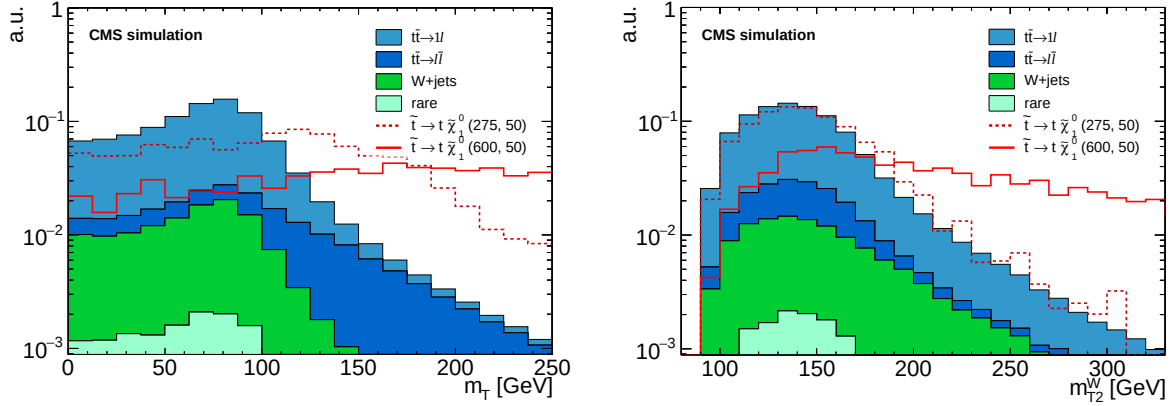


Fig. 5.9: the transverse mass and m_{T2}^W that are used in the present analysis. The y-axis shows the number of events (normalized to unity), and the x-axis shows the variable value for given bin. Closer to $\Delta M \doteq m_{\tilde{t}} - m_{\tilde{\chi}^0} \simeq m_t$, the signal shows similar characteristics to the background simulation. The simulated distributions from different background processes are stacked, while the signal distributions are individually presented.

Transverse mass m_T

The transverse mass of a mother particle for two body decay can be written as:

$$m_T = \sqrt{m_a^2 + m_b^2 + 2(E_{T,a} E_{T,b} - \vec{p}_{T,a} \cdot \vec{p}_{T,b})}, \quad (5.8)$$

where a and b are the decay products. The transverse mass of W boson given the decay $W \rightarrow l + \nu$, and neglecting the mass of lepton becomes:

$$m_T = \sqrt{2 p_{T,l} \cancel{E}_T (1 - \cos \Delta\phi(l, \cancel{E}_T))}. \quad (5.9)$$

In the search for W boson, the UA1 collaboration used the transverse component of this quantity to assign a minimum mass for the W boson [88], since this quantity is strictly smaller than m_W .

The transverse mass provides a process dependent discrimination, thus can be used to eliminate non-SM signals. In an ideal reconstruction scenario, the m_T distribution of $W + \text{jets}$ must have an endpoint at the m_W . A distribution with similar characteristic can be argued for the $t\bar{t} \rightarrow \ell + \text{jets}$ background as well, since the momentum of the hadronically-decaying leg is shared by jets and $\cancel{E}_T$, hence a small tail is expected due to the $\cancel{E}_T$ contribution of hadronically-decaying leg. On the other hand, $t\bar{t} \rightarrow \ell\bar{\ell}$ background with a missing lepton loses both outlying lepton energy for a given W decay and the correlation information.

m_T provides an excellent description of the kinematic topology, and thereby, an outstanding discrimination power as demonstrated in Fig. 5.9. Therefore, this variable is reserved for validating the background simulations and estimation. A threshold cut of 100 GeV is employed to remove a significant background contribution while keeping the signal efficiency in reasonably high value.

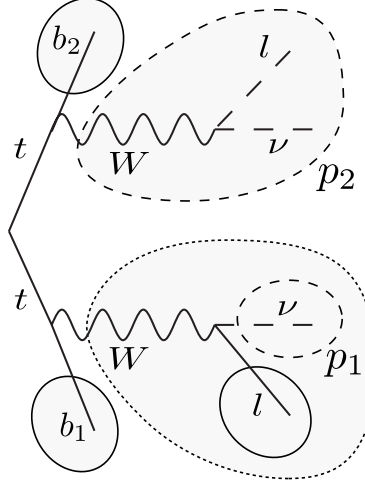


Fig. 5.10: $t\bar{t} \rightarrow l\bar{l}$ background with a missing lepton leg circled with dashed lines and denoted as p_2 . m_{T2}^W is used to eliminate this background by reconstructing the minimum transverse mass of the mother particle [89].

m_{T2}^W

$t\bar{t} \rightarrow l\bar{l}$ with one missing lepton shows a signal-like behavior due to a high $\cancel{E}_T$, and the lacking m_T reconstruction information. A more elaborate attempt of the signal discrimination can be made by exploiting the minimum transverse mass of the mother particle which is reconstructed by using the maximum kinematical information [89]. Such a quantity is given as:

$$m_{T2}^W = \min(m_y), \quad (5.10)$$

where m_y satisfies:

$$\begin{aligned} \vec{p}_{T,1} + \vec{p}_{T,2} = \cancel{E}_T, \quad p_1^2 = 0, \quad (p_1 + p_l)^2 = p_2^2 = m_W^2, \\ (p_1 + p_l + p_{b1})^2 = (p_2 + p_{b2})^2 = m_y^2, \end{aligned} \quad (5.11)$$

where sub-index 2 denotes objects related to the leg with the missing lepton as depicted in Fig. 5.10.

The discrimination power of this variable is especially powerful where the mother particle's mass is significantly different for two classes i.e. $\Delta M > m_t$ as can be observed in Fig. 5.9.

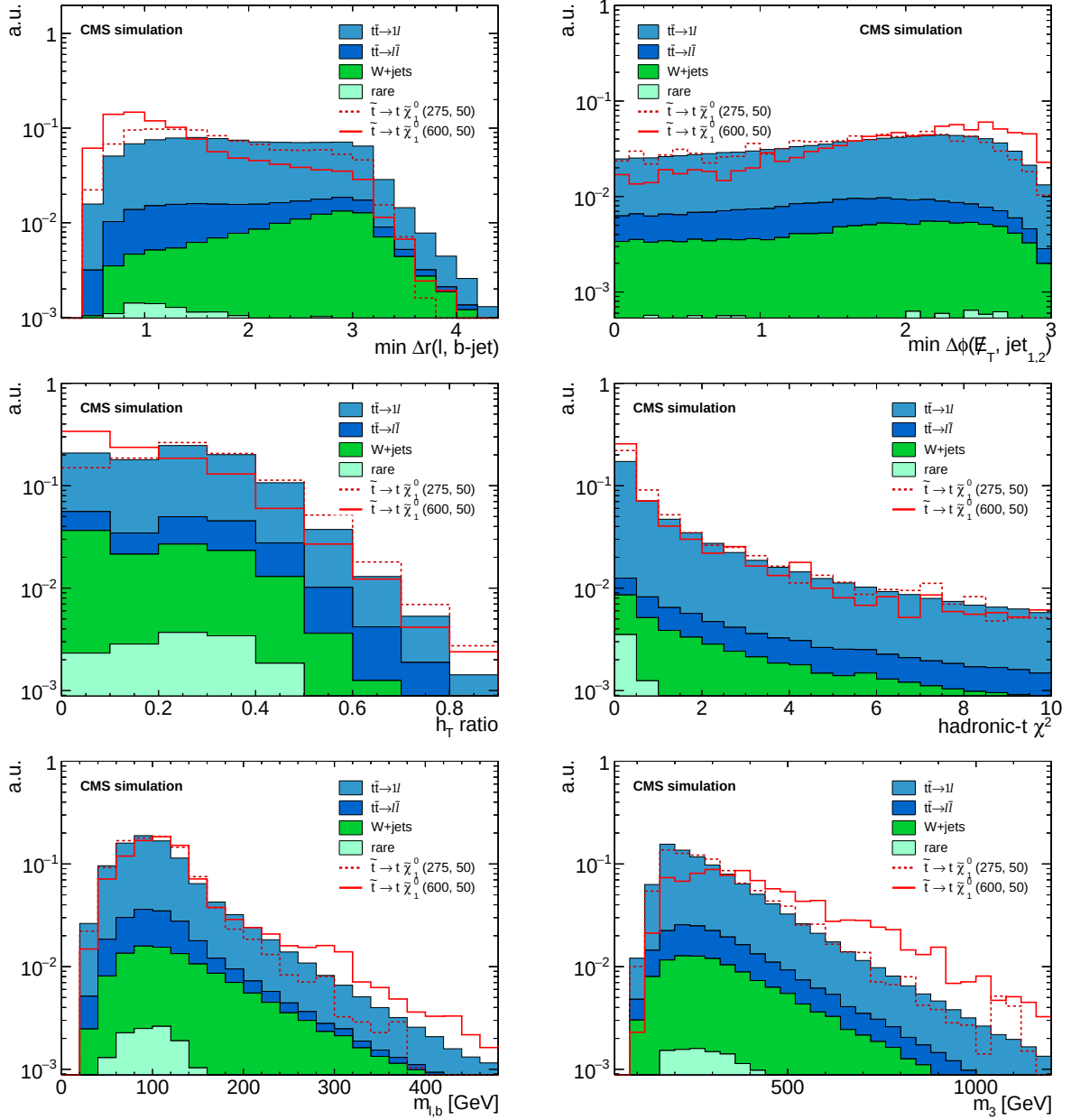


Fig. 5.11: A sub-set of high-level variables that are used in the present analysis. The y-axis shows the number of events (normalized to unity), and the x-axis shows the variable value for given bin. Two signal samples generated with different $\Delta M \doteq m_{\tilde{t}} - m_{\tilde{\chi}_0^0}$ are presented with the SM background simulations. Closer to $\Delta M \simeq m_t$, the signal distributions show similar characteristics to the background simulation. The simulated distributions from different background processes are stacked, while the signal distributions are individually presented.

Centrality

This topological variable exploits the angular dependency of the event yields of different processes along the beam-line axis. The signal events tend to be localized on the transverse plane, whereas the background processes are more uniformly distributed over the azimuthal angle. The centrality of an event is given by:

$$\text{Centrality} = \frac{\sum_{\text{jets}, l} p_T}{\sum_{\text{jets}, l} p} \quad (5.12)$$

Hadronic-top χ^2

A fit on the single-lepton events can be performed to test the presence of a hadronic leg of the $1l$ top background. The χ^2 value for such fit can be given by:

$$\chi^2 = \frac{(m_{j_1 j_2 j_3} - m_t)^2}{\sigma_{j_1 j_2 j_3}^2} + \frac{(m_{j_1 j_2} - m_W)^2}{\sigma_{j_1 j_2}^2}, \quad (5.13)$$

where m_t is the mass of t -quark and m_W is the mass of W-boson, $m_{j_1 j_2 j_3}$ and $m_{j_1 j_2}$ are the masses of jets minimizing Eq. 5.13 requiring at least one of the jets to appear in the term $m_{j_1 j_2 j_3}$ to be a b -jet, and $\sigma_{j_1 j_2 j_3}^2$ and $\sigma_{j_1 j_2}^2$ are the uncertainties on these masses calculated from the jet energy resolutions

Y and H_T -ratio

The variable Y is often referred to as $\cancel{E}_T$ significance, and is defined by:

$$Y = \frac{\cancel{E}_T}{\sqrt{H_T}}. \quad (5.14)$$

This value is usually proportional to the boost of visible objects, which are mostly due to the jet mismeasurements.

H_T -ratio presents a similar quantity in a topological aspect. It is the normalized hadronic activity in the hemisphere of $\cancel{E}_T$, and is presented in Fig. 5.11.

$m(l, b)$, m_3 , and $\Delta r_{\min}(l, b)$

Since it is not possible to estimate the missing energy coming from each of the mother particles in both signal and background processes, the mass of the b -jet and the closest lepton provides a feature proportional to the energy of the top quark for the given leg.

m_3 is the invariant mass of the three leading jets, which is observed to provide discrimination as shown in Fig. 5.11.

Furthermore the radial separation between these variables, $\Delta r_{\min}(l, b)$, provides an additional topological discriminator especially effective in suppressing the $W + \text{jets}$ background as shown in Fig. 5.11.

$\Delta\phi(W, l)$

In addition to the mass reconstruction of the W boson, the angle between the W boson and the lepton can be exploited for the classification. In the background processes the source of missing energy is the neutrinos originated from the W boson decay, whereas in the signal processes neutralinos also contribute to the missing energy. This variable is only employed in the case study to evaluate the performance of the machine learning algorithm in Chapter 6.

Support Vector Machines

Contents

6.1 A brief introduction to machine learning	84
6.2 Support vector machines	85
6.2.1 Linearly separable distributions	86
6.2.2 Overlapping distributions	89
6.2.3 Non-linear distributions	90
6.2.4 Computing SVM	92
6.2.5 Probability output	93
6.2.6 Hyper-parameter tuning	94
6.3 Binary Decision Trees	96
6.3.1 Boosting	97
6.4 Classifier implementations	98
6.4.1 Method of comparing performance of classifiers	98
6.4.2 SVM hyper-parameter optimization algorithm	99
6.4.3 Parallelization	100
6.5 Case studies	101
6.5.1 Performance comparison on a toy model	101
6.5.2 Third generation supersymmetric partner search	102
6.6 Summary	107

The idiom *looking for a needle in a haystack* does not suffice to describe the challenge of discriminating the expected SUSY *signal* events from the SM *background* at the CMS experiment. Due to the low cross-section of the SUSY signatures, the rarity of these events requires an intensive and long data taking period even at the luminosity frontier of the current high energy physics accelerators. Moreover, some SUSY models generate signal events that may have similar characteristics to those of SM background events, as discussed in Chapters 2 and 5. These events

can only be classified efficiently by taking advantage of all information obtained from the detector and using physics intuition. Therefore, most of the variables described in Chapter 5 need to be analyzed simultaneously. This cannot be accomplished using simple cut-based techniques, where each variable is considered separately, or by ad hoc combinatorial cut-based techniques that are limited by computational resources and time. *Machine learning* algorithms provide a computationally efficient, effective and generalized solution to this *classification* problem.

In this thesis, *Support Vector Machines (SVM)* are chosen to perform the classification of SUSY events in the data obtained from the CMS detector. A new software package, the SVM high energy physics interface (SVM-HINT), is introduced using a well established SVM implementation, LIB-SVM, to interface high energy physics classification problems. The performance of the SVM-HINT is compared to the *boosted binary decision trees (BDT)* of the TMVA framework [90].

In the first section, a brief introduction to machine learning (ML) and definition of machine learning concepts are given. It is followed by a detailed discussion of the SVM and a brief description of the BDT formalism. Finally, two different classifier implementations are tested and compared using two different case studies.

6.1 A brief introduction to machine learning

Machine learning cannot be introduced without the definition of learning. The concept of statistical learning can be described as guessing a target function f such that for given vector spaces of all possible inputs X and outputs Y ,

$$f : X \rightarrow Y \quad (6.1)$$

for a sample defined as

$$(\vec{x}_1, y_1), (\vec{x}_2, y_2), \dots, (\vec{x}_i, y_i), \dots, (\vec{x}_N, y_N) \quad (6.2)$$

where y_i for each $\vec{x}_i$ is either explicitly given or implicit to a *learner*. Using the sample or by *training* on the sample, the learner proposes an hypothesis g for f . If we define the amount of elements of a sample used by a learner as experience $\mathbf{E}$ and the task $\mathbf{T}$ of a machine as obtaining g with a performance $\mathbf{P}$, following the definition given by T. M. Mitchell [91], the machine learning can be defined as an autonomous machine process where the system increases $\mathbf{P}$ at a particular $\mathbf{T}$ with higher $\mathbf{E}$. If the possible values of y_i are discrete (and known beforehand), the task is called a classification problem. The case where y_i for each $\vec{x}_i$ is explicitly given to the learner is referred to as *supervised learning*.

One of the first *formal* classification algorithms was developed by R. A. Fisher [92] for classifying different species, solving a taxonomic problem. Fisher introduced a linear decision function to classify species and realized that the probability of misclassification of the algorithm relates to the small size of the given sample. Later Rosenblatt [93] proposed a learning machine called *perceptrons* which was inspired by neurodynamics. The perceptron learning algorithm assigns a weight for each feature¹ of the given sample. These weights are iteratively updated to minimize the misclassification rate in the training sample. The perceptron algorithm is the starting point

of the popular learning algorithm, the *neural network*. Furthermore, the concept of adaptively adjusting weights according to the misclassification rate is used in other more modern machine learning algorithms. Since the introduction of perceptrons, many different algorithms have been devised. In the following section two different algorithms are discussed, namely support vector machines, and binary decision trees with boosting.

The search for SUSY at the CMS detector can be described as a binary classification problem into SUSY signal events and SM background events. This classification can be performed by supervised machine learning algorithms which are trained on simulated MC samples.

6.2 Support vector machines

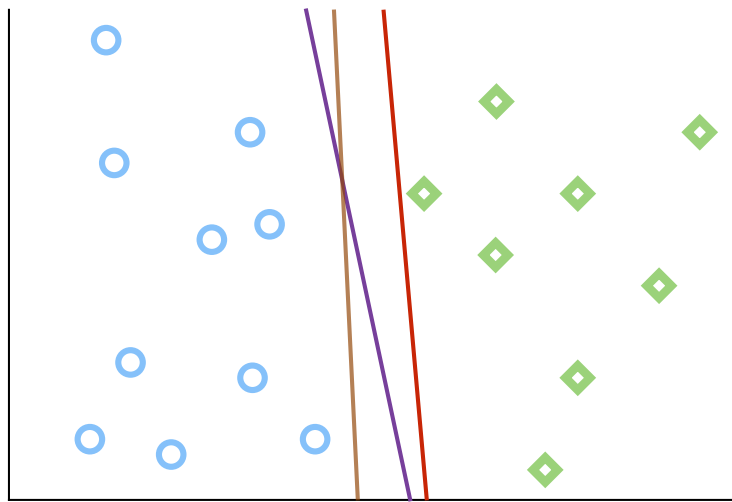


Fig. 6.1: The events represented by blue circles belong to the first class ($y = -1$), whereas the green rectangles belong to the second class ($y = 1$). The lines give examples of possible separation lines.

How can a unique separation plane that separates instances of two distinct classes be defined? In a small sample with two features, separating two distinct classes might seem trivial. However if Fig. 6.1 is considered, there exists infinitely many *lines* which can separate two distributions. For this particular problem, the question evolves to *which is the optimum line that separates two classes?* Vapnik, Lerner and Chervonenkis [94, 95] in 1964, introduced the *generalized portrait method* and answered the question by defining an optimal *hyperplane*, which maximizes the margin between two classes for the linearly separable case. This method was the first step towards support vector machines. In this thesis, for consistency and simplicity, the generalized portrait method is also referred to as support vector machines. In following sections, a throughout discussion of the SVM algorithm construction is given. In the simplest, linearly separable case, the SVM

¹In this chapter the term *feature* is interchangeably used with the term *variable*. The ambiguity arises from different conventions followed in Statistics and HEP. Mathematically, a *feature* is not a variable for a given problem, since an observable of a given particle, for instance p_T of a particle, is a constant. However, in HEP, a *variable* is considered as a property of an object which can change from an event to an another event.

detects support vectors on the boundary of the separation margin and maximizes the distance using the method of Lagrange multipliers. Later in 1995 Cortes and Vapnik [96] generalized the algorithm by allowing misclassified points (soft margin), so that the support vector network can be also successfully used in the classification of overlapping distributions. Finally, by transforming *input vectors* to *feature vectors*, the SVM can classify non-linear distributions as well.

6.2.1 Linearly separable distributions

The SVM uses hyperplanes to separate the elements of a sample with different class memberships. To construct an optimal hyperplane, the margin between the different classes are maximized using a smaller subsample of the training data, which are called support vectors. The training sample defined in Eq. 6.2 can be written for the two class problem as:

$$(\vec{x}_1, y_1), (\vec{x}_2, y_2), \dots, (\vec{x}_i, y_i), \dots, (\vec{x}_N, y_N) \quad y_i \in \{-1, 1\}. \quad (6.3)$$

The support vectors lie on the separation boundaries for this given training set. The linear separability condition ensures the existence of the following inequalities for each class

$$\begin{aligned} \vec{w} \cdot \vec{x} + b &\geq 1 & \text{where } y_i = 1, \\ \vec{w} \cdot \vec{x} + b &\leq -1 & \text{where } y_i = -1, \end{aligned} \quad (6.4)$$

where $\vec{w}$ is the weight vector orthogonal to the optimal hyperplane and b is a scalar bias parameter. The right-hand sides of the inequalities can be any scalar real value. For simplicity, the weight vectors and the bias term are chosen such that the right-hand sides of the inequalities are identical to y_i of the given class. Equation 6.4 can be rewritten in a more compact form

$$y_i(\vec{w} \cdot \vec{x} + b) - 1 \geq 0 \quad (6.5)$$

The equality case of Eq. 6.5 is satisfied by the support vectors, i.e.

$$y_i(\vec{w} \cdot \vec{x}_s + b) - 1 = 0. \quad (6.6)$$

Figure 6.2 shows two-dimensional linearly-separable distributions, where support vectors are forming two boundary lines of the margin between the two classes. The figure also visualizes the fact that once the support vectors are obtained, without taking into account all training data, the support vectors are sufficient to construct the optimal hyperplane.

The unique hyperplane is selected as

$$\vec{w}_0 \cdot \vec{x} + b_0 = 0. \quad (6.7)$$

In order to maximize the margin ρ between classes:

$$\max \rho(w, b) = \max \left(\min_{(y=1)} \frac{\vec{x} \cdot \vec{w}}{|\vec{w}|} - \max_{(y=-1)} \frac{\vec{x} \cdot \vec{w}}{|\vec{w}|} \right). \quad (6.8)$$

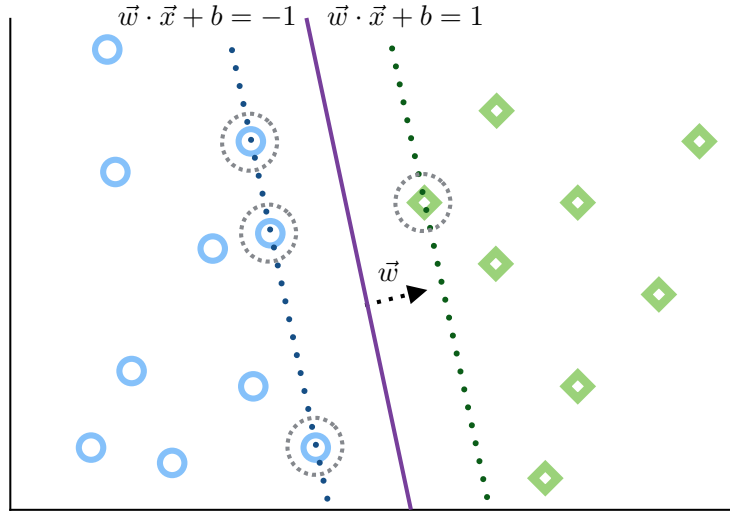


Fig. 6.2: The events represented by blue circles belong to the first class ($y = -1$), whereas the green rectangles belong to the second class ($y = 1$). The dashed lines represent the maximum margin boundaries, and the corresponding support vectors are marked by dashed circles. From all possible hyperplanes dividing the two samples, the one with the largest margin is chosen.

This equation is a restatement of the problem. In the given choice of weight vector and bias parameter, the right-hand side of the equation gives the distance between the closest elements of the class $y_i = 1$ and $y_i = -1$ to the margin (or to the support vectors). Thus, the equation maximizes the distance of the closest points and thereby the margin between distributions. The reason for taking the maximum margin instead of any of the arbitrary choices is given formally using the Vapnik - Chervonenkis dimension as described in the reference [97], but this proof is beyond the scope of this thesis. A simple and convincing argument can be provided as following: assuming that the margin is taken smaller than maximum value, hence the hyperplane is closer to class $y_i = -1$ in Fig. 6.2. Since the hyperplane is closer to the boundary, any fluctuation on the elements of $y_i = -1$ will have a greater chance of being misclassified. To minimize the misclassification probability, an optimal hyperplane satisfying the maximum margin is required.

Using equation 6.5, the condition for the unique optimal hyperplane can be rewritten as

$$\max \frac{2}{|\vec{w}|} = \frac{2}{|\vec{w}_0|} \quad (6.9)$$

or

$$\min \frac{1}{2} |\vec{w}|^2 = \frac{|\vec{w}_0|^2}{2}. \quad (6.10)$$

Equation 6.10 can be optimized using the method of Lagrange multipliers. The Lagrangian can be constructed by taking into account the constraint from Eq. 6.6 as

$$\mathcal{L}(\vec{w}, \vec{x}, \alpha) = \frac{1}{2} |\vec{w}|^2 - \sum \alpha_i [y_i (\vec{w} \cdot \vec{x} + b) - 1]. \quad (6.11)$$

This is the so-called *dual* Lagrangian corresponding to the *primal* Eq. 6.10 subject to constraint

presented in Eq. 6.6. To minimize the margin, the Lagrangian should be stationary with respect to w and b at the hyperplane:

$$\left. \frac{\partial \mathcal{L}}{\partial \vec{w}} \right|_{\vec{w}=\vec{w}_0} = 0 = \vec{w}_0 - \sum_{i=1}^N \alpha_i y_i \vec{x}_i, \quad (6.12)$$

$$\left. \frac{\partial \mathcal{L}}{\partial b} \right|_{b=b_0} = 0 = \sum_{i=1}^N \alpha_i y_i. \quad (6.13)$$

An expression for $\vec{w}_0$ is obtained from Eq. 6.12

$$\vec{w} = \sum \alpha_i y_i \vec{x}_i, \quad (6.14)$$

by substituting Eq. 6.14 to Eq. 6.11 and using Eq. 6.13, the following functional can be obtained

$$\mathcal{W}(\vec{\alpha}) = -\frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j (\vec{x}_i \cdot \vec{x}_j) + \sum \alpha_i. \quad (6.15)$$

This functional, however, is only valid under the constraints posed by Eq. 6.6. Generalizing the Lagrangian constructed in Eq. 6.11 for the constraint presented in Eq. 6.5 and meanwhile preserving *strong duality* requires a specific set of conditions. These conditions are named after W. Karush, H. W. Kuhn, and A. W. Tucker [98, 99]. The functional can be maximized with respect to α , at the same optimal point minimizing it with respect to $\vec{w}$ and b , given that it follows the Karush-Kuhn-Tucker (KKT) conditions:

$$\begin{aligned} \vec{\alpha} &\geq 0, \\ \alpha_i [y_i (\vec{w} \cdot \vec{x}_i + b) - 1] &= 0, \quad i = 1, \dots, N. \end{aligned} \quad (6.16)$$

An important consequence of the KKT conditions is that only support vectors are contributing to the maximization problem since all non-trivial solutions (i.e. $\alpha_i \neq 0$) requires Eq. 6.7 to hold. For an unknown sample element $\vec{u}$, the linear decision rules can be written as

$$\vec{w}_0 \cdot \vec{u} + b_0 \geq 0 \text{ for } y = 1 \quad \text{or} \quad \vec{w}_0 \cdot \vec{u} + b < 0 \text{ for } y = -1. \quad (6.17)$$

If we insert the equation 6.12 to SVM's decision function:

$$g(u) = \text{sign} \left(\sum \alpha_s y_s \vec{x}_s \cdot \vec{u} + b_0 \right) \quad (6.18)$$

This equation is the SVM's hypothesis for the target function. SVM, similar to perceptrons, has a linear decision function, moreover it ensures the maximum margin. A very important result that can be deduced from the equations 6.15 and 6.18 is that both the minimization and the selection criteria depend on the scalar-product of $\vec{x}_i \cdot \vec{x}_j$ and $\vec{x}_i \cdot \vec{u}$. This feature significantly reduces the computation time of the SVM algorithm as discussed in Sec. 6.2.4.

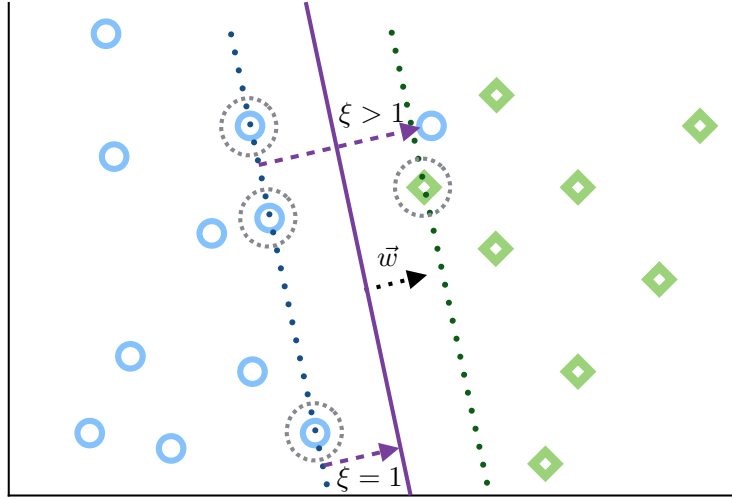


Fig. 6.3: The events represented by blue circles belong to first class ($y = -1$) whereas the green triangles belong to second class ($y = 1$). The dashed lines represent the maximum margin boundaries, and corresponding support vectors are circled by dashed lines.

6.2.2 Overlapping distributions

The method described up to this point works only for classification of two linearly separable distributions. Overlapping distributions require a different treatment. By allowing misclassification, the hard margin optimal hyperplane of SVM can be modified to a soft margin hyperplane. However, freely allowing misclassified points would decrease the classification accuracy of the SVM, hence they are required to be penalized. This requirement can be satisfied by assigning a slack variable (ξ_i) to each training point, which is linearly proportional to the distance of the element to the margin boundary of the given sample. Therefore, ξ updates the constraint given in Eq. 6.5 such that:

$$y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1 - \xi_i, \quad (6.19)$$

On the optimal hyperplane, ξ_i is equal to 1, as shown in Fig. 6.3), and when it is misclassified it takes a value larger than 1. Thus in this context, the margin is described as

- hard-margin when $\xi_i = 0$,
- soft-margin when $\xi_i \geq 0$.

The minimization problem given in Eq. 6.10 takes the following form:

$$C \sum_i \xi_i + \frac{1}{2} |\vec{w}|^2, \quad (6.20)$$

where C is the regularization parameter of the slack variables. The Lagrangian corresponding to this problem becomes

$$\mathcal{L} = \frac{1}{2} |\vec{w}|^2 + C \sum_i \xi_i - \sum_i \alpha_i [y_i(\vec{w} \cdot \vec{x} + b) - 1 + \xi_i] - \sum_i \beta_i \xi_i, \quad (6.21)$$

and the stability condition with respect to ξ gives the relation between the Lagrange multipliers:

$$\left. \frac{\partial \mathcal{L}}{\partial \xi_i} \right|_{\xi_i=1} = C - \alpha_i - \beta_i. \quad (6.22)$$

This optimization problem is subject to a larger set of KKT conditions which are given as:

$$\begin{aligned} \alpha_i &\geq 0, \\ \beta_i &\geq 0, \\ \xi_i &\geq 0, \\ \alpha_i[y_i(\vec{w} \cdot \vec{x}_i + b) - 1 - \xi_i] &= 0, \\ \beta_i \xi_i &= 0. \end{aligned} \quad (6.23)$$

where $i = 1, \dots, N$. The Lagrangian takes the same form as in Eq. 6.15, but with a different set of constraints. The regularization parameter does not explicitly appear in the maximization problem. However, considering the stability conditions, it adds a further constraint (an upper bound) to α_i :

$$C \geq \alpha_i \geq 0. \quad (6.24)$$

The problem is restored to the hard-margin case in the limit of $C \rightarrow \infty$.

6.2.3 Non-linear distributions

Non-trivial models considered for new physics searches are typically entangled with the background processes. These intricate distributions cannot be separated by linear hyperplanes discussed so far. In most cases, the SVM's linear decision function is insufficient to handle these problems even with the introduction of the soft-margin concept. However, the basic foundation of the SV machinery relies on these linear decision functions. This dilemma can be resolved by mapping the input space X to a higher dimensional feature space Z where the distributions are linearly separable, using the transformation:

$$\vec{x}_i \rightarrow \vec{\phi}(\vec{x}_i), \quad (6.25)$$

The optimization steps that we have discussed before can be repeated in the Z -space. As it is mentioned before, the optimization problem (the functional and the linear decision function) depends solely on the inner product of the $\vec{x}_i \cdot \vec{x}_j$ or in this case $\vec{\phi}(\vec{x}_i) \cdot \vec{\phi}(\vec{x}_j)$. Transforming the input space to a higher dimensional feature space is not always computationally effective or possible (for example mapping to an infinite dimensional feature space). If the existence of the space Z can be guaranteed, the inner products in this space can be calculated directly instead of mapping each individual element. This simplification can be upgraded to an advancement with the inclusion of infinite dimensional transformations. According to the Hilbert-Schmidt theory, a symmetric kernel function ($K(\vec{u}, \vec{v})$ which then can be expanded as $\sum \lambda_i \vec{\phi}(u) \cdot \vec{\phi}(v)$) can be used instead of the scalar-products of ϕ functions. By substituting this kernel, the functional can be

rewritten as

$$\mathcal{W} = -\frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j K(\vec{x}_i, \vec{x}_j) + \sum \alpha_i. \quad (6.26)$$

This substitution is referred to as *kernel trick*. There are certain conditions that need to be satisfied by the kernel to ensure the existence of the feature space Z . The kernel should be symmetric i.e. $K(\vec{x}_i, \vec{x}_j) = K(\vec{x}_j, \vec{x}_i)$ and positive semi-definite. These conditions are known as Mercer's conditions [100].

Even though these conditions provide a simple theoretical framework, there is no general way of obtaining kernel functions. Each proposed function should be tested individually for the Mercer's conditions.

Radial Basis Function kernel

In this thesis, a widely used *radial basis function (RBF)* kernel of the form

$$K_{RBF}(\vec{x}_i, \vec{x}_j) = e^{-\gamma|\vec{x}_i - \vec{x}_j|^2} \quad \gamma > 0 \quad (6.27)$$

is employed. The parameter γ adjusts the width of the kernel. In principle, no further information is required to employ this kernel in the SVM. To give a better understanding of the RBF kernel, Eq. 6.27 can be expanded for single dimension case

$$K_{RBF}(x_i, x_j) = e^{-\gamma|x_i - x_j|^2} = e^{-\gamma|x_i|^2} e^{-\gamma|x_j|^2} e^{2\gamma x_i \cdot x_j}, \quad (6.28)$$

$$= e^{-\gamma|x_i|^2} e^{-\gamma|x_j|^2} \left(\sum_{n=0}^{\infty} \sqrt{\frac{(2\gamma)^n}{n!}} (x_i)^n \cdot \sqrt{\frac{(2\gamma)^n}{n!}} (x_j)^n \right). \quad (6.29)$$

Thus the RBF kernel is a result of an infinite dimensional transformation, and can be given as:

$$\vec{\phi}(x_i) = e^{-\gamma|x_i|^2} \left(\sum_{n=0}^{\infty} \sqrt{\frac{(2\gamma)^n}{n!}} (x_i)^n \right). \quad (6.30)$$

The resulting transformation might not be meaningful in the context of SV machinery, and it is definitely not necessary due to Mercer's conditions, and Hilbert-Schmidt theorem. By considering the use of such transformations in *image processing*, an analogy can be drawn for motivating the use of RBF kernel in HEP problems. The transformation represents an iterative-Laplacian of Gaussian (LoG) filter. The LoG filter is employed to detect and separate the objects in image processing. Due to low light conditions and noise from the imaging sensors, the features of objects with rather smooth distributions become distorted. Furthermore, for overlapping images the feature edges (tails) are required to be amplified and emphasized. The Gaussian filter is used to reduce distortion (noise and fluctuations) of the underlying smooth distributions by eliminating the high frequency fluctuations, while the Laplacian filter is employed for improving the edge detection. In Chapter 5 (for 8 TeV pp collisions) and Section 6.5.2 the variable distributions for SM background and SUSY signal events which originate from smooth (i.e. not step functions) PDFs

are presented. Similar to the image processing case, in HEP the underlying smooth distributions are distorted by noise due to statistical and systematic factors and the goal is to classify these two classes. Thus, the RBF kernel is expected to perform effectively in new physics searches similar to their deployment for image processing.

6.2.4 Computing SVM

The SVM provides a linear decision function which depends solely on support vectors. However, determination of the Lagrange multipliers requires training over whole sample. These values are calculated with quadratic programming techniques. The functional can be rewritten in matrix representation as

$$\mathcal{W}(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^T \mathbf{1} - \frac{1}{2} \boldsymbol{\alpha}^T \mathbf{D} \boldsymbol{\alpha}, \quad (6.31)$$

where $\boldsymbol{\alpha}^T$ represents the transpose of $\boldsymbol{\alpha}$, and $\mathbf{D}$ is an $N \times N$ matrix with

$$D_{ij} = y_i y_j K(\vec{x}_i, \vec{x}_j). \quad (6.32)$$

In matrix notation, the maximization of the functional can be performed as

$$\max_{\boldsymbol{\alpha}} -\frac{1}{2} \boldsymbol{\alpha}^T \underbrace{\begin{bmatrix} y_1 y_1 K(\vec{x}_1, \vec{x}_1) & y_1 y_2 K(\vec{x}_1, \vec{x}_2) & \cdots & y_1 y_N K(\vec{x}_1, \vec{x}_N) \\ y_2 y_1 K(\vec{x}_2, \vec{x}_1) & y_2 y_2 K(\vec{x}_2, \vec{x}_2) & \cdots & y_2 y_N K(\vec{x}_2, \vec{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ y_N y_1 K(\vec{x}_N, \vec{x}_1) & y_N y_2 K(\vec{x}_N, \vec{x}_2) & \cdots & y_N y_N K(\vec{x}_N, \vec{x}_N) \end{bmatrix}}_{\text{quadratic maximization coefficients}} \boldsymbol{\alpha} + \boldsymbol{\alpha}^T \mathbf{1} \quad (6.33)$$

These quadratic maximization coefficients, i.e. the information obtained for the training sample, are the constants of this problem. Obtaining the α values for a large sample size is extremely cumbersome. Fortunately, due to the KKT conditions, some of the α values become zero (for training elements which violate the conditions). By estimating these values beforehand, and only focusing on a smaller subset of α parameters, the calculation time can be decreased significantly. This technique is known as *chunking* and was proposed by Vapnik. An even more effective method, *Sequential Minimal Optimization (SMO)*, introduced by Platt [101], takes the chunking method to extreme and only considers the minimum number of parameters.

By considering the stability condition given in Eq. 6.13, optimizing a non-zero α value is not possible due to

$$\alpha_m y_m = - \sum_{i=1, i \neq m}^N \alpha_i y_i, \quad (6.34)$$

where the right-hand side of the equation is determined by the constants of the problem. How-

ever, two α values can be optimized simultaneously:

$$\alpha_m y_m + \alpha_n y_n = - \sum_{i=1, i \neq m, n}^N \alpha_i y_i = a_{m,n}. \quad (6.35)$$

This equation can be rewritten as:

$$\alpha_m = y_m (a_{m,n} - \alpha_n y_n) \quad (6.36)$$

The algorithm iterates over α values and updates $a_{m,n}$, until it reaches a predefined *accuracy threshold*.

The LIBSVM software library that is implemented in the SVM-HINT uses a modified version of the SMO algorithm [102], which puts additional second order constraints on Eq. 6.36.

As a final remark, the quadratic maximization parameters are dominated by features with high values (i.e. $|\vec{x}_i| \gg |\vec{x}_j|$). The maximization algorithms become more sensitive to variables with larger values. In order to prevent such situations, the input variables are scaled in the training sample such that:

$$\vec{x}_i = (x_{i,0}, \dots, x_{i,n}) \rightarrow \vec{x}'_i = (c_0 x_{i,0}, \dots, c_n x_{i,n}), \quad i = 1 \dots N, \quad (6.37)$$

$$c_k = 1 / (\max_{i=1 \dots N} x_{i,k} - \min_{i=1 \dots N} x_{i,k}), \quad k = 1 \dots n. \quad (6.38)$$

The obtained scale parameters c_k are propagated to the test sample.

6.2.5 Probability output

For an unclassified sample element u , the decision function given in Eq. 6.18 provides a binary output, i.e. the class membership label. However, certainty of this decision or a posterior probability P that quantifies the belief in the class label is useful and offers an easier interpretation. In Section 6.5.2 and Chapter 7, the probability cut $P > P_0$ is used to modify the signal-to-background ratio and the total number of selected events. The probability distribution is obtained by using Platt scaling [103, 104] for each element. For the training sample element $(y_i, \vec{x}_i)$, a sigmoid model can be fitted such that:

$$P(y = 1|g) = \begin{cases} \frac{\exp(-t)}{1+\exp(-t)} & : t \equiv A + Bg \geq 0 \\ \frac{1}{1+\exp(t)} & : t < 0 \end{cases}, \quad (6.39)$$

where the decision value g is given in Eq. 6.18.

In general, especially for the non-linear SVM, the result will be biased if the SVM training data itself is used for the fit. In LIBSVM, the bias is avoided by a five-fold cross-validation. The procedure of k-fold cross-validation splits the training data randomly into k equally sized subsamples. One subsample is retained for validation while the remaining k-1 subsamples are used for the

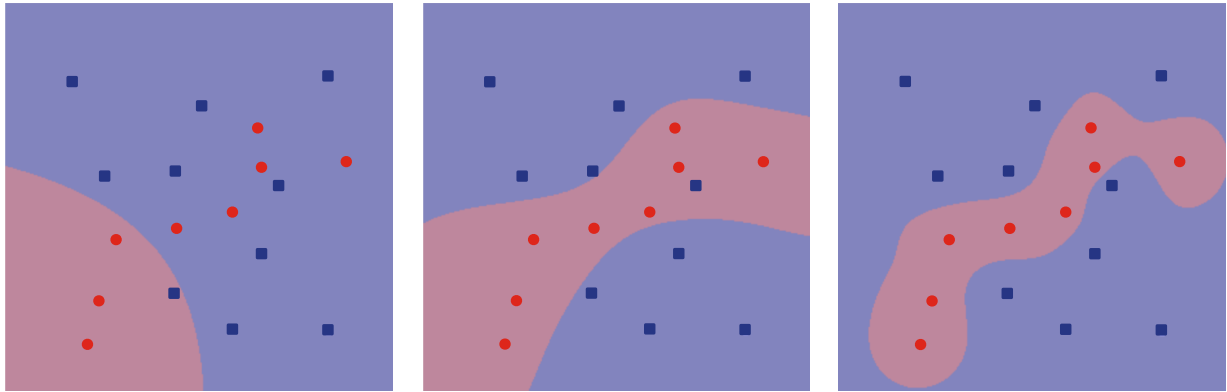


Fig. 6.4: Randomly generated samples with two distinct classes and two arbitrary features. The SVM instances are trained with $C = 5$ and γ values 1, 10, and 100 from left to right. With increasing γ value, the SVM classifier becomes more sensitive to the sub-structures in the training samples.

training procedure. The training is repeated k times with changing assignments such that each subsample is used exactly once for validation.

6.2.6 Hyper-parameter tuning

The SVM classifier with the RBF kernel has two independent hyper-parameters which are required to be tuned according to the problem of interest. The parameter C is independent of the choice of the kernel. It adjusts the penalty assigned to each misclassified element. High values of C result in a hard-margin-like SVM, which is not effective against fluctuations in the underlying distributions.

The parameter γ adjusts the width of the Gaussian smoothing. Higher values of γ correspond to a smaller gaussian width, hence the SVM becomes more sensitive to smaller sub-structures in the training sample as demonstrated in Fig. 6.4. The RBF kernel with low values of γ may result in a kernel which is too broad for the underlying distribution and in an inefficient SVM instance.

Hyper-parameters are required to be set before the classification. *Overtraining* on statistical fluctuations must be avoided while training. By using an independent test sample while optimizing the hyper-parameters, hence using a single-fold cross validation method.

Comparison between SVM hypotheses trained with a different set of hyper-parameters can be performed by assigning a performance measure to each hypothesis. There are various performance measures that can be used for this purpose as described in the following.

Machine learning performance measures

In machine learning, a classifier's performance is evaluated by how well it is able to separate the two distinct classes. In a test sample, the true class memberships of the elements are known. Therefore, the classification performance of a certain classifier can be obtained by comparing the actual memberships with the classifiers predictions. Typical ML performance measures are

the *accuracy* which gives the percentage of correctly predicted labels, or the *precision* which, in our case, is the percentage of correctly predicted signal events. Another frequently used measure is the *AUC*, the area under the receiver operator curve (ROC). The ROC curve represents the background rejection (false positive) against the signal efficiency (true positive) for the event yields after each probability cut on the decision functions. The area under ROC curve indicates the overall signal over background efficiency.

While the use of the described and other performance measures is common also in HEP [90], a different approach is followed in this thesis as described in the next paragraph.

Physics motivated performance measures

The performance of a new physics search can be characterized by its sensitivity and the potential to discover new physics. Therefore, the efficiency of a new physics search can be improved by optimizing the statistical discovery significance of the analysis. Typically, the search area is optimized for a statistically relevant signal to background ratio that allows to accept or reject a certain hypothesis. Here, we consider the case of a cut-and-count analysis for which several significance estimators are commonly used [105, 106]. Optimizing a certain, statistically motivated, figure of merit is common practice in HEP in order to select different ML algorithms or different sets of input variables. The new insight of this thesis is that such a procedure can successfully be applied in the stage of model selection, i.e. during the hyper-parameter tuning.

Asimov estimate: The exact numerical calculation of the statistical significance may become computationally costly due to the required toy Monte Carlo pseudo-experiments. A well performing estimate for the discovery significance is given in [105]. For the case of Poisson distributed background and signal events (s , b) with the background uncertainty σ_b , the approximated median discovery significance is given by

$$Z_A = \left[2 \left((s+b) \ln \left[\frac{(s+b)(b+\sigma_b^2)}{b^2 + (s+b)\sigma_b^2} \right] - \frac{b^2}{\sigma_b^2} \ln \left[1 + \frac{\sigma_b^2 s}{b(b+\sigma_b^2)} \right] \right) \right]^{1/2}. \quad (6.40)$$

Table 6.1 provides a comparison between the Asimov estimator and the calculated statistical significance quoted by various experiments. The Z_A values are calculated using the SVM-HINT software *from* class [107].

A wide collection of the significance estimators (except for the Asimov estimator) are compared in [106]. The estimations quoted in Table 6.1 indicate that the Asimov significance performs similar or better comparing to other significance estimators.

A ROOT interface for the LIBSVM library is developed together with a grid search algorithm, which is described elaborately in Section 6.4.2, to optimize the SVM hyper-parameters.

Table 6.1: Comparison of the Asimov estimator and significances quoted in different experiments. n is the event count, $\hat{\mu}_b$ is the background expectation, and the quoted Z (obtained via the log-likelihood ratios) is given with the Asimov estimator Z_A . The Asimov significance agrees with the quoted significance by 95% in the worst case.

$\hat{\mu}_b$	1.3	3.8	3.8	388.6	493434	2109732
$s = n - \hat{\mu}_b$	4.7	5.2	13.2	134	4992	9717
$f = \sigma_b / \hat{\mu}_b$	0.231	0.237	0.158	0.0207	0.00142	0.000206
Quoted Z	2.7	1.9	4.6	5.9	5.0	6.4
Z_A	2.8	2.0	4.6	5.9	5.0	6.4

6.3 Binary Decision Trees

Binary decision trees are one of the most commonly used machine learning algorithms for classification problems in experimental particle physics. Therefore, it is used to benchmark the performance of the SVM classifier and SVM-HINT. In this section, a brief discussion of the binary decision tree classifier and the classifier related subjects are presented.

A binary decision tree [108] is the simplest decision tree algorithm and is strictly bound to a structure with two branches. Each node of a binary decision tree is connected to two branches that are split with respect to a variable. The decision of the variable and its threshold value is adjusted by a performance measure. All variables are iteratively considered for each node, thus multiple cuts on the same variable is possible within the same tree. Even for relatively small tree sizes, simultaneously calculating all possible combinations of a given feature set to obtain the optimal separation is a lengthy calculation. Therefore, in practice a node based adjustment is mainly followed. For each node, the optimal feature and its cut value are obtained by calculating each possible variation in the feature set and then ranking them according to the performance measure. As a consequence, each branch leads to a differently optimized cubical separation in the multi-dimensional feature space. The Gini index of diversity is used as the performance measure in this thesis and it is defined as:

$$i(t) = \sum_{i \neq j} p(i|t)p(j|t), \quad (6.41)$$

where $p(y|t)$ is the proportion of the class y in the given node t . For binary trees, this can be rewritten as:

$$i(t) = p(y|t)(1 - p(y|t)). \quad (6.42)$$

One of the most important challenges for decision trees is the determination of the node depth (adding new nodes to the tree). The method followed in the implementation used for this thesis allows the trees to grow as much as possible by respecting a minimum node size and, after growth, pruning branches that increase the misclassification cost. There are different ways of enhancing the performance of binary decision trees. In this thesis, we focus on binary decision trees with boosting algorithm. The abbreviation *BDT* is used for boosted binary decision trees.

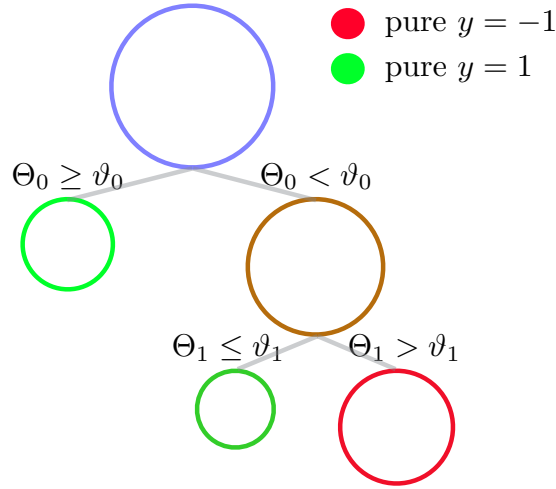


Fig. 6.5: Representation of a simple binary decision tree structure: Each node is split with respect to a feature Θ_i and a cut value θ_i determined by the performance measure.

6.3.1 Boosting

Boosting is a powerful iterative algorithm which improves the performance of the base (or weak) classifiers. After each training of the base algorithm, boosting increases the weights of the misclassified training points similar to the perceptron algorithm. After the final iteration, each classifier contributes to the final decision proportional to its accuracy. The AdaBoost algorithm developed by Freund and Schapire [109] is used for this study. For the training set defined in Eq. 6.3, the algorithm can be summarized as follows:

1. All events are initialized using the same weight $D_1(i) = 1/N$.
2. The base classifier is trained on the weighted distributions D_n .
3. The importance parameter α_n of the given base classifier is calculated by evaluating

$$\alpha_n = \frac{1}{2} \ln \left(\frac{1 - \epsilon_n}{\epsilon_n} \right), \quad (6.43)$$

where ϵ_n is the ratio of the misclassified events for the n th base classifier.

4. The event weights of the next iteration D_{n+1} is updated following

$$D_{n+1}(i) = \frac{D_n(i)e^{\alpha_n y_n}}{Z_n}, \quad (6.44)$$

where Z_n is the normalization term.

5. Finally, each weak classifier contributes to the final decision according to the classification

performance obtained in each iteration via

$$g(x) = \text{sign} \left(\sum_n \alpha_n y_n(x) \right) \quad (6.45)$$

Binary decision trees with high node multiplicity are able to provide a sophisticated and effective decision surface, but the calculation can be time consuming. The AdaBoost algorithm with simple binary decision trees generates a fuzzy decision surface where multiple different decision trees vote on a single element, thus a degree of belief in the class membership is obtained directly.

6.4 Classifier implementations

Modern HEP collider experiments pose a great challenge to physicists in terms of analysis and classification, of collected data or the simulated samples. The ROOT analysis framework is widely used in HEP to overcome some of these challenges of HEP collider experiments. In many cases, very rare processes of interests need to be separated from a large sample of background events. Therefore, physicists can benefit strongly from machine learning techniques like SVM and BDT. For this purpose we developed a new ROOT interface to LIBSVM. In the following we compare the performance and ease of use of the new LIBSVM-ROOT interface with the popular HEP multivariate analysis toolkit TMVA. The BDT implementation of the TMVA package is used as a reference of performance comparison since it is very popular among experimental HEP physicists and claimed to be one of the best out-of-the-box classifiers.

In the TMVA version used in this thesis (4.2.0), the SVM implementation includes only the RBF kernel, while the LIBSVM's implementation accompanies RBF, polynomial, sigmoidal, and linear kernels. Unlike TMVA, LIBSVM also contains different soft margin approaches. However, these differences are irrelevant to the context of the performance comparison. On the other hand, these two SVM implementations are expected to differ in performance due to different sequential minimal optimization methods followed in the minimization problem. Neither of these two packages have built in hyper-parameter search functionality, however, an external grid search script is available for the LIBSVM. TMVA offers a more complete framework, yet the trade off is being a very closed environment, which makes individual TMVA classifiers harder to use outside of the TMVA framework. Therefore, in terms of ease of development and performance, LIBSVM is used to develop the interface for the SVM.

6.4.1 Method of comparing performance of classifiers

Normally, the area under ROC curve is a common and reliable attribute to compare classifiers. However, in the classification problem discussed in the following, the aim is to obtain the maximum statistical significance of discovery of new physics. Thus, optimizing and comparing classifiers according to the maximum significance serves better the needs of experimental HEP. However, calculating the statistical significance may be computationally costly; therefore, in this pa-

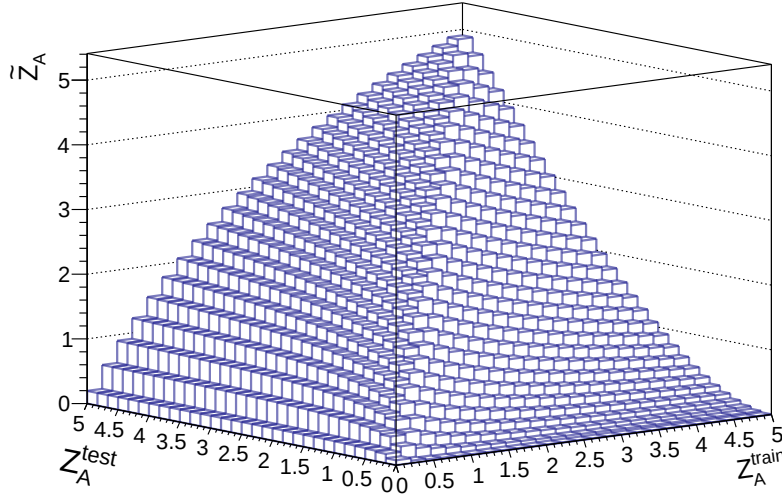


Fig. 6.6: Discrepancies between the Asimov significances obtained from training and test samples are asymmetrically penalized in the modified Asimov significance score. The score favors higher significance values on the Z_A^{test} . The highest score is obtained in the case where $\tilde{Z}_A = Z_A^{\text{train}} = Z_A^{\text{test}}$

per, the performance criterium of the classifiers is taken as the maximum value obtained from Asimov significance which is a very good approximation of the significance calculated with the likelihood methods.

6.4.2 SVM hyper-parameter optimization algorithm

SVM algorithm may have a weak classification power or may overtrain on the given training sample as observed in Figure 6.4, if the hyper-parameters are not tuned before training. While in principle a *brute force* grid search is sufficient to find the best hyper-parameters, it is computationally tiring. Therefore, an intuitive adaptive search strategy for the hyper-parameter tuning is introduced in the SVM-HINT software [107]. The SVM-HINT grid search algorithm uses a modified version of the Asimov Significance 6.40, a significance score $\tilde{Z}_A$ based on the difference between the significance value observed in the test sample and the significance value from the training sample.

$$\tilde{Z}_A = Z_A^{\text{test}} \left[1 - \frac{|Z_A^{\text{test}} - Z_A^{\text{train}}|}{Z_A^{\text{test}} + Z_A^{\text{train}}} \right]. \quad (6.46)$$

This way, the extreme significance values observed due to fluctuations or overtraining can be penalized without a high computational effort. Behavior of the modified Asimov score with respect to significance estimations obtained from test and training samples can be seen in Fig. 6.6. If the significance estimated from the test and training sample is equal to each other, the modified score is equal to this value as well. Other conditions are penalized asymmetrically while disfavoring the overtraining case i.e. $Z_A^{\text{train}} > Z_A^{\text{test}}$. The search algorithm can be formalized as follows:

1. For the given initial parameters $\gamma_{initial}$ and $C_{initial}$, the iterative grid search algorithm produces an array of logarithmically spaced γ values with a step size K_t around the mid-value $\gamma_m^{(1)} = \gamma_{initial}$ such that:

$$\gamma_k^{(l)} = K_t \cdot \gamma_{k-1}^{(l)}, \quad \text{where} \quad K_t = \frac{1}{2}(1 + \ln(t/2)), \quad (6.47)$$

$$k = 0, \dots, m, \dots, 2m = 18, \quad t = (l/4), \quad l = 1, \dots, 20$$

where l indicates the number of iterations, the variable t is a *focus parameter* that decreases the step size factor K_t every fourth iteration.

$\tilde{Z}_A$ is then evaluated for all of these C - γ -pairs.

2. For the next step C is increased to $C^{(l+1)} = 1.5 \cdot C^{(l)}$ and $\tilde{Z}_A$ is again calculated with each value in the γ array.
3. If the maximum $\tilde{Z}_A^{(l)}$ value is at least 30 % larger than the best $\tilde{Z}_A^{l-1}$ from the previous iteration the higher C parameter is accepted. The 30 % hurdle is introduced to stabilize against fluctuations.
4. After each fourth iteration, the C - γ -pair corresponding to the highest significance score is taken as the new initial γ and the algorithm returns to the first step; now with a smaller step size factor K_t such that the new γ array has a tighter stepping around the new initial γ value.
5. When the number of iterations reaches the pre-defined maximum value, the algorithm enters an optional fine-tuning step. The γ - C -pair with the maximum $\tilde{Z}_A$ in the final iteration are taken as initial values of the fine tuning step.
6. In the last step, a finer C - γ -pair grid is created by taking the optimal values as medians of the grid which has a smaller spread steps (i.e. $C^{(l+1)} = 1.25 \cdot C^{(l)}$). The C - γ -pair with the highest $\tilde{Z}_A$ score are chosen as the optimal hyperparameters of the given problem.

The procedure assumes that a sufficiently small $C_{initial}$ had been chosen. In case that the found best C value is identical with the $C_{initial}$ the algorithm is restarted with a smaller value of $C_{initial}$. Since models with smaller C value forms a *softer* margin, they are less susceptible to overtraining. Thus, choosing a sufficiently small C as the initial value ensures the tuning algorithm to start with a more general solution.

6.4.3 Parallelization

Training and testing multiple hyper-parameter settings can be computationally costly. The LIB-SVM library provides a built-in multi-threaded parallelization in the training step of the SVM classifier. Moreover, the models trained with different gamma values can be tested in parallel in the SVM-HINT software. Therefore, even without a large-scale computing facility, the SVM-HINT framework can be used efficiently with a multicore processor. The effectiveness of the parallelization is discussed in the next section.

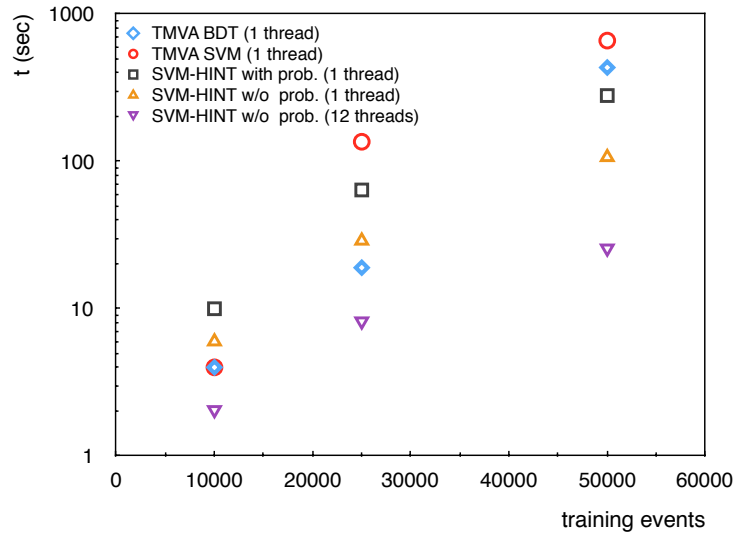


Fig. 6.7: The timing performance of the classifiers are compared on a computer with two Intel® Xeon® E5-2440 CPUs and 12 physical threads running at 2.40 GHz clock speed. Number of threads used by each classifier implementations are stated within parentheses in the legend.

6.5 Case studies

6.5.1 Performance comparison on a toy model

Comparing the performance of the two different classifier implementations in terms of speed and classification performance is not straightforward. The comparison results heavily depend on the initial classification problem. Thus, we compare the TMVA's BDT and the SVM-HINT in terms of timing performance starting from the same initial conditions and problem which lead to a similar statistical significance. Samples with different number of events are generated to investigate the effect of event multiplicity on timing performance. The events are weighted as common for MC samples of real HEP experiments with average weights for signal around 0.5 and background around 0.01. The following distribution functions are deployed as features in the generated samples:

$$\begin{aligned}
 V_1 &= \sin(x_1); & x_1 &\sim g(x_1|a, b) \\
 V_2 &= x_2; & x_2 &\sim \exp(-x_2/c) \\
 V_3 &= x_3; & x_3 &\sim g(x_3|d, e) \\
 V_4 &= \sqrt{x_4}; & x_4 &\sim \exp(-x_4/f)
 \end{aligned} \tag{6.48}$$

where $g(x, \sigma) = \exp\left(-\frac{x^2}{2\sigma^2}\right)$ and a, b, c, d, e, f are constants that have different values for the signal and background samples.

Unlike SVM-HINT, TMVA does not provide a hyper-parameter optimization method. In order to have similar results with the SVM-HINT, TMVA's BDT classifier is manually optimized. In the TMVA's SVM implementation, the hyper-parameters that are obtained from the SVM-HINT

grid-search algorithm are used. Overall, the multi-threaded performance of the SVM-HINT with twelve threads has the best timing performance of all classifier implementations. At low number of training events, the BDT performs better than other classifiers. For higher number of training events, the number of trees is increased correspondingly resulting in a slower training for the BDT. TMVA's SVM implementation does not scale well in terms of timing performance with increasing number of training events, and it performs poorly in bigger samples. Overall the SVM-HINT performs similar or better than TMVA's BDT and SVM implementations.

6.5.2 Third generation supersymmetric partner search

A more realistic comparison study is accomplished by considering a future new physics search. By the end of the LHC run in 2023, 300 fb^{-1} 14 TeV pp -collision data is expected to be collected. Search for direct top-squark pair production in the single-lepton final state is performed to explore the reach of the SVM's performance and investigate effects of variable choice. As discussed in Chapter 5, the dominant irreducible background is the top-antitop production and for simplicity only this background is considered. LSP mass of 100 GeV and top-squark mass of 900 GeV are considered as the signal sample in this case study.

Analysis strategy

The baseline selection given in Table 6.2 are applied on the background and signal samples in order to reduce the time required for training and optimization of the classifiers. To investigate the effect of variable selection, four different sets of variables, each containing different variable numbers with different complexity, are used. These variables are discussed in detail in Chapter 5 and the sets are given in Table 6.3.

Table 6.2: List of baseline cuts used for 14 TeV analysis.

$ \eta_{l,\text{jet}} $	$<$	2.4
$p_{T,l}$	$>$	30 GeV
$p_{T,\text{jet}}$	$>$	40 GeV
$p_{T,\text{jet1}}$	$>$	80 GeV
$p_{T,\text{jet2}}$	$>$	60 GeV
$\cancel{E}_T$	$>$	200 GeV
H_T	$>$	300 GeV
n_{jet}	$>$	3
$n_{b\text{jet}}$	$>$	0

After the baseline selection, the event yields of background process is several orders of magnitude higher than the signal process as expected. A subset of low- and high-level variables (labeled as Set-4) are presented in Fig. 6.8.

Table 6.3: Summary of all low-level and high-level variables used in the analysis. Set-1 includes all variables. Set-2 and set-3 consist of low- and high-level variables, respectively. Set-4 is a smaller subset of high- and low-level variables.

	Variable	Set 1	Set 2	Set 3	Set 4
low-level	$p_{T,l}$	•	•		
	η_l	•	•		
	$p_{T,jet(1,2,3,4)}$	•	•		
	$\eta_{jet(1,2,3,4)}$	•	•		
	$p_{T,b\,jet1}$	•	•		
	$\eta_{b\,jet1}$	•	•		
	n_{jet}	•	•		
	$n_{b\,jet}$	•	•		
	$\cancel{E}_T$	•	•		•
	H_T	•	•		•
high-level	m_T	•		•	•
	m_{T2}^W	•		•	•
	$\Delta\phi(W, l)$	•		•	
	$m(l, b)$	•		•	
	Centrality	•		•	
	Y	•		•	
	H_T -ratio	•		•	
	$\Delta r_{\min}(l, b)$	•		•	
	$\Delta\phi_{\min}(j_{1,2}, \cancel{E}_T)$	•		•	

The samples are separated into three independent subsamples: training, test and evaluation samples. Each classification method is optimized over training and test samples. The optimal hyper-parameters and optimal discriminator output cuts obtained from this step are used to reach final results from the evaluation sample for SVM-HINT. The TMVA's BDT (AdaBoost) is trained and tested with 8 different settings for each variable set to obtain optimal parameters (i.e. the number of trees, minimum node size, cut values etc). The best performing configuration set and the optimal discriminator output cut are directly used for evaluation.

Results

In the case study, in addition to providing the comparison of performance of the classifiers in a real-world like analysis scenario, we investigated different feature selections. Although a direct comparison of the classifiers are not possible due to lack of parameter optimization functionality in the TMVA package, we managed to manually optimize it (with 32 different configurations)

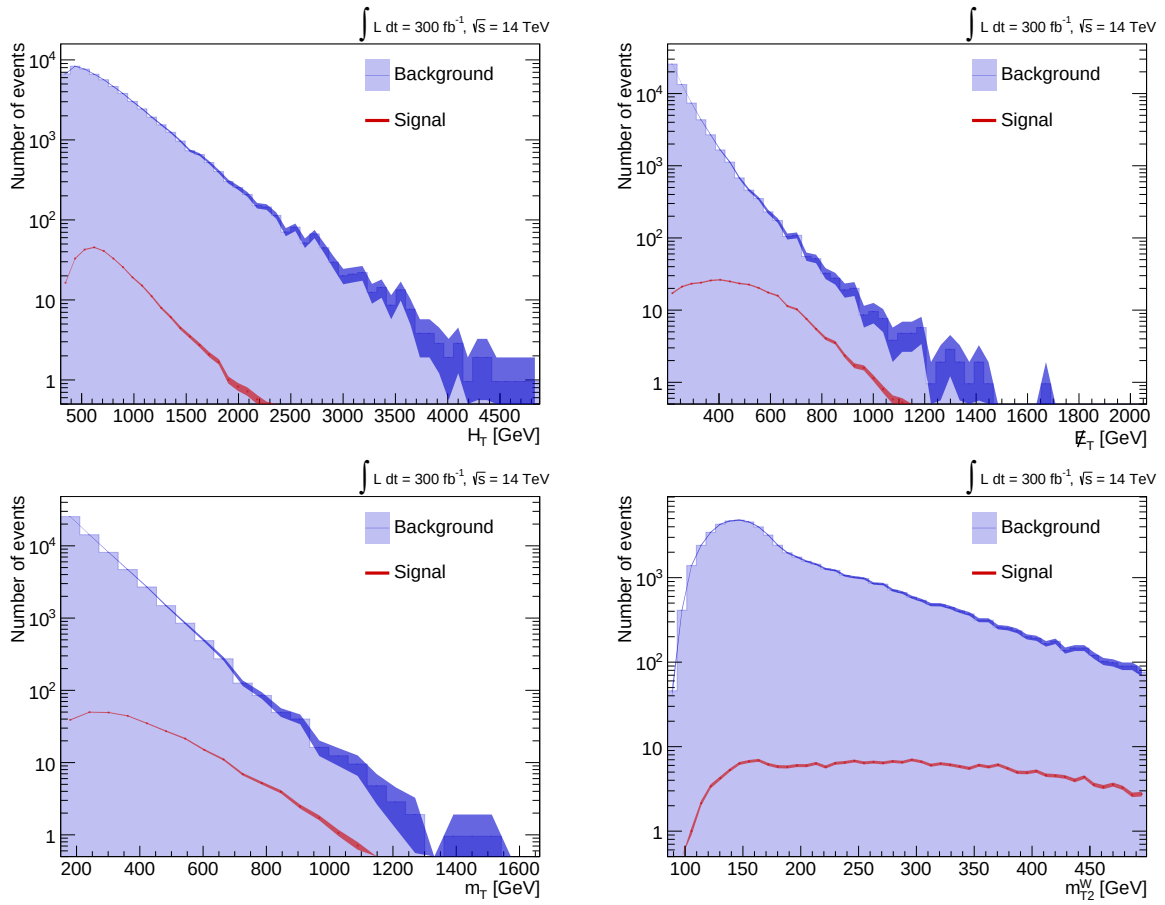


Fig. 6.8: The relative distribution of the sub-set of variables that are used in the analysis. The y-axis shows the number of events (normalized to the integrated luminosity), and the x-axis shows the variable value for given bin. Statistical errors are represented by transparent bands.

in order to obtain a performance baseline. One should also note that further improvements are possible in both classifier implementations.

Without modifying the default SVM-HINT settings, the two step grid search hyper-parameter optimization function provided the optimal parameters using test and training samples and evaluate an independent sample to give the final significance. One of the most important results is the fact that SVM-HINT performs outstandingly in a real-world like analysis scenario and can make use of a high number of features (high number of features in HEP is regarded as low number of features in the field of machine learning) simultaneously with an increasing classification power. Another important conclusion is that the SVM-HINT scales efficiently with the increasing number of features and the negative impact of the variables with small or no discrimination power does not have a drastic impact on accuracy. This can be seen in the contribution of the feature set-2 in the SVM-HINT trained with the feature set-1, since $\text{set-1} = \text{set-2} \cup \text{set-3}$. The SVM-HINT trained with set-2 does not perform outstandingly comparing to other configurations, regardless, contribution of the features set-2 is visible in the SVM-HINT trained with set-1. This implies that even

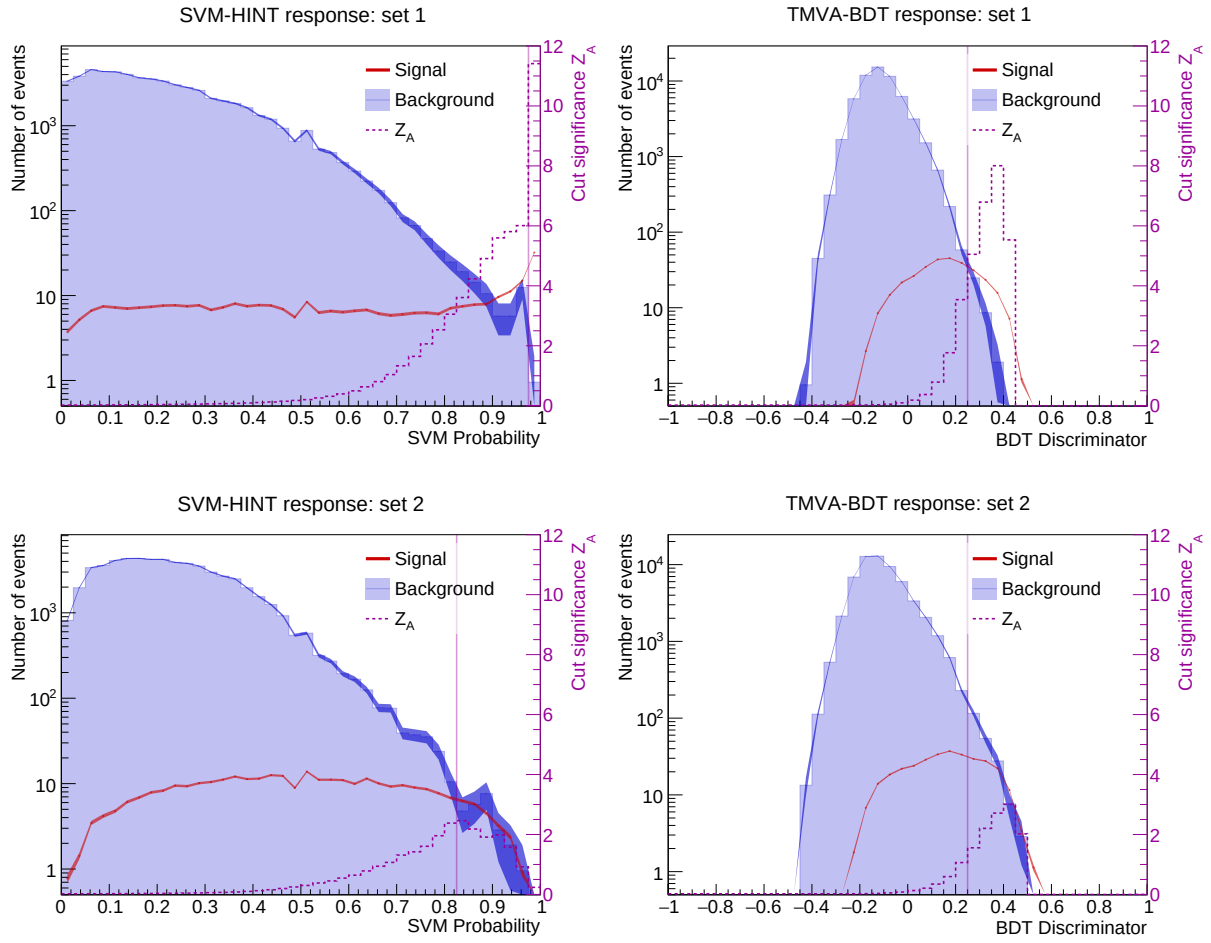


Fig. 6.9: The SVM-HINT and TMVA BDT responses trained with feature set-1 (top) and set-2 (bottom) defined in 6.3. Even though the optimal Z_A efficiency information is not available in collected data, it is included to observe the reliability of the optimal discriminator cut output from the classifier implementations. The y-axis on left shows the number of events normalized to aimed integrated luminosity, whereas the y-axis on right shows the Asimov significance for the discriminator cut for given bin. The SVM-HINT obtains the highest accuracy with the largest number of features (set-1). The SVM-HINT classifier that trained with set-2 does not perform as well as other configurations, on the other hand, set-2 contributes the overall accuracy on the inclusive set-1.

with a feature multiplicity of 25, SVM-HINT does not require a preselection of features.

As a final remark, the importance of the performance measures is visible on the discriminator cut decisions by the classifier implementations. The TMVA-BDT uses $\frac{S}{\sqrt{S+B}}$ as the performance measure, unfortunately, this estimator is unable to produce results similar to log likelihood significance calculation. Therefore, the optimal cut provided by TMVA reduces the significance obtained from the classifier implementation.

SVM-HINT uses the Asimov significance which gives very similar results to the log-likelihood calculation, therefore, the results obtained from SVM-HINT not only provides out of the box good

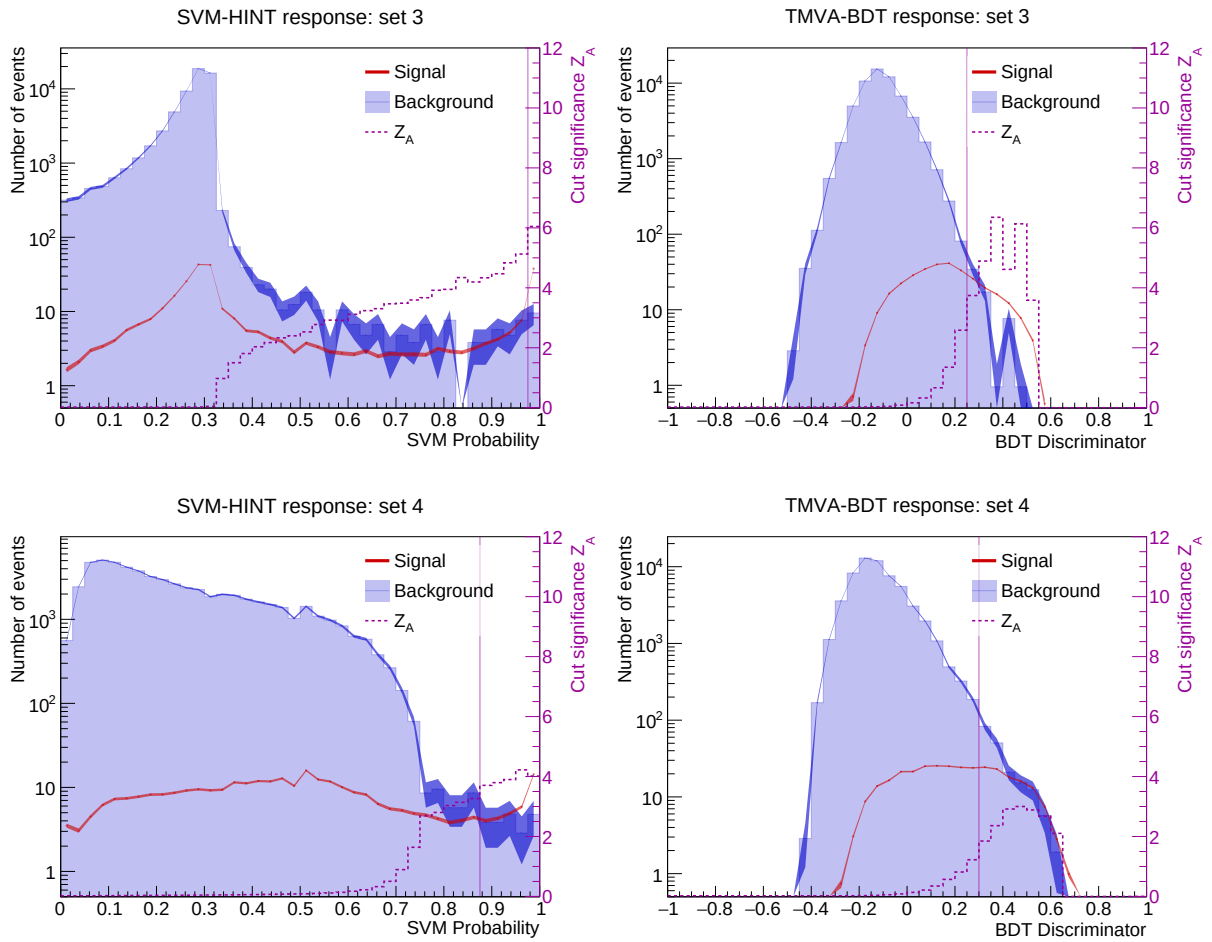


Fig. 6.10: The SVM-HINT and TMVA BDT responses trained with feature set-3 (top) and set-4 (bottom) defined in 6.3. Even though the optimal Z_A efficiency information is not available in collected data, it is included to observe the reliability of the optimal discriminator cut output from the classifier implementations. The y-axis on left shows the number of events normalized to the integrated luminosity, whereas the y-axis on right shows the Asimov significance for the discriminator cut for given bin. The statistical errors are presented with the darker bands.

estimation of the actual significance, but the discriminator cut given by SVM-HINT maximizes the significance between background and signal.

Table 6.4: Summary of the number of signal (N_s) and background (N_b) events expected for the different variable sets and the two multivariate methods under study. In addition, we quote the achievable discovery significance ($Z_A^{\max}$). For comparison, we also quote the significance for the cut value suggested by SVM-HINT and TMVA-BDT, respectively (Z_A^{prov}).

Set	SVM-HINT				TMVA-BDT			
	s	b	$Z_A^{\max}$	Z_A^{prov}	s	b	$Z_A^{\max}$	Z_A^{prov}
1	32.1 ± 0.6	1.0 ± 1.0	11.4	11.4	24.1 ± 0.5	1.9 ± 1.4	8.0	5.1
2	23.2 ± 0.5	23.9 ± 4.8	2.5	2.5	16.2 ± 0.4	10.5 ± 3.2	3.0	1.5
3	37.8 ± 0.6	9.6 ± 3.0	6.1	6.1	40.5 ± 0.7	9.6 ± 3.0	6.4	3.7
4	33.4 ± 0.6	20.1 ± 4.4	4.2	3.7	40.5 ± 0.7	35.4 ± 5.8	3.0	1.8

6.6 Summary

SVM is an intuitively understandable, but rigorously constructed machine learning algorithm which has not been widely used in the field of HEP. The SVM-HINT software with the statistical-significance based hyper-parameter tuning provides a state of the art performance. It autonomously adjusts itself to benefit from large number of variables available to describe the underlying physics processes. With the insight presented in the case studies, the SVM-HINT is employed to search for SUSY events in the following sections.

In the next section, for 8 TeV pp -collisions, the SVM-HINT is used to separate signal from background in the single lepton final state. After optimizing the hyper-parameters for each LSP mass and top-squark partner's mass combination, the SVM discriminator cuts are applied on both data and MC samples. By taking the MC samples as a reference, the background yields are estimated using a data driven method described in [79, 110].

Analysis strategy

Contents

7.1	SVM hyperparameter tuning	110
7.2	Simulation validation	115
7.2.1	W + jets and 1l top validation in CR1: 0b jet final state	116
7.2.2	$t\bar{t} \rightarrow l\bar{l}$ validation in CR2: 2l final states and CR3: $l\tau$ final states	117
7.3	Background Estimation	119

The search for extremely rare top-quark SUSY partners requires a good description of the SM background processes and an effective classification methodology. The powerful machine learning tool, SVM-HINT, is created with this purpose, and hence it is employed in the present analysis to find the *slightly different and rare hays in the haystack*.

The methodology for analyzing 19.5 fb^{-1} pp collision data collected in the CMS detector at 8 TeV to search for top-quark SUSY partners is discussed in this chapter. For each *signal* sample corresponding to different $m_{\tilde{t}}$ and $m_{\tilde{\chi}^0}$ configurations, SVM-HINT tunes the hyperparameters on independent training and test subsamples. An optimal threshold maximizing the discovery significance for each signal sample is determined automatically.

SVM-HINT relies on the labeled MC samples in order to accurately discriminate signal processes from the SM background. The SVM-discriminator thresholds and the RBF-SVM models are selected to maximize the discovery significance, resulting in a low number of events mostly occupying the tail of the variable distributions. Therefore, the MC subsamples passing the SVM discriminator threshold are required to be validated.

In Section 7.1 the optimization methodology is described. The validation and correction of the background samples are discussed in Section 7.2 and 7.3.

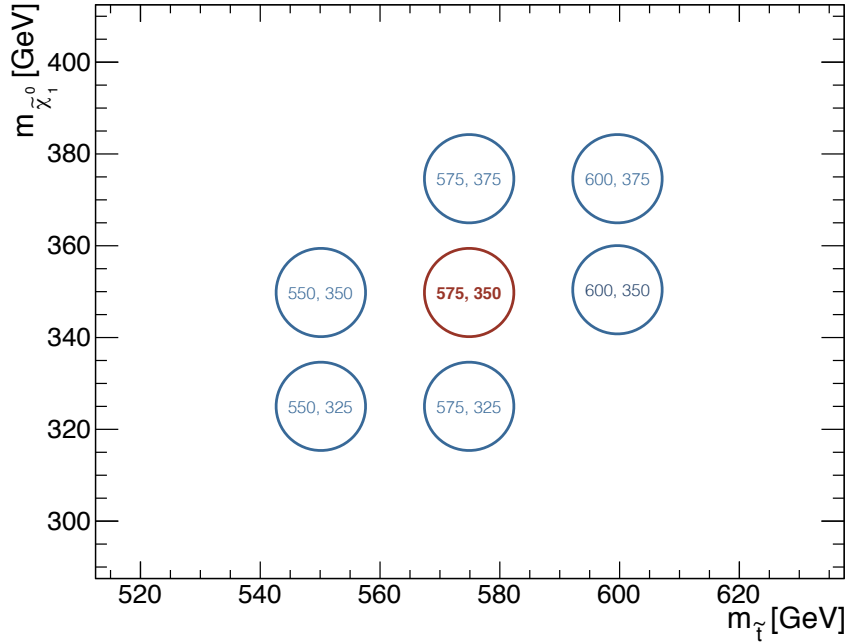


Fig. 7.1: For an example signal sample $m_{\tilde{t}} = 575$ GeV, $m_{\tilde{\chi}^0} = 350$ GeV, the training and test subsamples enriched by the neighboring points are shown. The red circle represents the signal point of interest and the blue circles represent the neighboring points that are used for enriching the training and test samples for the point of interest.

7.1 SVM hyperparameter tuning

The underlying physics of the T2tt signal model may vary significantly with respect to the difference between the top squark and the neutralino (ΔM). The samples where $\Delta M \simeq m_t$ or $\Delta M < m_t$ lead to smaller phase spaces for the mediator particles or may require virtual mediators. Moreover, even the signal samples with the same ΔM , but different sparticle masses, may present themselves with different event topologies. Therefore, an effective classification of the signal processes can be achieved by optimizing the algorithm, hence the SVM hyperparameters, for each mass configuration of the signal model.

As discussed in Chapter 6, the SVM hyperparameter tuning is performed using two independent subsamples. However, dividing the signal MC sample into three subsamples (training, test and evaluation) results in low sample size for training, test and evaluation of the signal samples. To avoid the impact of low sample size the following methodology is employed: The samples corresponding to each mass point is divided into two equal subsamples for hyperparameter tuning and evaluation of the models, and the number of events in the sample reserved for the tuning procedure is increased by merging the samples of the neighboring mass configurations which differ from either $m_{\tilde{t}}$ or $m_{\tilde{\chi}^0}$ or ΔM by 25 GeV as depicted in Fig. 7.1. Even though the kinematic distributions may not be exactly the same as the point of interest, they are nevertheless similar enough to justify this strategy, as shown in Fig. 7.2 for two different signal samples with different

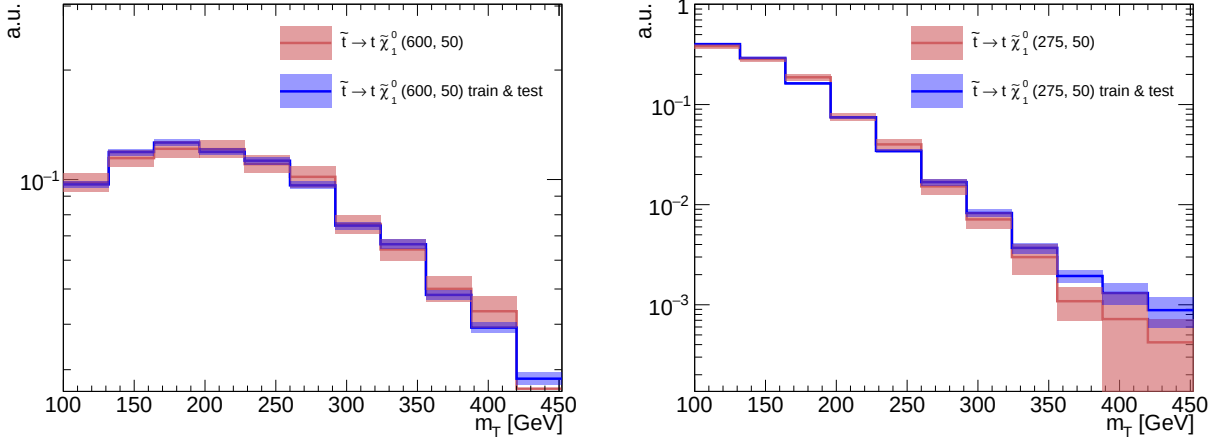


Fig. 7.2: m_T distributions of two example signal samples (blue line) and the enriched samples with the neighboring $(m_{\tilde{t}}, m_{\tilde{\chi}^0})$ mass configurations (red line). The resulting distributions are within the statistical uncertainties indicated by the transparent bands.

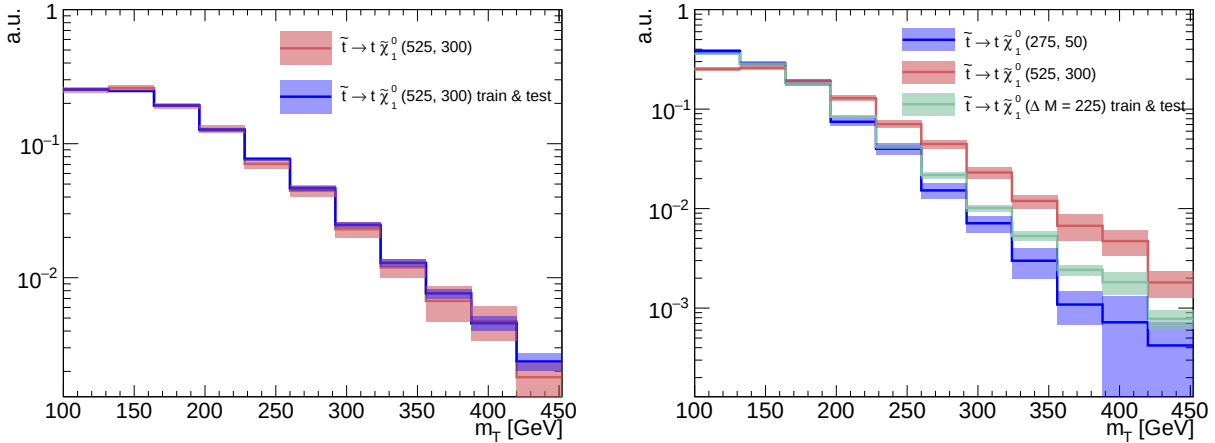


Fig. 7.3: m_T distributions comparison of two different signal sample enrichment approaches. The method presented on left combines the $(m_{\tilde{t}}, m_{\tilde{\chi}^0})$ neighboring samples, whereas the other method provides a common training and test sample for a given ΔM configuration. Even though the latter method provides higher sample size, the underlying physics is not described as good as the method that is used in the present analysis.

mass configurations.

Another approach to enhance the size of the training and test samples that can be employed is to combine signal samples with the same ΔM values [79]. The resulting distributions, however, may differ significantly from the ones of the point of interest, as shown in Fig. 7.3.

In the present analysis a large set of high- and low-level variables are considered as listed in Table 7.1. The variable m_T is excluded, since it is used for the background evaluation. A minimum of $m_T > 100$ GeV is required for the search regions. After enriching the signal samples, following a similar procedure as described in Chapter 6, the SVM hyper-parameters are optimized based

Table 7.1: Summary of all low-level and high-level variables used in the analysis. The m_T variable is excluded, since it is used for the background validation.

low-level variables	high-level variables
$p_{T,l}$	Y
η_l	m_{T2}^W
$p_{T,jet(1,2,3)}$	$\Delta\phi_{\min}(j_{1,2}, \cancel{E}_T)$
H_T	$m(l, b)$
$p_{T,b\,jet1}$	Centrality
n_{jet}	H_T -ratio
$n_{b\,jet}$	$\Delta r_{\min}(l, b)$
$\cancel{E}_T$	Hadronic top χ^2

on the highest Asimov significance score obtained using SVM-HINT. A systematic uncertainty of 25% is assumed for the significance estimation (cf. Eq. 6.40). Each signal sample for the specific mass configuration is individually considered, resulting in different topologies, hence decision surfaces. A minimum signal presence of 0.1% acceptance times efficiency or 2 events is imposed.

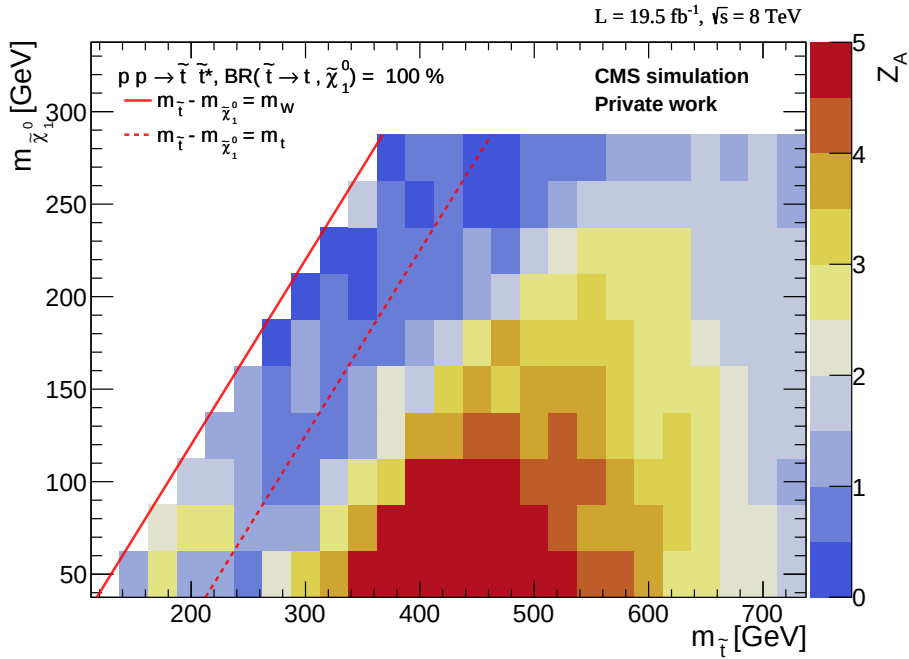


Fig. 7.4: Resulting Z_A for the signal and background samples classified by SVM-HINT optimized for each mass configuration $(m_{\tilde{t}}, m_{\tilde{\chi}^0})$.

Each of these topologies correspond to a phase space or, in other words, a search region. In the range of $100 \leq \Delta M \leq 750$ and $m_{\tilde{\chi}^0} \leq 325$ GeV, more than 250 signal samples are individually considered. SVM-HINT was unable obtain optimal hyperplane within the imposed efficiency con-

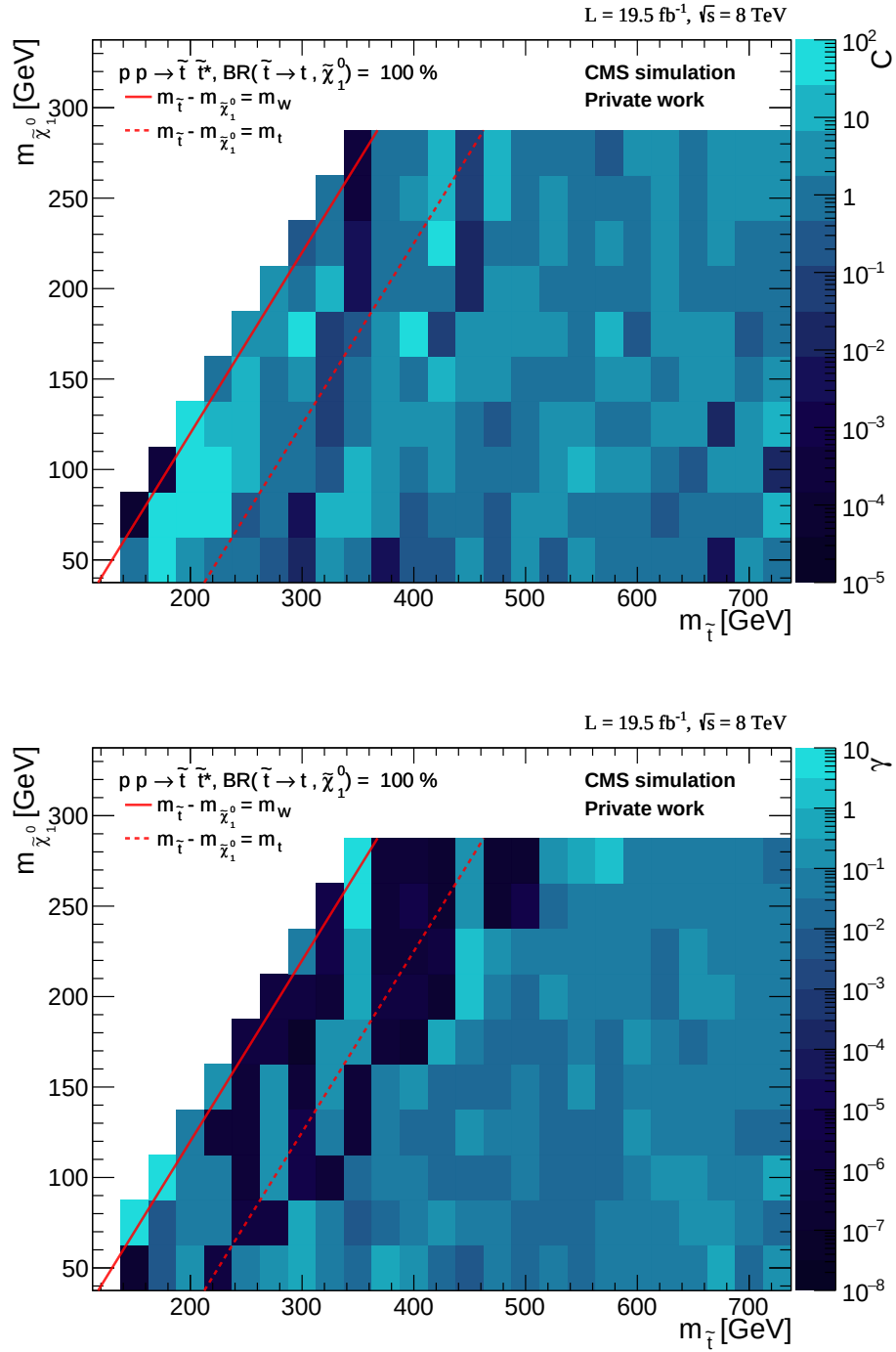


Fig. 7.5: The optimized C (top) and γ (bottom) hyperparameters varies for the different signal samples ($m_{\tilde{t}}, m_{\tilde{\chi}_1^0}$). An effective use of the SVM algorithm requires a hyperparameter tuning algorithm. The tuned hyperparameters for each signal sample results with the Z_A shown in Fig.7.4

straints on the minimum number of signal events. Therefore, some SVM models are incapable of discriminating the signal from the background. These models are especially localized around the signal samples with low ΔM and high $m_{\tilde{\chi}^0}$, and high $m_{\tilde{t}}$ as shown in Fig. 7.4. The SVM-HINT training on the remaining signal samples is converged with good estimated statistical significances, thus providing 212 search regions. The validation and estimation methodology described in the following is repeated for each of these search regions.

The corresponding multidimensional phase spaces may coincide. Regardless, a generalization of the hyperparameters is not possible, as can be clearly observed in Fig. 7.5. Each C (top) and γ (bottom) pair shown in Fig. 7.5 corresponds to different decision hyperplanes, thereby also differing event selection. To achieve the highest discovery significance, the SVM-HINT eliminates the majority of the background events, thus only encapsulating small twenty-dimensional phase spaces where the signal purity is the highest. Even though the MC generators described in Chapter 3 give a good description of the background processes, validation, and if necessary, correction of these samples in the small phase spaces is crucial for the interpretation of the results.

7.2 Simulation validation

The classification ability of the SVM relies on the labeled simulated samples, i.e. the samples where the class membership (signal or background) is known. The SVM-HINT aggressively constrains the sample space to maximize the discovery significance of the signal model. However, the MC generators and the detector modeling software may become insufficient to flawlessly explain the intricate structure of the underlying physics in these small samples. In addition, the description of the investigated phase space may not be perfectly described by simulation. Therefore, validation of the background simulations is crucial for the accuracy of the physics results.

The background simulation can be validated in phase spaces where the signal presence is not expected, yet the underlying background topologies are similar to the search region. In the present analysis, these regions are referred to as *control regions (CRs)*, and they are selected to be statistically orthogonal to the search region.

This orthogonality can be achieved by introducing different preselection criteria varying the b -jet and lepton multiplicity. These regions are chosen such that each background contribution is enriched, and, therefore, the major contribution is due to the background model under investigation.

As discussed in Chapter 5, there are four main types of background samples that need to be considered: $t\bar{t} \rightarrow l\bar{l}$, $1l$ top, $W + \text{jets}$, and rare backgrounds. The m_T variable described in Chapter 5 is used to validate the background simulation samples. Similar to the signal processes, most of the background processes include a W decay chain, thus m_T provides an excellent description of the underlying physics. The simulated samples are validated after normalizing the m_T distributions on the peak region ($50 \text{ GeV} < m_T < 80 \text{ GeV}$) that is dominated by the SM processes due to W mass peak. Weighting the background samples in this SM dominated region to data eliminates simulation related uncertainties, yet introduces systematic uncertainties related to possibly low sample size.

The $W + \text{jets}$ process is investigated in a control region where the t quark related background events are suppressed. This is achieved by vetoing all events containing b -tagged jets (since b quarks are the main decay products of t quarks), and keeping the same selection criteria with the search regions (CR1). The $t\bar{t} \rightarrow l\bar{l}$ background is investigated by requiring exactly two leptons in the final state (CR2), and hence suppressing the $W + \text{jets}$ contribution. An additional control region for the $t\bar{t} \rightarrow l\bar{l}$ background is obtained by reverting the isolated-track and τ -lepton rejection filters in the search region (CR3). The control regions used in the present analysis are summarized in Table 7.2.

It is not possible, however, to obtain such a control region for the $1l$ top background process. Therefore, CR1 is also used for the validation of the $1l$ top background sample. The contributions from the rare processes are mostly very small or even negligible, therefore, a validation with data for each individual rare process is not possible, hence the predictions for the rare processes are taken from simulation with a conservative uncertainty of 50% on the cross-sections.

Table 7.2: Overview of the control regions, and background samples validated in each region. In CR3, the isolated track and the τ lepton is counted as the second lepton.

$n_{b \text{ jet}}$	1 μ or 1 e	2 leptons
0	CR1 - W + jets and $1l$ top	
≥ 1	Search region	CR2 and CR3 - $t\bar{t} \rightarrow l\bar{l}$

7.2.1 W + jets and $1l$ top validation in CR1: 0b jet final state

CR1 is used to validate the W + jets and $1l$ top MC samples shown in Fig. 7.6 (left). While all other MC samples are normalized to the theoretical cross-sections, only the W + jets sample is reweighted to match the background samples to the data in the m_T -peak region. After normalizing the peak region, the tail of the m_T distributions is compared to data in order to validate the W + jets simulations. The MC simulation underestimates the m_T tail as shown in (middle) Fig. 7.6. Therefore, the ratio between the tail region ($m_T > 100$ GeV) and the peak region is used to quantify this discrepancy. The ratio (r_{W+jets}^{MC}) obtained from the MC sample is compared to the one obtained for the data (r_{W+jets}^{data}). The incompatibility between these two values is quantified with a scaling factor $SF \cdot r_{W+jets}$. As mentioned earlier, the SVM-HINT aggressively eliminates the background events resulting in very low event yields in the high $m_{\tilde{t}}$ region.

Thus, considering each mass point individually causes the scale factors to be dominated by statistical fluctuations. As discussed in Chapter 2, with the increasing top squark masses, the cross-section of the T2tt models steeply falls. Therefore, in order to account for steeply falling cross-sections and keeping similar event topology, the signal regions are separated into two groups with respect to the ΔM (below and above $\Delta M = 400$ GeV), and a linear χ^2 fit is performed to estimate the scale factors. Muons and electrons are considered individually for two extreme cases, i.e. either by taking the W + jets sample as the single source of discrepancy, or by assuming that all samples contribute to the discrepancy. An average of these two extreme cases is taken. A conservative uncertainty of 20% is assigned to the obtained fit values to account for possible correlation effects. The resulting scale factors ($SF \cdot r_{W+jets}$) are presented in Table 7.3 and reweighted simulation compared to data is shown in Fig. 7.6.

Table 7.3: $SF \cdot r_{W+jets}$ indicating the discrepancy between the data and MC samples' tail-to-peak ratio. ΔM represents the difference between the mass of top squark and neutralino. The search regions are separated into two groups with respect to ΔM (below and above $\Delta M = 400$ GeV). A linear χ^2 fit is performed to obtain the $SF \cdot r_{W+jets}$ within each group. The procedure is applied to single electron and single muon samples separately.

lepton	$\Delta M < 400$ GeV	$\Delta M > 400$ GeV
Electron	1.25 ± 0.25	1.50 ± 0.30
Muon	1.24 ± 0.25	1.24 ± 0.25

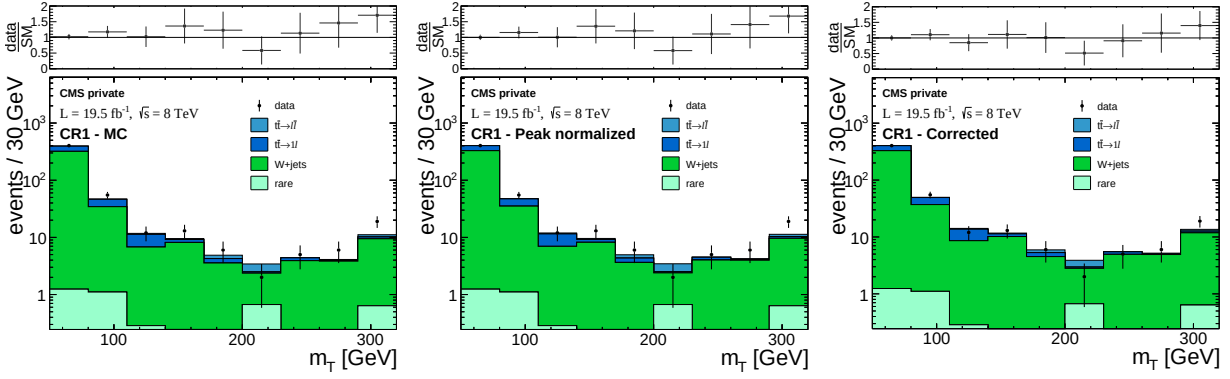


Fig. 7.6: Comparison of the data and the SM m_T distributions for the benchmark signal sample ($m_{\tilde{t}} = 275$ GeV, $m_{\tilde{\chi}^0} = 50$ GeV) in CR1. The discrepancy between the data and MC samples is visible in the uncorrected plot on the left. The figure in the middle shows the distributions normalized in the m_T -peak region. The dominant process, $W + \text{jets}$, and the second most dominant process, $l\bar{l} \text{ top}$, are corrected by scaling the samples to the ratio of tail-to-peak yields ($SF \cdot r_{W+\text{jets}}$ and $SF \cdot r_{l\bar{l} \text{ top}}$), shown in the right figure. The last bin includes overflow events. The simulated distributions from different background processes are stacked, while the data is individually presented.

A specific control region for the $l\bar{l} \text{ top}$ process is not possible to identify. However, by considering mis-reconstructed or out of acceptance b -jets, the CR1 region can be used to validate the $l\bar{l} \text{ top}$ background simulations as well. After correcting the $W + \text{jets}$ backgrounds with the scale factors given in Table 7.3, the $l\bar{l} \text{ top}$ background is validated following a similar procedure. The discrepancy is also visible in the tails of the $l\bar{l} \text{ top}$ m_T distributions, and is quantified by the scale of the m_T tail-to-peak ratios of the data and background simulation samples. In order to be robust against statistical fluctuations, a common fit is applied following the same procedure with $SF \cdot r_{W+\text{jets}}$. The resulting scale factors are quoted in Table 7.4.

Table 7.4: $SF \cdot r_{l\bar{l} \text{ top}}$ for the single electron and single muon final states. The discrepancy between the data and the $l\bar{l} \text{ top}$ MC sample in CR1 is corrected by scaling the tail-to-peak ratios of the MC-samples' m_T distribution with the $SF \cdot r_{l\bar{l} \text{ top}}$ quoted below.

Lepton	$\Delta M < 400$ GeV	$\Delta M > 400$ GeV
Electron	1.12 ± 0.23	0.94 ± 0.19
Muon	1.15 ± 0.23	1.67 ± 0.33

7.2.2 $t\bar{t} \rightarrow l\bar{l}$ validation in CR2: $2l$ final states and CR3: $l\tau$ final states

The $t\bar{t} \rightarrow l\bar{l}$ is the major background process of the present analysis. $t\bar{t} \rightarrow l\bar{l}$ events with a second electron and muon that does not pass the selection criteria described in Chapter 5, or is not identified as a lepton, have similar kinematic distributions as the signal processes due to less

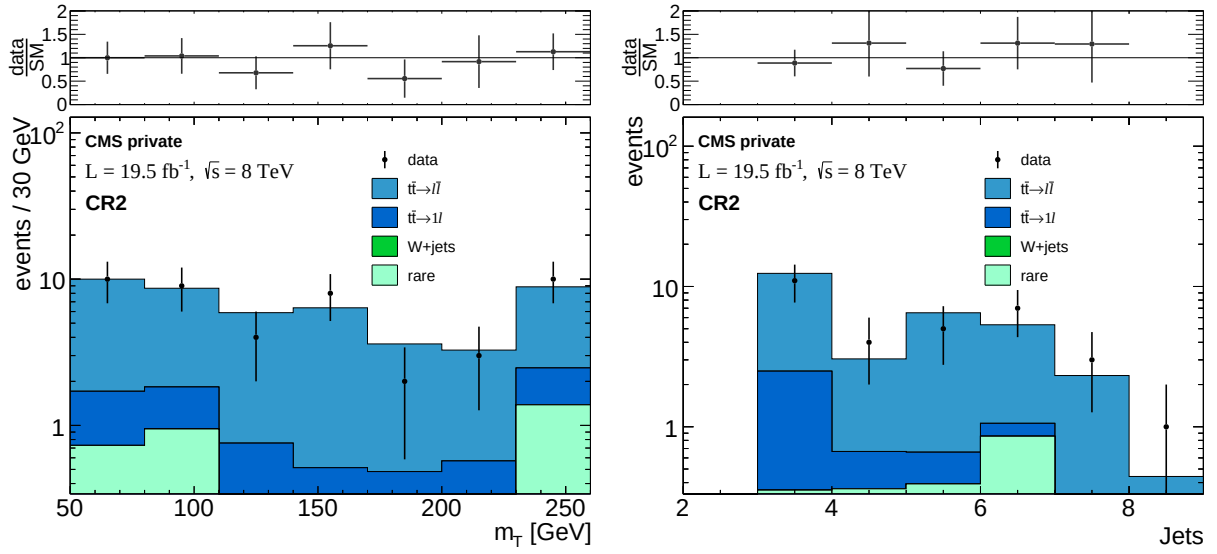


Fig. 7.7: The m_T (left) and number of jets (right) are presented for the benchmark signal sample ($m_{\tilde{t}} = 275$ GeV, $m_{\tilde{\chi}^0} = 50$ GeV) in CR2. Overall, a good agreement between the data and the SM prediction is observed even with low sample size. The last bin includes overflow events. The simulated distributions from different background processes are stacked, while the data is individually presented.

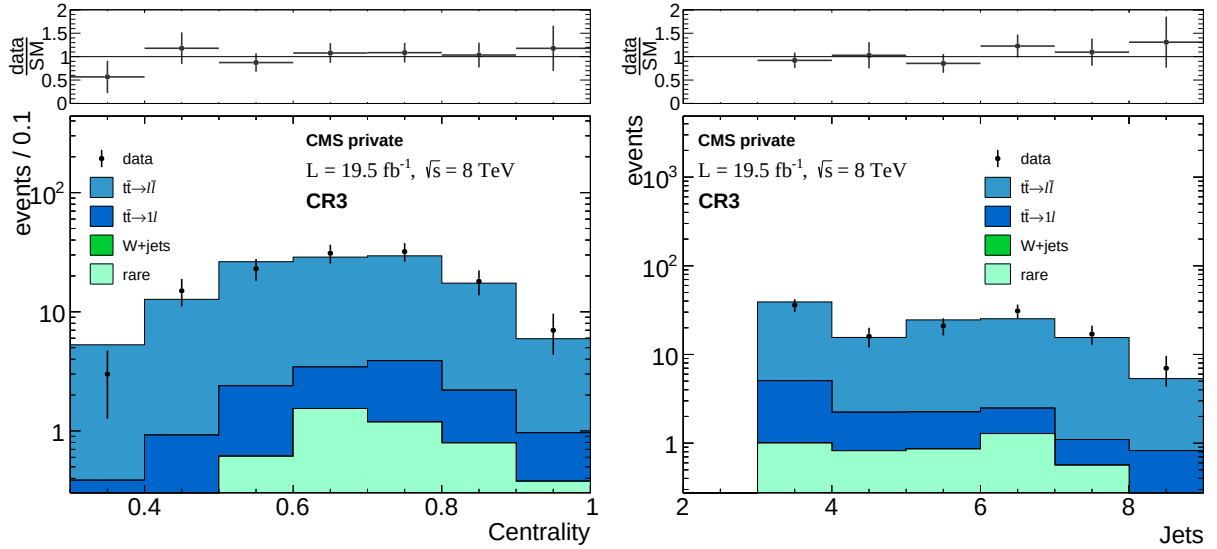


Fig. 7.8: The centrality (left) and number of jets (right) distributions are presented for the benchmark signal sample ($m_{\tilde{t}} = 275$ GeV, $m_{\tilde{\chi}^0} = 50$ GeV) in CR3. Overall, a good agreement between the data and the SM prediction is observed. The last bin includes overflow events. The simulated distributions from different background processes are stacked, while the data is individually presented.

correlated $\cancel{E}_T$ and the single lepton topology. An additional similar contribution is caused by the second lepton is a hadronically decaying τ -lepton that escapes from the τ -lepton rejection.

A two-fold approach is required to validate the $t\bar{t} \rightarrow l\bar{l}$ sample in order to address these different contributions. The single-lepton requirement is relaxed to two leptons forming CR2 which is the nominal number of final-state leptons for the $t\bar{t} \rightarrow l\bar{l}$ processes. However, this selection causes an ambiguity in the construction of the m_T variable, since it is obtained using a single lepton. This is solved by constructing the m_T distribution using only the leading lepton, i.e. the lepton with the highest p_T .

The Drell-Yan process dominates events with dilepton final states around the Z boson mass. Therefore, events with opposite-charge same-flavor leptons within the invariant mass range of Z -boson, i.e. $76 \text{ GeV} < m_{ll} < 106 \text{ GeV}$ are rejected in order to obtain a rather pure $t\bar{t} \rightarrow l\bar{l}$ yield.

The $t\bar{t} \rightarrow l\bar{l}$ process is expected to have two jets from the hard interaction in the final state, thus the events passing the preselection contain additional jets due to initial and final state radiation. A good agreement between the two jet-multiplicity distributions is observed in the constrained signal regions and is demonstrated in Fig. 7.7 for one benchmark point.

In CR3, the $t\bar{t} \rightarrow l\bar{l}$ modeling is tested without the isolated-track and τ -lepton rejection by reverting these veto algorithms. Tail-to-peak ratios obtained in CR3 are very close to unity. Small deviations are observed in the samples with very low event yields. These deviations are well within the statistical uncertainties. The centrality and number of jets distributions for a benchmark point is shown in Fig. 7.8.

Overall a good description of the $t\bar{t} \rightarrow l\bar{l}$ process is observed in the MC samples.

7.3 Background Estimation

The discrepancies between the data and MC samples in CR1 are corrected by scaling the MC samples with the tail-to-peak ratio obtained from the data. These scale factors are also applied to the search regions. All background MC samples are normalized to data in the m_T -peak region. A larger sample size for the m_T -peak is obtained for the $t\bar{t} \rightarrow l\bar{l}$ sample by removing the isolated track and τ -lepton rejection requirements, hence reducing the statistical uncertainty. The background validation studies in CR3 ensure that a good description of the data is given by the $t\bar{t} \rightarrow l\bar{l}$ MC simulation of the *pre- τ* lepton and isolated track rejection region.

The following methodology used to estimate all background yields is discussed in detail below for the benchmark point ($m_{\tilde{t}} = 275 \text{ GeV}$, $m_{\tilde{\chi}^0} = 50 \text{ GeV}$) for the muon sample:

- The m_T -peak normalization factors for $t\bar{t} \rightarrow l\bar{l}$ are calculated without the isolation criteria to obtain a larger $t\bar{t} \rightarrow l\bar{l}$ sample size:

$$\begin{aligned} f_{\text{pre}}^{\mu} &= (\text{Data}^{\text{peak}} - \text{MC}_{\text{Rare}}^{\text{peak}}) / (\text{MC}_{t\bar{t} \rightarrow l\bar{l}} + \text{MC}_{l\text{top}}^{\text{peak}} + \text{MC}_{W+\text{jets}}^{\text{peak}}) \\ &= (397 - 10.72) / (24.71 + 270.7 + 28.80) = 1.19. \end{aligned} \quad (7.1a)$$

- After normalizing the $t\bar{t} \rightarrow l\bar{l}$, a similar factor (f) is obtained for the $l\text{top}$ and $W + \text{jets}$ in

the SR:

$$\begin{aligned} f^\mu &= (\text{Data}^{\text{peak}} - f_{\text{pre}}^\mu \times \text{MC}_{\text{tt} \rightarrow \text{ll}}^{\text{peak}} - \text{MC}_{\text{Rare}}^{\text{peak}}) / (\text{MC}_{\text{ll top}}^{\text{peak}} + \text{MC}_{\text{W+jets}}^{\text{peak}}) \\ &= (347 - 1.19 \times 9.21 - 7.20) / (242.21 + 27.51) = 1.22. \end{aligned} \quad (7.2a)$$

- The obtained normalization factors (f , f_{pre}) together with the tail-to-peak ratios ($r_{\text{W+jets}}$, $r_{\text{ll top}}$) are used in the estimation of the final yields in the search regions:

$$\text{tt} \rightarrow \text{ll} = f_{\text{pre}} \times \text{MC}_{\text{tt} \rightarrow \text{ll}}^{\text{tail}} = 1.19 \times 28.92 = 34.42 \quad (7.3a)$$

$$\text{ll top} = f \times r_{\text{ll top}} \times \text{MC}_{\text{ll top}}^{\text{peak}} = 1.22 \times 0.08 \times 242.21 = 23.64 \quad (7.3b)$$

$$\text{W+jets} = f \times r_{\text{W+jets}} \times \text{MC}_{\text{W+jets}}^{\text{peak}} = 1.22 \times 0.18 \times 27.51 = 6.04 \quad (7.3c)$$

$$\text{Rare} = \text{MC}_{\text{Rare}} = 4.53 \quad (7.3d)$$

Therefore, the single-muon SM prediction for the SR ($m_{\tilde{t}} = 275 \text{ GeV}$, $m_{\tilde{\chi}^0} = 50 \text{ GeV}$) is 68.63 events, whereas the corresponding data yield is 67 events. The combined electron and muon sample distributions for the benchmark point is shown in Fig. 7.9.

An important result that can be drawn from Fig. 7.9 is that following the rough similarity between the signal and background distributions, the background events passing the SVM-HINT

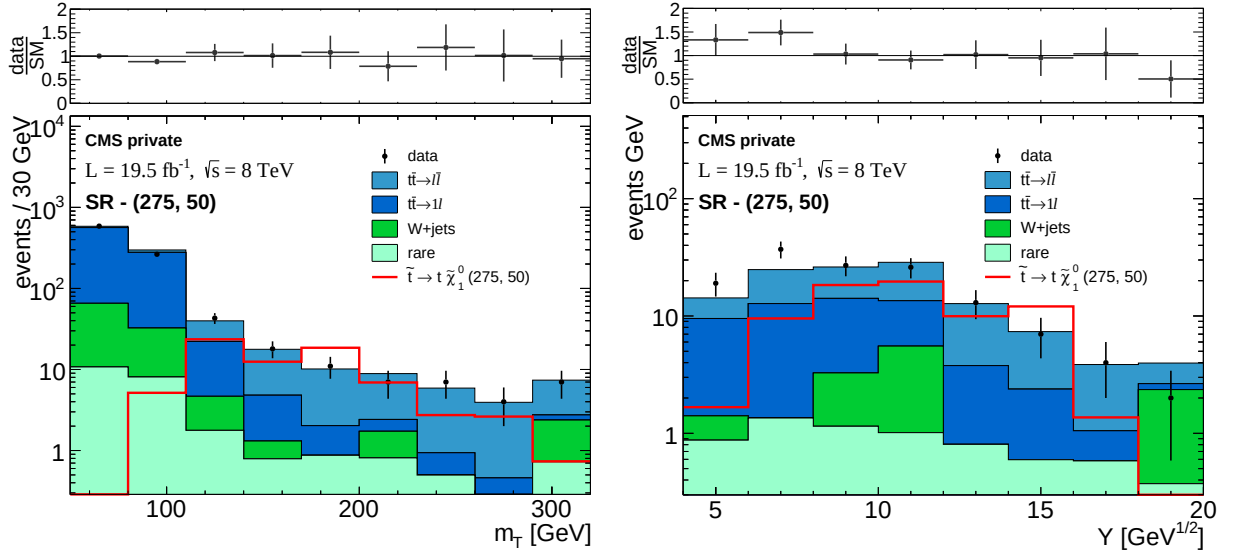


Fig. 7.9: Distributions of the m_T (left) and Y (right) are presented for the benchmark signal sample ($m_{\tilde{t}} = 275 \text{ GeV}$, $m_{\tilde{\chi}^0} = 50 \text{ GeV}$) in the SR. The tail distributions are normalized to the estimated event multiplicities using the $r_{\text{ll top}}$, $r_{\text{W+jets}}$ and f for the ll top and W+jets backgrounds, and f_{pre} for the $\text{tt} \rightarrow \text{ll}$ background samples. The last bin includes overflow events. The estimated distributions from different background processes are stacked, while the data and signal distributions are individually presented.

discriminator threshold, or the events selected within the SVM hyperplane, have similar topologies to the signal processes.

The background estimation procedure is applied for all search regions to obtain the background estimations. The interpretation of the results with the corresponding uncertainties is discussed in the next chapter.

Results and Interpretation

Contents

8.1	Systematic uncertainties	123
8.1.1	Systematic uncertainties on the background estimation	124
8.1.2	Systematical uncertainties on the signal samples	126
8.2	Interpretation and comparison	129
8.2.1	Comparison with similar studies	131

SVM-HINT encapsulates multidimensional *hyperspaces* with highest purity of the rare signal events. The SM yields in these small hyperspaces are estimated using a data-driven technique. In this chapter, as the last part of the quest, the SM predictions are tested with the observed data, and the results are interpreted within the SUSY signal models.

Such an interpretation would not be meaningful without the understanding of the uncertainties. Therefore, in the first section, the systematic uncertainties are estimated for both signal and background yields, which is then followed by the statistical inference of the obtained results. In the last section, similar searches are compared to the present analysis.

8.1 Systematic uncertainties

The systematic uncertainties contain all uncertainties that are not due to finite size of the estimated event yields. However, the distinction between statistical and systematic uncertainties can be blurry in particular cases due to finite size of the samples used in the event yield estimation. In such cases, the statistical uncertainties should be considered and propagated accordingly. To prevent such an ambiguity, each uncertainty source that is not due to the finite sample yields in the search regions is referred to as systematic uncertainty in this thesis.

Systematical uncertainties can be estimated with an ML classifier following two different approaches: A rather optimistic recipe requires re-training of the classification models for each

systematic uncertainty variation and evaluating spreads around the expected value. In the conservative method such variations are only applied on the evaluation samples, and may result with suboptimal training, resulting in wider spreads. In the present analysis, the latter approach is followed, leading to higher systematic uncertainties compared to the rather simple threshold based analysis (also known as *cut-and-count*) [79, 110]. Nevertheless, due to low event yields, the uncertainties on the estimated SM yields are dominated by the statistical fluctuations.

The majority of the uncertainties described in this section depend on the individual events yields of different search regions. Therefore, these uncertainties have to be determined in each 212 search regions separately, as each SR might contain a different event content due to the individual SVM training. The non-negligible sources of systematic uncertainties are described in the following.

8.1.1 Systematic uncertainties on the background estimation

The systematic uncertainties in the expected background yields for each SR are individually estimated. The effects of the uncertainties that propagate to the estimation of the related uncertainty sources are considered.

Uncertainty due to m_T peak region choice

The effect of the m_T peak choice is investigated by varying the m_T -peak region to $40 \geq m_T \geq 100$. Since the event yield in the initial peak region is at least two orders of magnitude higher than the region around the peak, all deviations due to such a variation are within the statistical uncertainty.

Statistical uncertainties due to the limited number of events in the MT peak regions

As mentioned in Chapter 7, normalizing background events in the m_T peak region eliminates some of the systematic uncertainty sources, however, statistical errors due to the finite statistic in the m_T peak-regions propagates to the background estimations in the search regions. These uncertainties also affect the tail-over-peak ratios obtained from CR1.

Uncertainty on the tail-over-peak ratios for the $W + \text{jets}$ and $1l$ top samples

The background estimation for the $W + \text{jets}$ and $1l$ top processes relies on the tail-to-peak ratio of the corresponding MC samples in the control regions. The statistical uncertainties on these ratios are considered in the search regions. The statistical uncertainties in the control regions are lower than in the search regions due to separation of the training, test, and evaluation samples in the search regions. Therefore, the uncertainties propagated from the control regions $r_{W+\text{jets}}$ and $r_{1l\text{top}}$ as well as $\text{SF} \cdot r_{W+\text{jets}}$ and $\text{SF} \cdot r_{1l\text{top}}$ have a lower impact on the overall uncertainty.

Uncertainty on the $W + \text{jets}$ and rare processes cross-sections

A 50% uncertainty is assigned to the cross-sections of the $W + \text{jets}$ and rare processes.

$t\bar{t} \rightarrow l\bar{l}$ background studies

The modeling of the underlying physics processes in the POWHEG generated $t\bar{t} \rightarrow l\bar{l}$ background samples can be tested using alternative MADGRAPH samples. Following the background estimation methodology described in Chapter 7, the expected event yields are determined for the following systematic variations:

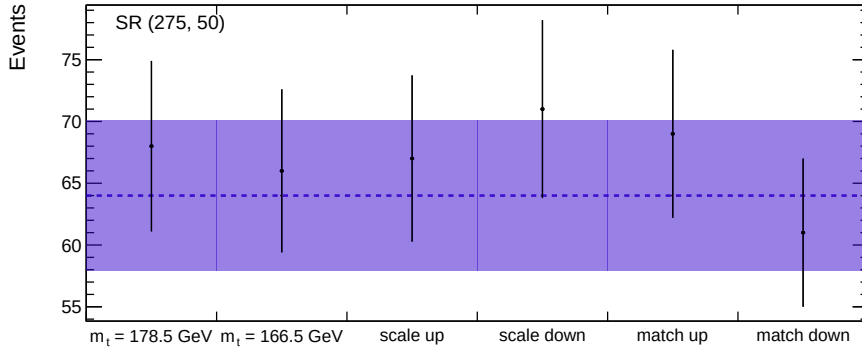


Fig. 8.1: The $t\bar{t} \rightarrow l\bar{l}$ yields for different background modeling studies are presented for the benchmark search region $m_{\tilde{t}} = 275 \text{ GeV}$, $m_{\tilde{\chi}^0} = 50 \text{ GeV}$. The blue dashed-line represents the $t\bar{t} \rightarrow l\bar{l}$ yield estimated from the default POWHEG sample, and the corresponding uncertainty is represented by the blue transparent band. Variations on the MADGRAPH samples are within the statistical uncertainty of the default POWHEG sample.

- The top-quark mass is varied around the central value by 6 GeV: $m_t = 166.5 \text{ GeV}$ and $m_t = 178.5 \text{ GeV}$.
- Alternative samples with variations $\times 2$ and $\times 0.5$ from the central value given by the geometric mean of $Q^2 = m_t^2 + \sum_{\text{jets}} p_T^2$ for each particle in the final state are used to estimate the effect of renormalization and factorization scales' variations.
- The matching scale described in Chapter 3 has the nominal value of $x_q > 40 \text{ GeV}$. The background estimation is repeated for the alternative values $x_q > 30 \text{ GeV}$ and $x_q > 60 \text{ GeV}$.
- The modeling of the τ -lepton decay is simulated with the TAUOLA MC generator. A reference MADGRAPH sample without TAUOLA is used to investigate the impact of the TAUOLA.

The variations are within the statistical uncertainties of the POWHEG sample. Therefore, no additional uncertainty is accounted for these variations. The $t\bar{t} \rightarrow l\bar{l}$ background studies are presented for the benchmark point in Fig. 8.1.

Summary of the systematic uncertainties combined with the statistical uncertainties is shown

in Fig. 8.2. An uncertainty of 40% is applied to the points where the SVM-HINT optimization does not convert or no background event is expected.

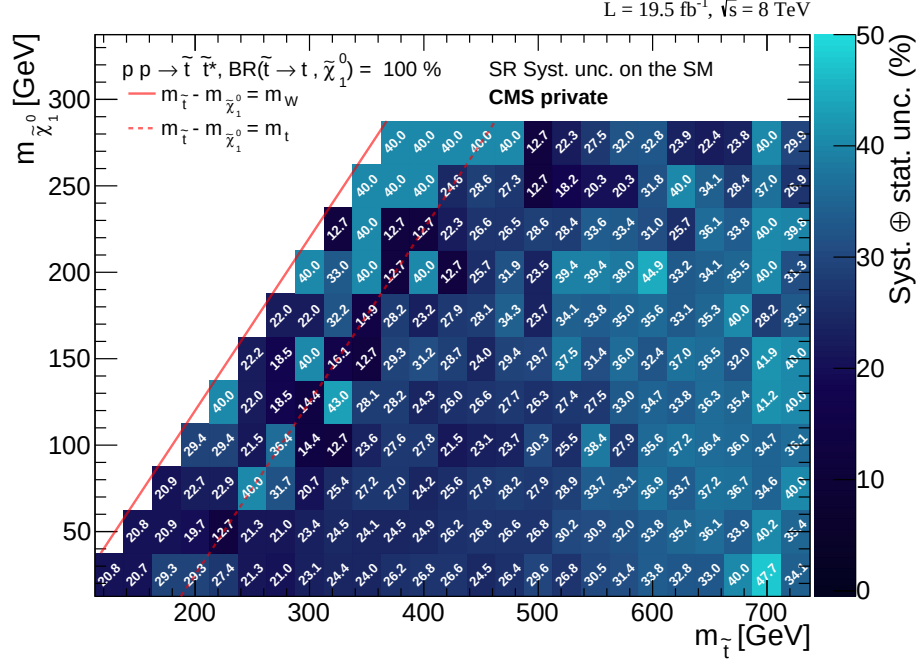


Fig. 8.2: The combination of systematic and statistical uncertainties on the background yields for the search regions. The high- ΔM region is dominated by the statistical uncertainties reaching up to 47%.

8.1.2 Systematical uncertainties on the signal samples

Sources of the systematic uncertainties considered in the present analysis are given in the following:

Integrated luminosity

The luminosity measurement in the CMS experiment has an uncertainty of 2.6%(cf. Chapter 3).

Trigger, lepton-identification and isolation efficiencies

An overall uncertainty of 3% is assigned due to slight variations with respect to lepton p_T and η for the trigger efficiency as described in Section 5.2.

An uncertainty of 5% is applied for the lepton-identification and isolation efficiency due to the 5% discrepancy presented in Section 5.2, which is measured using a tag-and-prob method on $Z^0 \rightarrow ll$ events [79, 80].

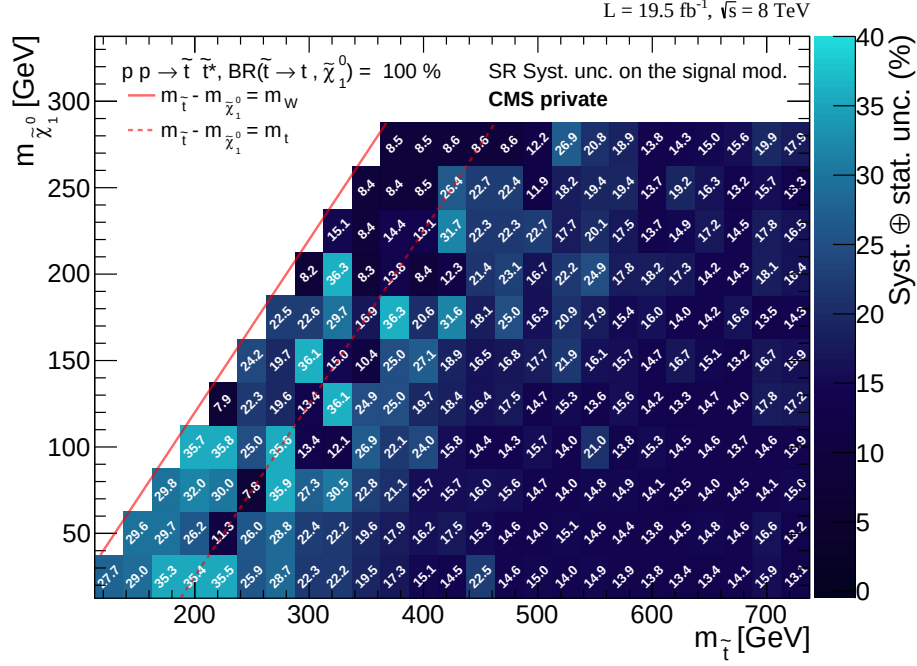


Fig. 8.3: The combination of systematic and statistical uncertainties on the signal yields. The low- ΔM region is dominated by the statistical uncertainties due to higher event weights in the signal samples, reaching up to 40%. The uncertainty on the signal model cross-sections are not included.

b-tagging efficiencies

Following the recommended recipe from the CMS b-tag and vertexing POG [111], the uncertainty on the b-tagging efficiency is taken into account by varying the b-tagging discriminator reshaping parameters for each signal sample. These variations are propagated to all variables used in the evaluation of the SVM models and the background estimation. The signal efficiency is estimated for each b-tagging variation and the quoted uncertainty is defined as half of the spread between the extrema of these values.

Jet energy scale

After applying JEC (cf. 5), the discrepancies in the jet energy measurements of the MC samples and the data in the predefined control regions are quantified with the Jet Energy Scale (JES) factors. The energy of the jets is scaled up and down within 1σ Gaussian spread following the recommendations of the CMS Jet POG [112]. The signal efficiency is estimated for both up and down variations and the uncertainty is defined as half spread between the extrema.

Initial state radiation

Due to the imperfections in the initial state radiation modeling, the MADGRAPH p_T spectrum of the system recoiling against initial-state radiation jets is compared to the data in Z/γ^* , $t\bar{t}$, and

WZ samples [79]. Weights ranging from 0 – 20% depending on the p_T of the recoiling system are applied to the MADGRAPH signal samples used in the present analysis. The full values of the corrections are taken as systematic uncertainty.

The combination of the statistical and systematic uncertainties is shown in Fig. 8.3.

8.2 Interpretation and comparison

Following the analysis strategy described in Chapter 7, and the estimated uncertainties discussed in the previous section, the SM background estimation and the data compatibility is represented in Fig. 8.4.

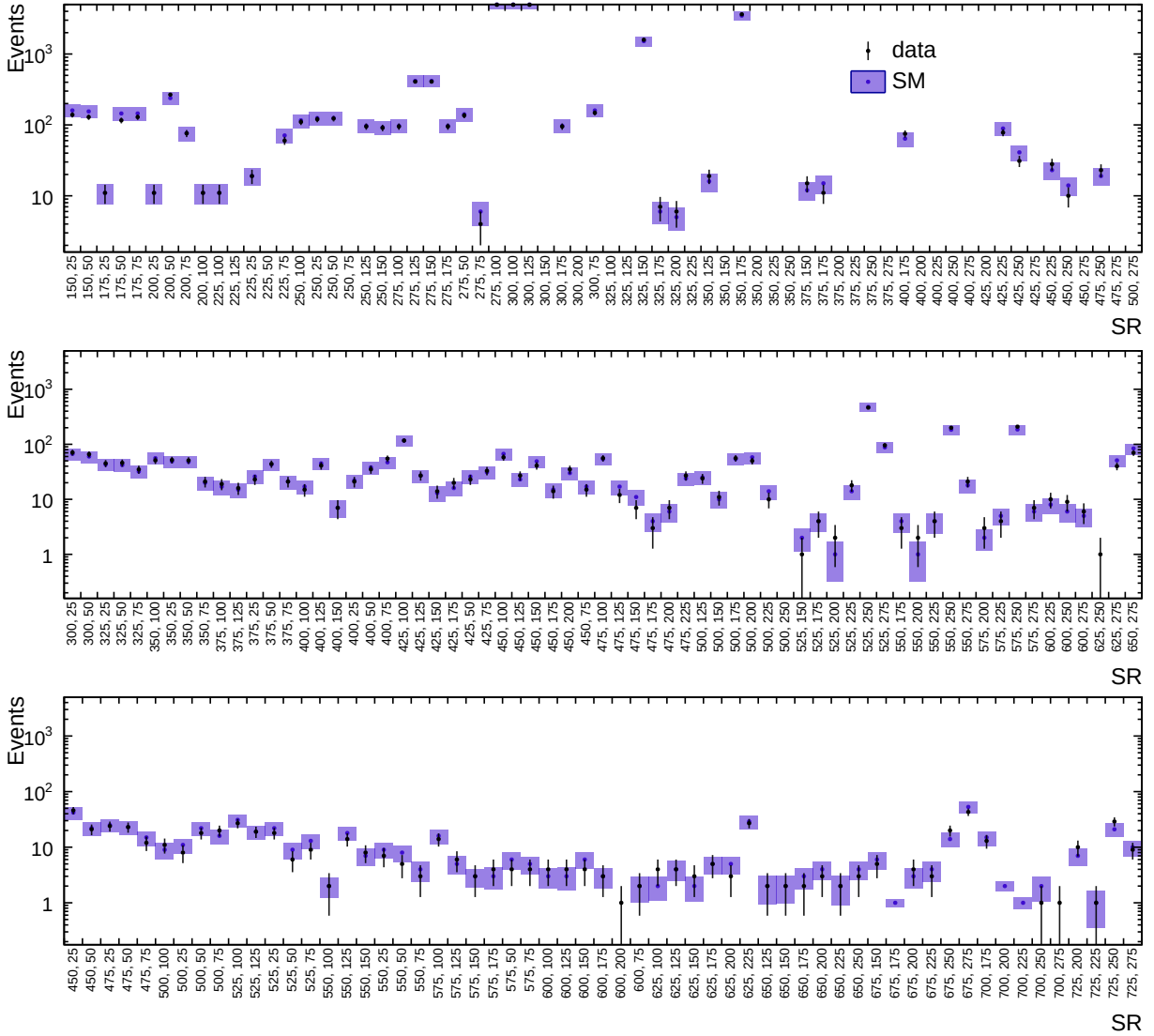


Fig. 8.4: The event yields for the SM background prediction and the observed data for all search regions optimized with SVM-HINT. The combined statistical and systematic uncertainties are represented by the transparent rectangles. The data show good agreement with the background estimation. The event yields are decreasing with the higher $m_{\tilde{t}}$ following the signal model cross-sections. Empty SR bins correspond to the SVM models that could not converge within the signal efficiency restrictions or they do converge with zero background estimation. signal samples with lower sensitivity causes suboptimal hyperparameter tuning, thus resulting in relatively high event yields (i.e. higher than 4000 events).

The observed data yields' compatibility are well within two σ spread (reaching the maximum discrepancy, and exceeding the expected yields in SRs $m_{\tilde{t}} = 675$ GeV, $m_{\tilde{\chi}^0} = 250$ GeV and $m_{\tilde{t}} = 725$ GeV, $m_{\tilde{\chi}^0} = 250$ GeV). Therefore, following the limit setting procedure described in Appendix C, an upper limit on the expected T2tt signal models can be obtained.

The observed CL_s limits with the $\pm 1\sigma$ acceptance band corresponding to about 15% theoretical uncertainty on the signal model's cross-sections and the expected CL_s within the $\pm 1\sigma$ range are shown in Fig 8.5.

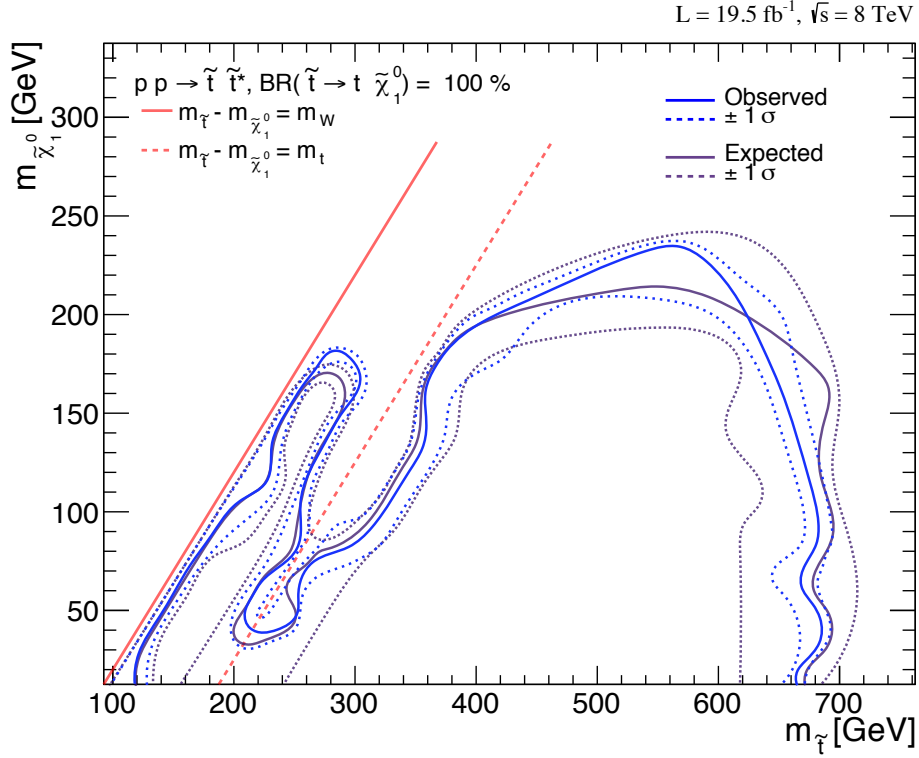


Fig. 8.5: Exclusion limits at 95% CL for direct top-squark production with decay $\tilde{t} \rightarrow t \tilde{\chi}^0$ at 8 TeV center-of-mass energy. The interpretation is done within the T2tt simplified model. The solid blue curve shows the observed limit, whereas the solid purple curve represents the expected limits in the two dimensional space of $m_{\tilde{t}}$ and $m_{\tilde{\chi}^0}$. The $\pm 1\sigma$ uncertainty bands are represented by the dashed lines.

As expected the sensitivity of the exclusion heavily depends on the ΔM : closer to the top-quark mass, there is rather weak classification power of the present analysis. In the signal samples with higher ΔM , the phase space for the natural SUSY models are constrained up to 700 GeV in $m_{\tilde{t}}$ and up to 250 GeV in $m_{\tilde{\chi}^0}$.

The expected limits, thereby the statistical significance boundary presented in Fig. 8.5 show similar characteristic to the Asimov significance obtained in Chapter 7. The differences between Fig. 7.4 and Fig. 8.5 are due to the discrepancies between the background MC samples and the SM, as well as the constant systematic uncertainties (25%) assumed in SVM-HINT.

8.2.1 Comparison with similar studies

Independent direct $\tilde{t}$ searches in the single lepton final states have been performed by both the CMS and ATLAS collaborations. These searches have in general similar sensitivities.

Figure 8.6 gives an overview of the CMS direct $\tilde{t}$ searches with various final states. The analysis of interest (represented by the orange line) uses BDTs (cf. Chapter 6) for the classification of the signal models. The training samples are enriched by merging signal samples within a ΔM range. Even though this training strategy simplifies the background estimation procedure and increases the training sample event size significantly, it fails to describe the underlying distributions for each signal sample within the whole ΔM range as described in Chapter 7. The state-of-the-art

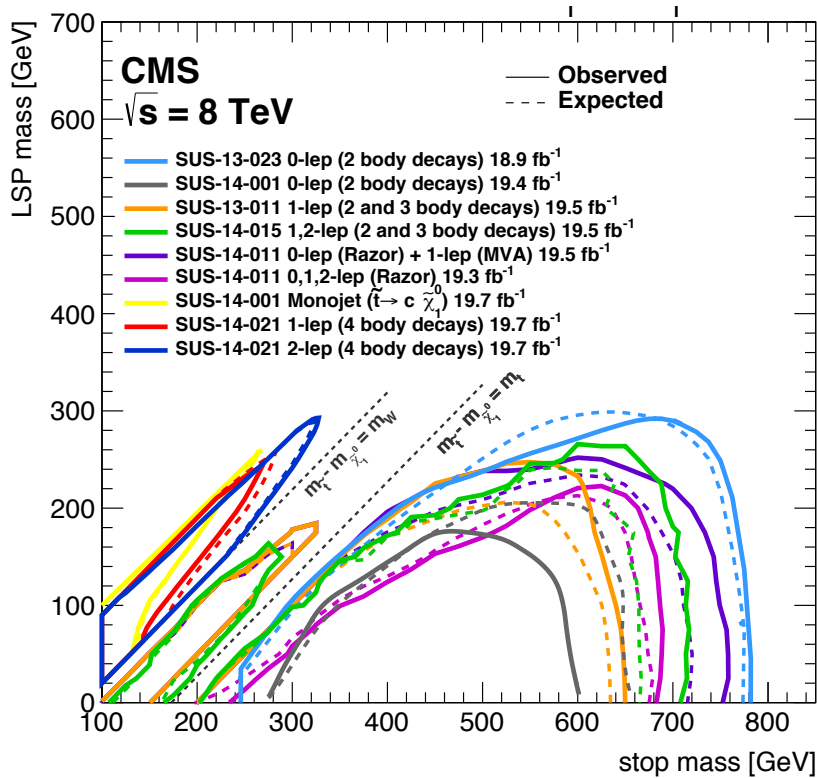


Fig. 8.6: gives an overview of the CMS direct $\tilde{t}$ searches with various final states [25]. The results of the analysis in the single lepton final states is represented by the orange line.

classification algorithm implemented in SVM-HINT together with a more granular optimization strategy provides a better sensitivity overall in both low and high ΔM regions even with higher uncertainties (due to separation of the training, test, and evaluation samples to prevent bias in the ML algorithm). The sensitivity is improved by as much as 75 GeV at the high $m_{\tilde{t}}$ region. Even though very small, an additional sensitivity is gained in the $\Delta M = m_t$ region.

The comparison of the present analysis to the searches in different final states is considered as follows:

- 0-lepton: Due to the $t \rightarrow bW$ decay, the branching ratio of W boson to hadron and lepton

final states significantly affect the event rates of these two different final states. The W boson is more than two times more likely to decay into all-hadronic than leptonic final states. This difference is especially pronounced in the high $m_{\tilde{t}}$ region as shown in Fig. 8.6. In the low ΔM region, these analyses have little to no sensitivity.

- 2-lepton: Following the same argument, because of rather low cross section observed in the dilepton final states, the searches are mainly covering very low ΔM regions.

The present analysis interpreted within $m_W < \Delta M$ parameter space of the T2tt simplified model. The searches in the hadronic final states provide higher sensitivity in high $m_{\tilde{t}}$ region. The present analysis constrains the T2tt model in a larger configuration space and including the compressed region as shown in Fig. 8.6.

A similar study is performed by the ATLAS collaboration, leading to comparable results. Similarly, the sensitivity of the present analysis exceeds the observed limits in both low and high ΔM samples comparing to this analysis as seen in Fig. 8.7.

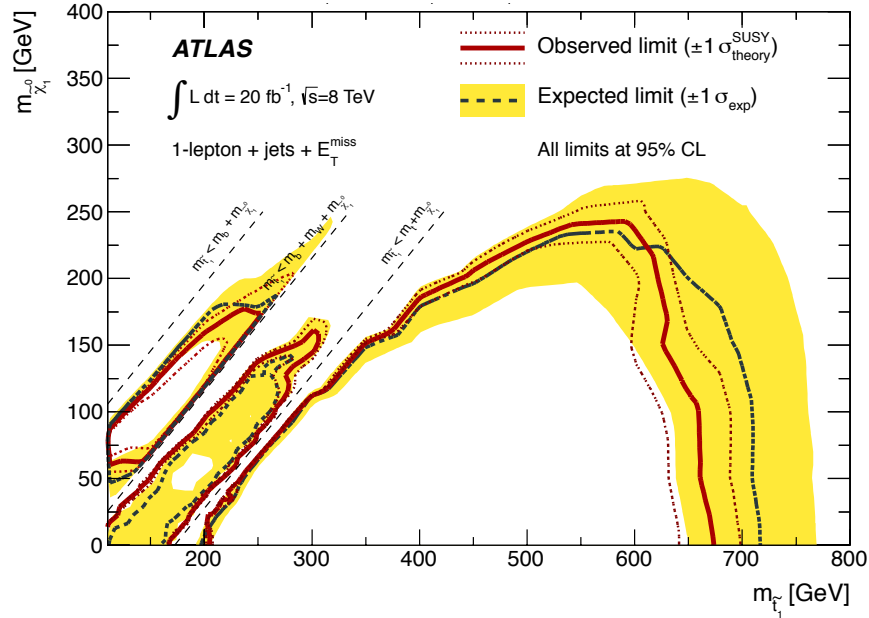


Fig. 8.7: The direct $\tilde{t}$ search in the single lepton final states [113]. Observed limits are represented with the solid red line, whereas the black dashed line shows the expected limits.

In conclusion, the present analysis provides the best sensitivity among direct $\tilde{t}$ searches in single lepton final states at 8 TeV center of mass energy. The natural SUSY models predicting $m_{\tilde{\chi}^0} < 50$ GeV are constrained significantly.

Conclusion and Outlook

The Standard Model (SM) of particle physics is a breathtakingly successful model, and its predictions have been confirmed by various experiments. After the discovery of the Higgs boson, the particle content of the SM is completed. However, the SM is unable to answer questions related to unaccounted experimental observations. Combined with the extreme fine-tuning requirements on the Higgs boson mass, the SM shows similar characteristics with effective theories with a limited validity. By imposing the naturalness argument, the validity scale of the SM can be determined as up to the TeV scale. Therefore, exploring the TeV scale is both exciting and promising for experimental physicists.

Various theories beyond the SM (BSM) are hypothesized. One of the most appealing BSM is the supersymmetric extension of the SM (SUSY) which relates the fermions to bosonic superpartners and bosons to fermionic superpartners. SUSY not only offers explanations to the unaccounted experimental observations, but also provides an aesthetically pleasing theory. Scalar top-quark partners with their mass around or less than a TeV cancel the divergent corrections on the Higgs mass, and with the R-parity conservation, SUSY provides a cold-dark matter candidate in the form of lightest SUSY particles. Therefore, in this thesis, a search for pair production of the top quark supersymmetric partners has been performed.

The cross section of the top-squark pair production decreases steeply with the increasing top-squark mass, resulting in extremely scarce signal event multiplicity. Therefore, making the most out of the data is absolutely crucial. Within the thesis study, a new support vector machine HEP interface (SVM-HINT) has been developed in order to efficiently extract information from the data. A novel hyperparameter optimization algorithm based on the Asimov discovery significance estimator has been introduced with SVM-HINT. A 14 TeV benchmark study has been performed to compare SVM-HINT with the popular TMVA-BDT implementation. SVM-HINT achieves up to 40% higher statistical significance compared to the TMVA-BDT implementation.

SVM-HINT is employed in the search for top-squark pair production in the pp-collision events of a center-of-mass energy of 8 TeV in the CMS experiment. SVM-HINT optimizes SVM models for each signal sample with different $m_{\tilde{\chi}^0}$ and $m_{\tilde{t}}$ configurations. Each SVM model reserves a different phase space or search region in the Monte Carlo background samples. Due to discrepancies

observed in statistically independent control regions, a data-driven method is used to estimate the background yields in each of the 212 search regions.

The results are interpreted within a simplified model describing direct top-squark pair production, where each top-squark decays to a top quark and a neutralino. Considering systematic and statistical uncertainties, the observed data are compatible with the SM background estimation. Therefore, upper limits are set on the masses of the top-squark ($m_{\tilde{\chi}^0}$) and the neutralino ($m_{\tilde{t}}$) of up to $m_{\tilde{t}} = 675$ GeV with 10 fb signal cross-section, and up to $m_{\tilde{\chi}^0} = 225$ GeV are excluded with 95% confidence. The signal configurations where the mass difference between $m_{\tilde{t}}$ and $m_{\tilde{\chi}^0}$ is close to t -quark mass (also named as *compressed* region) have similar topologies as the $t\bar{t}$ background. The present analysis constrains the compressed region up to $m_{\tilde{t}} = 200$ GeV and $m_{\tilde{\chi}^0} = 25$ GeV. These results make the present analysis the most sensitive search for pair production of the top-squarks in single-lepton final states at 8 TeV. Comparing to the CMS and ATLAS analyses, the sensitivity is improved up to 75 GeV in $m_{\tilde{t}}$ and $m_{\tilde{\chi}^0}$.

The observed results of the present analysis strengthen the necessity to have higher luminosity and center-of-mass energies at the LHC experiments. With this purpose, the LHC and the CMS detector have been upgraded. The outdated CMS-HCAL's electronics infrastructure is replaced with faster and more reliable modules. Within this upgrade effort, the front-end readout control system has been replaced with a new μ TCA based infrastructure. The next generation Front End Control (ngFEC) system has been developed and a new firmware for the ngFEC FPGAs has been designed within the thesis studies. The ngFEC provides a fast, reliable, compact and unified control structure for the HCAL front-end readout electronics. It simultaneously setups and controls all front-end readout, and it recovers diagnostic information and responses immediately in case of unexpected events. The system is successfully tested in various test stands and irradiation tests. It is proven to be capable of withstanding irradiation levels equivalent to a radiation exposure at the CMS in 3000 fb^{-1} at 13 TeV.

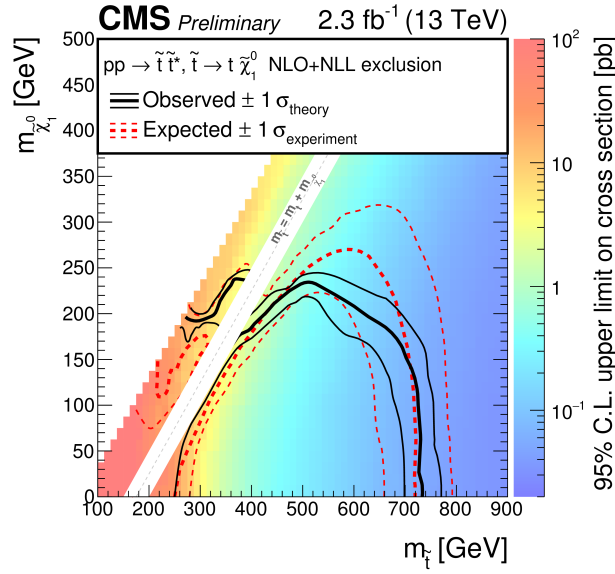


Fig. 9.1: Exclusion limits at 95% CL for direct top-squark production with decay $\tilde{t} \rightarrow t \tilde{\chi}^0$ at 13 TeV center-of-mass energy. The interpretation is done within the T2tt simplified model. The solid black curve shows the observed limit in the two dimensional space of $m_{\tilde{t}}$ and $m_{\tilde{\chi}^0}$ [23].

Outlook

The top-squark pair production search has been repeated with 2.3 fb^{-1} pp-collision data at the 13 TeV center-of-mass energy [23] by the CMS collaboration. The observed limits are shown in Fig. 9.1. Due to relatively low integrated luminosity and steeply falling $\tilde{t}$ pair-production cross-section, a limited sensitivity improvement in the high $m_{\tilde{t}}$ region has been accomplished in spite of a $1.6\times$ center-of-mass energy increase. Therefore, a higher integrated luminosity is necessary to unveil the top-squark possibly residing around TeV energy scale.

In the next few years the LHC is expected to deliver 300 fb^{-1} pp collision data at 13 TeV center-of-mass energy. The pair production cross-section of top-squarks with $m_{\tilde{t}} = 1 \text{ TeV}$ is 6 fb at this energy. Therefore, sensitivity of the top-squark analysis is naively expected to reach $m_{\tilde{t}} = 1 \text{ TeV}$ region. However, the compressed region will require a dedicated analysis. Use of Machine Learning techniques (like Support Vector Machines) may significantly boost the sensitivity in this region.

Overall, whether to continue with natural SUSY models or not will be decided at the end of the LHC run in a few years, and maybe new scenarios will be constructed and considered. One thing to be sure is that exciting times are ahead, and closer than ever.

Appendices

Gigabit Transceiver Protocol

The Gigabit Transceiver (GBT) [60] protocol is specifically designed to establish communication with the radiation tolerant GBTx ASICs. The GBT provides a gigabit communication link between GBTx chip placed in the front-end detector electronics and the back-end FPGA which is responsible for the front-end control system. It can be also employed in the scenarios where the communication system is lacking the GBTx chip.

The GBT logic transfers 120-bit frame (parallel) at 40 MHz. The transmitter scrambles the frame to prevent consecutive errors due to possible problems in the system, serializes (in FPGAs using on chip SerDes functionality), and send the serialized bits to the receiving side via an optical fiber cable at 4.8 Gbps. The receiver logic follows the same order reversely. Since it is particularly designed for the irradiated environments, it employs a 32-bit Reed-Solomon forward error correction (RS-fec) algorithm.

The suggested GBT-frame scheme consists of 88 bits user data (including 4 bits reserved for the slow controls), 4 bits header, and 32 bits RS-fec as shown in Fig. A.1. The 32-bit RS-fec is particularly efficient against the bursts of bit errors caused by single event upsets. It can correct up to 20 consecutive erroneous bits. The GBT protocols offers other schemes all providing similar functionality but different scrambling and forward error correction strategies. In the present work, the GBT frame is used.

The ngFEC-GBT frame includes both slow and fast control signals and is delivered to the front-end modules. A novel I²C-over-GBT algorithm is implemented to simultaneously control multiple

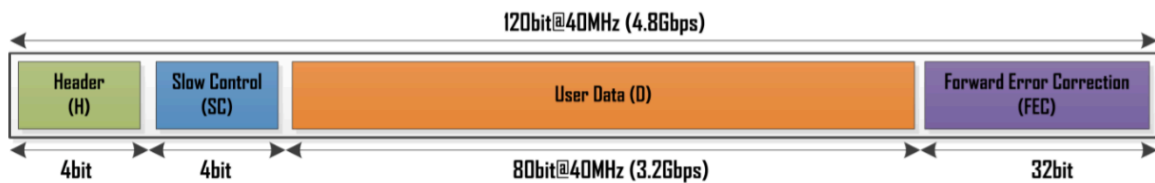


Fig. A.1: The 120-bit frame of the GBT protocol. The bits are refreshed with 40 MHz clock, thus, reaching a transfer rate of 4.8 Gbps [60].

front-end I²C buses as discussed in Chapter 4. This communication can be tested by monitoring the slow and fast control signals from the front-end crate's backplane slots as well as by checking the debugging registers in the ngCCM and ngFEC. Figure A.2 shows the slow and fast control signals monitored by 1 GHz sampling rate oscilloscope.

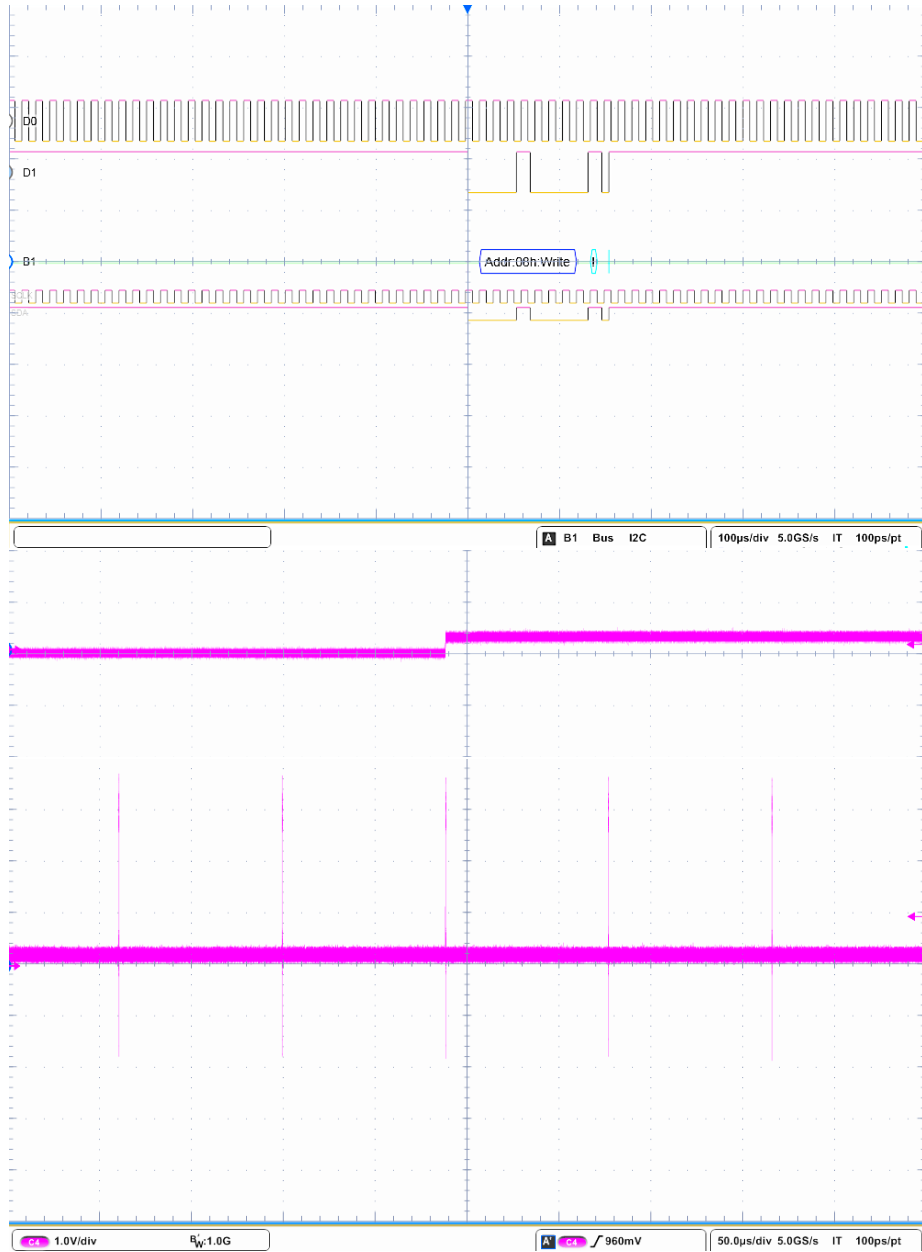


Fig. A.2: Screenshots of the 1 GHz sampling rate oscilloscope which is probing one of the front-end slots. The figure on top is the oscilloscope screenshot showing a write request from the ngFEC I²C master corresponding to the front-end crate slot that the oscilloscope probes are connected to. D1 is the SDA and D0 is the SCL line of the I²C bus. The figure in the middle shows the power enable signal and the bottom figure is the QIE reset signal (40 ns width) with 88.93 µs period.

In order to detect errors that pass the RS-fec, an additional safety measure in the form of Pseudo Random Bit Stream (PRBS) - bit error rate test (BERT) is implemented in the ngFEC firmware. Various length of PRBS (23 bits for the HF and 15 bits for HE/HB) inserted into the GBT-frame user data to spot errors that cannot be corrected or mis-corrected by the RS-fec.

Lepton Identification and Isolation Efficiency

The identification and isolation efficiencies for muons and electrons are measured using a tag-and-probe method. In the tag-and-probe method, events passing a looser selection criteria (probe) are compared to a subset of these events passing through a tighter selection criteria (tag). The tagged events are required to pass the pre-selection criteria and be matched to the single-lepton triggers with a p_T threshold of minimum 20 GeV, whereas a lower p_T threshold is imposed on the probe events: the efficiencies are measured by requiring isolation or identification criteria separately for the probe events. The discrepancies in the data and MC are taken as scale factors and are quoted in Table B.1 and Table B.2 for muons, and Table B.3 and Table B.4 for electrons.

Table B.1: Muon identification scale factors [79, 80].

Scale Factor ID p_T [GeV]	$ \eta < 0.8$	$0.8 < \eta < 1.5$	$1.5 < \eta < 2.1$
20 - 30	0.9839 ± 0.0006	0.9850 ± 0.0008	0.9876 ± 0.0010
30 - 40	0.9850 ± 0.0003	0.9846 ± 0.0004	0.9890 ± 0.0006
40 - 50	0.9865 ± 0.0003	0.9866 ± 0.0003	0.9902 ± 0.0005
50 - 60	0.9829 ± 0.0006	0.9834 ± 0.0007	0.9864 ± 0.0012
60 - 80	0.9835 ± 0.0012	0.9818 ± 0.0015	0.9909 ± 0.0024
80 - 100	0.9785 ± 0.0031	0.9803 ± 0.0039	0.9995 ± 0.0070
100 - 150	0.9847 ± 0.0042	0.9765 ± 0.0054	0.9884 ± 0.0102
150 - 200	0.9958 ± 0.0101	1.0064 ± 0.0145	0.9613 ± 0.0279
200 - 300	0.9937 ± 0.0215	0.9867 ± 0.0339	0.9652 ± 0.0720
300 - ∞	0.9754 ± 0.0663	1.0348 ± 0.1693	0.4286 ± 0.4676

Table B.2: Muon isolation scale factors [79, 80].

Scale factor iso p_T [GeV]	$ \eta < 0.8$	$0.8 < \eta < 1.5$	$1.5 < \eta < 2.1$
20 - 30	0.9934 ± 0.0010	0.9974 ± 0.0011	1.0068 ± 0.0011
30 - 40	0.9969 ± 0.0003	1.0004 ± 0.0004	1.0039 ± 0.0004
40 - 50	0.9979 ± 0.0002	1.0001 ± 0.0002	1.0023 ± 0.0003
50 - 60	0.9985 ± 0.0005	1.0007 ± 0.0005	1.0042 ± 0.0006
60 - 80	0.9989 ± 0.0011	0.9997 ± 0.0011	1.0046 ± 0.0013
80 - 100	0.9999 ± 0.0031	1.0075 ± 0.0034	1.0086 ± 0.0042
100 - 150	1.0014 ± 0.0043	1.0056 ± 0.0049	1.0071 ± 0.0053
150 - 200	0.9802 ± 0.0109	1.0203 ± 0.0139	0.9582 ± 0.0129
200 - 300	1.0016 ± 0.0171	1.0059 ± 0.0200	1.0261 ± 0.0398
300 - ∞	0.9923 ± 0.0377	0.9822 ± 0.0681	1.0000 ± 0.0000

Table B.3: Electron identification scale factors [79, 80].

Scale Factor ID p_T range [GeV]	$ \eta < 0.8$	$0.8 < \eta < 1.4442$
20 - 30	0.9923 ± 0.0022	0.9632 ± 0.0022
30 - 40	0.9883 ± 0.0008	0.9707 ± 0.0010
40 - 50	0.9900 ± 0.0006	0.9755 ± 0.0008
50 - 60	0.9880 ± 0.0012	0.9777 ± 0.0017
60 - 80	0.9847 ± 0.0024	0.9797 ± 0.0032
80 - 100	0.9924 ± 0.0062	0.9687 ± 0.0081
100 - 150	0.9892 ± 0.0081	0.9813 ± 0.0110
150 - 200	1.0216 ± 0.0191	0.9940 ± 0.0286
200 - 300	0.9869 ± 0.0320	0.8853 ± 0.0408
300 - ∞	1.0789 ± 0.0854	1.0286 ± 0.1733

Table B.4: Electron isolation scale factors [79,80].

Scale Factor ISO p_T range [GeV]	$ \eta < 0.8$	$0.8 < \eta < 1.4442$
20 - 30	0.9938 ± 0.0015	0.9939 ± 0.0012
30 - 40	0.9968 ± 0.0004	0.9963 ± 0.0005
40 - 50	0.9973 ± 0.0002	0.9965 ± 0.0003
50 - 60	0.9957 ± 0.0005	0.9963 ± 0.0008
60 - 80	0.9962 ± 0.0012	0.9952 ± 0.0017
80 - 100	0.9992 ± 0.0035	1.0013 ± 0.0055
100 - 150	0.9964 ± 0.0052	0.9882 ± 0.0077
150 - 200	0.9861 ± 0.0117	1.0068 ± 0.0196
200 - 300	1.0025 ± 0.0256	1.0076 ± 0.0344
300 - ∞	1.1525 ± 0.0944	1.0084 ± 0.0926

Statistical Inference

The compatibility of the SM predictions and the collected data allows discovery or the rejection of the signal models within an exclusion boundary. A likelihood model can be constructed to quantify these models for the expected event yields.

This can be regarded as the test of two alternative models where the SM background represents the null hypothesis (H_0), and the SM background+SUSY signal models represents the hypothesis under test (H_1). According to the Neyman-Pearson lemma, the test can be achieved most efficiently using the log-likelihood ratio providing the highest discrepancy between H_0 and H_1 .

Event yields for the problem of interest can be parametrized as

$$\nu = \mu \cdot S(\theta) + B(\theta), \quad (\text{C.1})$$

where the $S(\theta)$ and $B(\theta)$ signal and background yields respectively, μ is the strength modifier of the signal model, and the θ is referred to as *nuisance* parameter. The degree of belief on θ to take a particular value can be summarized in terms of probability distribution functions $\rho(\theta|\tilde{\theta})$. Within the context of the present analysis, the systematical uncertainties discussed in Chapter 8, can be accounted as the nuisance parameters of the $\rho(\theta|\tilde{\theta})$.

The probability distribution function with uncertainties can be interpreted using the Bayes' theorem. By neglecting the normalization to the data prior

$$\rho(\text{theory}|\text{data}) \sim \rho(\text{data}|\text{theory}) \cdot \rho(\text{theory}), \quad (\text{C.2})$$

where $\rho(\text{theory})$ is the prior *probability* indicating the confidence on the theory estimation, $\rho(\text{data}|\text{theory})$ is called *likelihood* and $\rho(\text{theory}|\text{data})$ is the probability of obtaining the observed data yield given that the theory is correct. However, this quantity carries no value in the interpretation of the systematical uncertainties since the desired measure is the $\rho(\text{data}|\text{theory})$. Luckily, a similar relation is given as:

$$\rho(\text{data}|\text{theory}) \sim \rho(\text{theory}|\text{data}) \cdot \rho(\text{data}), \quad (\text{C.3})$$

The estimated uncertainty $\tilde{\theta}$ can be incorporated to a flat prior $\pi_{\theta}(\theta)$ to obtain :

$$\rho(\theta|\tilde{\theta}) \sim p(\tilde{\theta}|\theta) \cdot \pi_{\theta}(\theta) \quad (\text{C.4})$$

This Bayesian interpretation allows the treatment of the systematic uncertainties following a full frequentist approach. In this thesis, all background and signal uncertainties are treated as uncorrelated and each of them is described by a log-normal $\rho(\theta|\tilde{\theta})$:

$$\rho(\theta|\tilde{\theta}) = \frac{1}{\sqrt{2\pi} \ln(k)} \exp\left(-\frac{(\ln(\theta/\tilde{\theta}))^2}{2(\ln k)^2}\right) \frac{1}{\theta} \quad (\text{C.5})$$

where k is the width of the log-normal distribution which is set to be equal to the assessed uncertainty. The log-normal $\rho(\theta|\tilde{\theta})$ is chosen for its similarity to the Gaussian distribution and because it can describe the positively defined sources of the signal and background uncertainties.

Assuming a Poissonian spread for the counting experiment, the likelihood function $\mathcal{L}(\text{data}|\nu(\mu, \theta))$ is defined as:

$$\mathcal{L}(\text{data}|\nu(\mu, \theta)) = \text{Poisson}(\text{data}|\nu(\theta)) \cdot p(\tilde{\theta}|\theta) \quad (\text{C.6})$$

As discussed earlier the likelihood function can be interpreted as the probability of observing *data* as the outcome of the measurement, given the $\nu(\mu, \theta)$ is correct.

The test statistic providing the maximum discrepancy which compresses all the information available to discriminate the H_1 from the H_0 hypothesis is based on the profile likelihood ratio. This ratio is defined as [114]:

$$\tilde{q}_{\mu} = -2 \ln \frac{\mathcal{L}(\text{data}|\mu, \hat{\theta}_{\mu})}{\mathcal{L}(\text{data}|\hat{\mu}, \hat{\theta})}, \quad \text{with a constraint } 0 \leq \hat{\mu} \leq \mu \quad (\text{C.7})$$

where $\hat{\theta}_{\mu}$ are the conditional maximum likelihood estimators of θ , given the signal strength parameter μ and *data*. The pair of parameter estimators $\hat{\mu}$ and $\hat{\theta}$ maximize the likelihood. The constraint $0 \leq \hat{\mu}$ exclude unphysical negative signals, while the upper limit $\hat{\mu} \leq \mu$ is imposed by hand to ensures a one-sided confidence interval.

At this point, it is possible to calculate the observed value of the test statistic $\tilde{q}_{\mu}^{obs}$ as a function of μ .

Maximizing the likelihood function in Eq. C.6 allows the calculation of the value of the nuisance parameters $\hat{\theta}_0^{obs}$ and $\hat{\theta}_{\mu}^{obs}$ that best describe the observed data for the background-only and signal+background hypotheses respectively.

Since it is not possible to repeat the experiment as its required by the frequentist approach, the probability density functions are estimated via the toy MC generation $f(\tilde{q}_{\mu}|\mu, \hat{\theta}_{\mu}, \mu^{obs})$ and $f(\tilde{q}_{\mu}|0, \hat{\theta}_0^{obs})$. In order to satisfy the arguments of the frequentist approach, large enough number of toy-samples can be generated via the reinterpretation of the systematic uncertainties *pdfs* as the posteriors $p(\tilde{\theta}|\theta)$.

With $f(\tilde{q}_\mu|\mu, \hat{\theta}_\mu^{obs})$ and $f(\tilde{q}_\mu|0, \hat{\theta}_0^{obs})$ obtained, two p-values is for H_1, p_μ ; H_0, p_b :

$$p_\mu = P(\tilde{q}_\mu \geq \tilde{q}_\mu^{obs} | \text{signal+background}) = \int_{\tilde{q}_\mu^{obs}}^{\infty} f(\tilde{q}_\mu|\mu, \hat{\theta}_\mu^{obs}) d\tilde{q}_\mu, \quad (\text{C.8a})$$

$$1 - p_b = P(\tilde{q}_\mu \geq \tilde{q}_\mu^{obs} | \text{background}) = \int_{\tilde{q}_\mu^{obs}}^{\infty} f(\tilde{q}_\mu|0, \hat{\theta}_0^{obs}) d\tilde{q}_\mu. \quad (\text{C.8b})$$

The $\text{CL}_s(\mu)$ ratio is given as:

$$\text{CL}_s(\mu) = \frac{p_\mu}{1 - p_b} \quad (\text{C.9})$$

Even though this is a dimensionless quantity, it gives an unbiased confidence level on the H_0 . if $\text{CL}_s \leq \alpha$ for $\mu = 1$, the signal is *excluded* with a $(1 - \alpha)$ CL_s confidence level. In the present analysis, following the convention commonly used in HEP, a 95% confidence level (CL) is required to consider a signal model to be excluded, hence $\text{CL}_s = 0.05$. The upper limit at 95% CL on μ , is quoted as $\mu^{95\% \text{CL}}$, and hence defined as the value of μ for which $\text{CL}_s = 0.05$.

Bibliography

- [1] D. N. Spergel *et al.*, “First-year wilkinson microwave anisotropy probe (wmap) observations: Determination of cosmological parameters,” *The Astrophysical Journal Supplement Series* **148** no. 1, (2003) 175. <http://stacks.iop.org/0067-0049/148/i=1/a=175>.
- [2] N. J. Cornish, D. N. Spergel, G. D. Starkman, and E. Komatsu, “Constraining the topology of the universe,” *Phys. Rev. Lett.* **92** (May, 2004) 201302.
<http://link.aps.org/doi/10.1103/PhysRevLett.92.201302>.
- [3] I. Bars, L. Krauss, F. Nekoogar, and J. Terning, *Extra Dimensions in Space and Time*. Multiversal Journeys. Springer New York, 2009.
<https://books.google.de/books?id=fFSMatekilIC>.
- [4] F. Halzen and A. D. Martin, *Quarks and Leptons: an introductory course in modern Particle Physics*. John Wiley & Sons, 1984.
- [5] D. Perkins, *Introduction to high energy physics*. Cambridge University Press, 1982.
- [6] **Particle Data Group** Collaboration, J. Beringer *et al.*, “Review of Particle Physics (RPP),” *Phys.Rev.* **D86** (2012) 010001.
- [7] H. Baer and X. Tata, *Weak scale supersymmetry: From superfields to scattering events*. Cambridge University Press, 2006.
- [8] S. P. Martin, “A Supersymmetry primer,” *Adv.Ser.Direct.High Energy Phys.* **21** (2010) 1–153, [arXiv:hep-ph/9709356](https://arxiv.org/abs/hep-ph/9709356) [hep-ph].
- [9] **ATLAS** Collaboration, G. Aad *et al.*, “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC,” *Phys.Lett.* **B716** (2012) 1–29, [arXiv:1207.7214](https://arxiv.org/abs/1207.7214) [hep-ex].
- [10] **CMS** Collaboration, S. Chatrchyan *et al.*, “Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC,” *Phys.Lett.* **B716** (2012) 30–61, [arXiv:1207.7235](https://arxiv.org/abs/1207.7235) [hep-ex].
- [11] H. D. Politzer, “Reliable perturbative results for strong interactions?,” *Phys. Rev. Lett.* **30** (Jun, 1973) 1346–1349. <http://link.aps.org/doi/10.1103/PhysRevLett.30.1346>.
- [12] D. J. Gross and F. Wilczek, “Ultraviolet behavior of non-abelian gauge theories,” *Phys. Rev. Lett.* **30** (Jun, 1973) 1343–1346. <http://link.aps.org/doi/10.1103/PhysRevLett.30.1343>.

- [13] “Higgs potential – Goldstone mode – Higgs mode.”
<https://francisthemulenews.files.wordpress.com/2013/07/dibujo20130704-higgs-potential-goldstone-mode-higgs-mode.jpg?w=824>.
 Accessed: 21.03.2016.
- [14] F. Zwicky, “Die Rotverschiebung von extragalaktischen Nebeln,” *Helvetica Physica Acta* **6** (1933) 110–127.
- [15] N. Jarosik, C. Bennett, J. Dunkley, B. Gold, M. Greason, *et al.*, “Seven-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Sky Maps, Systematic Errors, and Basic Results,” *Astrophys.J.Suppl.* **192** (2011) 14, arXiv:1001.4744 [astro-ph.CO].
- [16] **Planck** Collaboration, P. Ade *et al.*, “Planck 2013 results. XVI. Cosmological parameters,” *Astron.Astrophys.* (2014) , arXiv:1303.5076 [astro-ph.CO].
- [17] **Supernova Search Team** Collaboration, A. G. Riess *et al.*, “Observational evidence from supernovae for an accelerating universe and a cosmological constant,” *Astron.J.* **116** (1998) 1009–1038, arXiv:astro-ph/9805201 [astro-ph].
- [18] G. Paal, I. Horvath, and B. Lukacs, “Inflation and compactification from galaxy redshifts?,”.
- [19] **Super-Kamiokande** Collaboration, “Evidence for oscillation of atmospheric neutrinos,” *Phys. Rev. Lett.* **81** (Aug, 1998) 1562–1567.
<http://link.aps.org/doi/10.1103/PhysRevLett.81.1562>.
- [20] R. Davis, D. S. Harmer, and K. C. Hoffman, “Search for neutrinos from the sun,” *Phys. Rev. Lett.* **20** (May, 1968) 1205–1209. <http://link.aps.org/doi/10.1103/PhysRevLett.20.1205>.
- [21] **DAYA-BAY** Collaboration, F. An *et al.*, “Observation of electron-antineutrino disappearance at Daya Bay,” *Phys.Rev.Lett.* **108** (2012) 171803, arXiv:1203.1669 [hep-ex].
- [22] **OPERA** Collaboration, “Observation of a first candidate event in the OPERA experiment in the CNGS beam,” *Physics Letters B* **691** no. 3, (2010) 138 – 145.
<http://www.sciencedirect.com/science/article/pii/S0370269310007537>.
- [23] **CMS** Collaboration, “Search for direct top squark pair production in the single lepton final state at $\sqrt{s} = 13$ TeV,” 2016.
- [24] L. Edelhäuser, J. Heisig, M. Krämer, L. Oymanns, and J. Sonneveld, “Constraining supersymmetry at the lhc with simplified models for squark production,” *Journal of High Energy Physics* **2014** no. 12, (2014) 1–14. [http://dx.doi.org/10.1007/JHEP12\(2014\)022](http://dx.doi.org/10.1007/JHEP12(2014)022).
- [25] **CMS** Collaboration, “CMS Supersymmetry Physics Results.”
<https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsSUS>.
 Accessed: 10.02.2016.
- [26] O. Buchmueller, R. Cavanaugh, A. De Roeck, J. Ellis, H. Flacher, *et al.*, “Likelihood Functions for Supersymmetric Observables in Frequentist Analyses of the CMSSM and NUHM1,” *Eur.Phys.J.* **C64** (2009) 391–415, arXiv:0907.5568 [hep-ph].
- [27] D. Green, ed., *At the leading edge: The ATLAS and CMS LHC experiments*. 2010.
<http://www.worldscibooks.com/physics/7349.html>.
- [28] L. Evans and P. Bryant, “LHC Machine,” *Journal of Instrumentation* **3** (2008) S08001.

-
- [29] “The CERN accelerator complex.” <https://cds.cern.ch/record/1621583/files/CERN's-accelerator-complex2013.jpg>. Accessed: 16.12.2015.
- [30] “The 15-m long LHC cryodipole.” <http://lh-machine-outreach.web.cern.ch/lhc-machine-outreach/images/cryodipole.jpg>. Accessed: 16.12.2015.
- [31] **CMS** Collaboration, S. Chatrchyan *et al.*, “The CMS experiment at the CERN LHC,” *Journal of Instrumentation* **3** (2008) S08004.
- [32] **ATLAS** Collaboration, G. Aad *et al.*, “The ATLAS Experiment at the CERN Large Hadron Collider,” *Journal of Instrumentation* **3** (2008) S08003.
- [33] **ALICE** Collaboration, K. Aamodt *et al.*, “The ALICE experiment at the CERN LHC,” *Journal of Instrumentation* **3** (2008) S08002.
- [34] **LHCb** Collaboration, J. Alves, A. Augusto *et al.*, “The LHCb Detector at the LHC,” *Journal of Instrumentation* **3** (2008) S08005.
- [35] **CMS** Collaboration, S. Bayatian *et al.*, *CMS Physics: Technical Design Report Volume 1: Detector Performance and Software*. Technical Design Report CMS. CERN, Geneva, 2006.
- [36] **CMS** Collaboration, “CMS technical design report, volume II: Physics performance,” *J. Phys. G* **34** (2007) 995–1579.
- [37] **CMS** Collaboration, “The CMS quadrant figure.” <https://twiki.cern.ch/twiki/bin/view/CMS/CSCDPGFigures#CMSQuadrantFigure>. Accessed: 16.12.2015.
- [38] **CMS** Collaboration, *CMS, tracker technical design report*. 1998.
<http://weblib.cern.ch/abstract?CERN-LHCC-98-6>. CERN-LHCC-98-06, CMS-TDR-5.
- [39] **CMS** Collaboration, “CMS: The electromagnetic calorimeter. Technical design report,”. CERN-LHCC-97-33, CMS-TDR-4.
- [40] **CMS** Collaboration, “CMS: The hadron calorimeter technical design report,”. CERN-LHCC-97-31.
- [41] **CMS** Collaboration, “CMS, the Compact Muon Solenoid. Muon technical design report,”. CERN-LHCC-97-32.
- [42] **CMS** Collaboration, “CMS Luminosity Based on Pixel Cluster Counting - Summer 2013 Update,”.
- [43] T. Sjostrand, S. Mrenna, and P. Z. Skands, “PYTHIA 6.4 Physics and Manual,” *JHEP* **0605** (2006) 026, [arXiv:hep-ph/0603175](https://arxiv.org/abs/hep-ph/0603175) [hep-ph].
- [44] B. Andersson, G. Gustafson, G. Ingelman, and T. Sjostrand, “Parton Fragmentation and String Dynamics,” *Phys.Rept.* **97** (1983) 31–145.
- [45] R. Field, “Early LHC Underlying Event Data - Findings and Surprises,” [arXiv:1010.3558](https://arxiv.org/abs/1010.3558) [hep-ph].
- [46] **CMS** Collaboration, S. Chatrchyan *et al.*, “Measurement of the Underlying Event Activity at the LHC with $\sqrt{s} = 7$ TeV and Comparison with $\sqrt{s} = 0.9$ TeV,” *JHEP* **1109** (2011) 109, [arXiv:1107.0330](https://arxiv.org/abs/1107.0330) [hep-ex].
- [47] **CMS** Collaboration, S. Chatrchyan *et al.*, “Jet and underlying event properties as a function of charged-particle multiplicity in proton–proton collisions at $\sqrt{s} = 7$ TeV,” *Eur.Phys.J.* **C73** (2013) 2674, [arXiv:1310.4554](https://arxiv.org/abs/1310.4554) [hep-ex].

- [48] F. Maltoni and T. Stelzer, “MadEvent: Automatic event generation with MadGraph,” *JHEP* **0302** (2003) 027, [arXiv:hep-ph/0208156](#) [hep-ph].
- [49] P. Nason, “A New method for combining NLO QCD with shower Monte Carlo algorithms,” *JHEP* **0411** (2004) 040, [arXiv:hep-ph/0409146](#) [hep-ph].
- [50] S. Frixione, P. Nason, and C. Oleari, “Matching NLO QCD computations with Parton Shower simulations: the POWHEG method,” *JHEP* **0711** (2007) 070, [arXiv:0709.2092](#) [hep-ph].
- [51] S. Alioli, P. Nason, C. Oleari, and E. Re, “NLO vector-boson production matched with shower in POWHEG,” *JHEP* **0807** (2008) 060, [arXiv:0805.4802](#) [hep-ph].
- [52] Z. Was, “TAUOLA for simulation of tau decay and production: perspectives for precision low energy and LHC applications,” *Nucl.Phys.Proc.Suppl.* **218** (2011) 249–255, [arXiv:1101.1652](#) [hep-ph].
- [53] CMS Collaboration, “Comparison of the Fast Simulation of CMS with the first LHC data,” 2010.
- [54] S. Ovin, X. Rouby, and V. Lemaitre, “DELPHES, a framework for fast simulation of a generic collider experiment,” [arXiv:0903.2225](#) [hep-ph].
- [55] “Ansi/vita, ansi/vita 1-1994, vme64 standard.”
www.vita.com/home/Specification/Specifications.html. Accessed: 30.10.2015.
- [56] CMS Collaboration, J. Mans *et al.*, “CMS Technical Design Report for the Phase 1 Upgrade of the Hadron Calorimeter,”.
- [57] “Micro telecommunications computing architecture base specification r1.0.”
<https://www.picmg.org/openstandards/microtca/>. Accessed: 30.10.2015.
- [58] E. Hazen *et al.*, “The AMC13XG: a new generation clock/timing/DAQ module for CMS MicroTCA,” *Journal of Instrumentation* **8** (2013) C12036.
- [59] E. Hazen *et al.*, “The FC7 AMC for generic DAQ and control applications in CMS,” *Journal of Instrumentation* **10** (2015) C03036.
- [60] M. B. Marin *et al.*, “The GBT-FPGA core: features and challenges,” *Journal of Instrumentation* **10** (2015) C03021.
- [61] “Kintex fpga overview and specifications.”
<http://www.xilinx.com/products/silicon-devices/fpga/kintex-7.html>.
Accessed: 30.10.2015.
- [62] C. G. Larrea, K. Harder, D. Newbold, D. Sankey, A. Rose, A. Thea, and T. Williams, “Ipbu: a flexible ethernet-based control system for xtca hardware,” *Journal of Instrumentation* **10** no. 02, (2015) C02019. <http://stacks.iop.org/1748-0221/10/i=02/a=C02019>.
- [63] S. Goadhouse, “The CMS HCAL FEE Control Module.” Accessed: 09.11.2015.
- [64] A. Baumbaugh *et al.*, “QIE10: a new front-end custom integrated circuit for high-rate experiments,” *Journal of Instrumentation* **9** (2014) C01062.
- [65] “Proasic3l fpgas.” <http://www.microsemi.com/products/fpga-soc/fpga/proasic3l>.
Accessed: 08.12.2015.

-
- [66] L. Amaral, S. Dris, A. Gerardin, T. Huffman, C. Issever, A. J. Pacheco, M. Jones, S. Kwan, S. C. Lee, Z. Liang, T. Liu, Z. Meng, A. Prosser, S. Padadopoulos, I. Papakonstantinou, C. Sigaud, S. Silva, C. Soos, P. Stejskal, J. Troska, F. Vasey, P. Vichoudis, T. Weidberg, A. Xiang, and J. Ye, “The versatile link, a common project for super-lhc,” *Journal of Instrumentation* **4** no. 12, (2009) P12003. <http://stacks.iop.org/1748-0221/4/i=12/a=P12003>.
- [67] “Igloo2 fpgas.” <http://www.microsemi.com/products/fpga-soc/fpga/igloo2-fpga>. Accessed: 08.12.2015.
- [68] CMS Collaboration, “Commissioning of the Particle-flow Event Reconstruction with the first LHC collisions recorded in the CMS detector,” 2010.
- [69] CMS Collaboration, “Tracking and Primary Vertex Results in First 7 TeV Collisions,” 2010.
- [70] CMS Collaboration, “CMS Luminosity - Public Results.”
- [71] R. Fruhwirth, W. Waltenberger, and P. Vanlaer, “Adaptive vertex fitting,” *J.Phys.* **G34** (2007) N343.
- [72] CMS Collaboration, S. Chatrchyan *et al.*, “Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$ TeV,” *Journal of Instrumentation* **7** (2012) P10002, [arXiv:1206.4071](https://arxiv.org/abs/1206.4071) [physics.ins-det].
- [73] W. Adam, R. Fruhwirth, A. Strandlie, and T. Todorov, “Reconstruction of electrons with the Gaussian sum filter in the CMS tracker at LHC,” *eConf C0303241* (2003) TULT009, [arXiv:physics/0306087](https://arxiv.org/abs/physics/0306087) [physics].
- [74] CMS Collaboration, S. Chatrchyan *et al.*, “Performance of tau-lepton reconstruction and identification in CMS,” *Journal of Instrumentation* **7** (2012) P01001, [arXiv:1109.6034](https://arxiv.org/abs/1109.6034) [physics.ins-det].
- [75] Salam, Gavin P., “Towards jetography,” *Eur. Phys. J. C* **67** no. 3, (2010) . <http://dx.doi.org/10.1140/epjc/s10052-010-1314-6>.
- [76] CMS Collaboration, “Jet Performance in pp Collisions at 7 TeV,” 2010.
- [77] CMS Collaboration, S. Chatrchyan *et al.*, “Determination of Jet Energy Calibration and Transverse Momentum Resolution in CMS,” *Journal of Instrumentation* **6** (2011) P11002, [arXiv:1107.4277](https://arxiv.org/abs/1107.4277) [physics.ins-det].
- [78] CMS Collaboration, S. Chatrchyan *et al.*, “Identification of b-quark jets with the CMS experiment,” *Journal of Instrumentation* **8** (2013) P04013, [arXiv:1211.4462](https://arxiv.org/abs/1211.4462) [hep-ex].
- [79] CMS Collaboration, S. Chatrchyan *et al.*, “Search for top-squark pair production in the single-lepton final state in pp collisions at $\sqrt{s} = 8$ TeV,” *Eur.Phys.J.* **C73** (2013) 2677, [arXiv:1308.1586](https://arxiv.org/abs/1308.1586) [hep-ex].
- [80] CMS Collaboration, “Search for Direct Top Squark Pair Production in the Single Lepton Channel with Transverse Mass at 8 TeV,” 2013. Unpublished. Available at http://cms.cern.ch/iCMS/jsp/openfile.jsp?tp=draft&files=AN2013_089_v9.pdf.
- [81] CMS Collaboration, S. Chatrchyan *et al.*, “Measurement of differential top-quark pair production cross sections in pp collisions at $\sqrt{s} = 7$ TeV,” *Eur.Phys.J.* **C73** (2013) 2339, [arXiv:1211.2220](https://arxiv.org/abs/1211.2220) [hep-ex].

- [82] **CMS** Collaboration, “Measurement of differential top-quark pair production cross sections in the lepton+jets channel in pp collisions at 8 TeV,” 2013.
- [83] **CMS** Collaboration, “Measurement of the differential top-quark pair production cross section in the dilepton channel in pp collisions at $\sqrt{s} = 8$ TeV,” 2013.
- [84] N. Kidonakis, “NNLL threshold resummation for top-pair and single-top production,” *Phys.Part.Nucl.* **45** (2014) 714, [arXiv:1210.7813 \[hep-ph\]](#).
- [85] **CMS** Collaboration, “Top p_T reweighting twiki.”
- [86] **CMS** Collaboration, “MET Optional Filters.”
- [87] M. Cacciari, G. P. Salam, and G. Soyez, “FastJet User Manual,” *Eur. Phys. J. C* **72** (2012) 1896, [arXiv:1111.6097 \[hep-ph\]](#).
- [88] **UA1** Collaboration, G. Arnison *et al.*, “Experimental Observation of Isolated Large Transverse Energy Electrons with Associated Missing Energy at $\sqrt{s} = 540$ -GeV,” *Phys.Lett. B* **127** (1983) 103–116.
- [89] Y. Bai, H.-C. Cheng, J. Gallicchio, and J. Gu, “Stop the Top Background of the Stop Search,” *JHEP* **1207** (2012) 110, [arXiv:1203.4813 \[hep-ph\]](#).
- [90] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, and H. Voss, “TMVA: Toolkit for Multivariate Data Analysis,” *PoS ACAT* (2007) 040, [arXiv:physics/0703039](#).
- [91] T. M. Mitchell, *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 ed., 1997.
- [92] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals of Eugenics* **7** no. 2, (1936) 179–188. <http://dx.doi.org/10.1111/j.1469-1809.1936.tb02137.x>.
- [93] F. Rosenblatt, *Principles of neurodynamics: perceptrons and the theory of brain mechanisms*. Spartan Books, Washington, DC, USA, 1962.
- [94] V. Vapnik and A. Lerner, “Pattern recognition using generalized portrait method,” *Automation and Remote Control* **24** (1963) 774–780.
- [95] V. Vapnik and A. Chervonenkis, “A note on one class of perceptrons,” *Automation and Remote Control* **25** (1963) 103.
- [96] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning* **20** no. 3, (1995) 273–297. <http://dx.doi.org/10.1023/A%3A1022627411411>.
- [97] V. Vapnik, *Estimation of Dependences Based on Empirical Data*. Springer Verlag, 1982.
- [98] W. Karush, *Minima of Functions of Several Variables with Inequalities as Side Constraints*. PhD thesis, University of Chicago, 1939.
- [99] H. W. Kuhn and A. W. Tucker, “Nonlinear programming,” in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, pp. 481–492. University of California Press, Berkeley and Los Angeles, 1951.
- [100] J. Mercer, “Functions of positive and negative type, and their connection with the theory of integral equations,” *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **209** no. 441-458, (1909) 415–446.

-
- [101] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods - Support Vector Learning*. MIT Press, January, 1998.
<http://research.microsoft.com/apps/pubs/default.aspx?id=68391>.
- [102] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology* **2** (2011) 27:1–27:27. Software available at
<http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [103] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in large margin classifiers*, pp. 61–74. 1999.
- [104] T. F. Wu, C. J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *J. Mach. Learn. Res.* **5** (2004) 975–1005.
- [105] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics," *European Physical Journal C* **71** (Feb., 2011) 1554, [arXiv:1007.1727](https://arxiv.org/abs/1007.1727) [physics.data-an].
- [106] 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," *Nuclear Instruments and Methods in Physics Research A* **595** (Oct., 2008) 480–501, [physics/0702156](https://arxiv.org/abs/physics/0702156).
- [107] M. O. Sahin, D. Kruecker, and I. A. Melzer-Pellmann, "SVM-HEP Interface," 2015.
<https://www.github.com/ml-hint/svm-hint>.
- [108] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.
- [109] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences* **55** no. 1, (1997) 119 – 139.
- [110] F. Costanza, *Search for Pair Production of Supersymmetric Top-Quark Partners in Events with a Single Lepton at CMS*. PhD thesis, University of Hamburg, 2014.
- [111] CMS Collaboration, "Methods to apply b-tagging efficiency scale factors."
- [112] CMS Collaboration, "Jet Energy Scale - Jet Energy Resolution and Corrections."
- [113] ATLAS Collaboration, "ATLAS Supersymmetry Physics Results." <https://twiki.cern.ch/twiki/bin/view/AtlasPublic/SupersymmetryPublicResults>. Accessed: 10.02.2016.
- [114] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics," *Eur. Phys. J.* **C71** (2011) 1554, [arXiv:1007.1727](https://arxiv.org/abs/1007.1727) [physics.data-an].
[Erratum: *Eur. Phys. J.* **C73**,2501(2013)].

Acknowledgments

During my PhD studies I had a chance to work with excellent and inspiring people. Even though it is not possible to present my appreciation to all of them, I would like to mention the most representative (Asimov) set in the following few sentences.

I am more than grateful to my supervisor Dr Isabell Melzer-Pellman. She provided me an unparalleled guidance in my research and career. Under her supervision, I had the inspiration and freedom to exceed my limits. I would like to also thank my co-supervisor Prof Dr Peter Schleper, for our motivational meetings, and his invaluable wisdom and advises.

I feel myself privileged to work in the DESY-CMS group, particularly the SUSY and mTCA workgroups. I am grateful to Altan for his support and friendship, and to Dirk and Alan for their guidance and the excellent collaboration. I would like to thank to our staff members and postdocs: Dean, Elias, Ulf, Ingo, Claudia and Alexis.

It is not possible to mention all the DESY-CMS group, therefore, I would like to thank Prof Dr Kerstin Borras and Dr Matthias Kasemann as the team leaders. A special mention goes to Birgit and Gabriele for their support and understanding.

I really appreciate the support from Joachim Herz Stiftung and PIER. Especially Dr Eva Ackermann, Dr Jörg Maxton-Küchenmeister and Stefanie Tepass provided me the excellent environment for the research and study. I would like to also thank to other JHS fellows: Clemens, Matthias and Nele: organizing READi 2014 with them was a tremendous joy. I would like to express my gratitude to Petra and late Joachim Herz for their vision and support for science.

I have worked and collaborated with excellent scientists and engineers at CERN. I would like to thank Sezen, Buğra, Efe, Stephen, Tullio and Dick for the knowledge and inspiration they shared with me.

I was really lucky to be officemates with excellent people including Francesco, Hannes, Niklas, Artur, Karim, Tuğba and Aksansh. I really appreciate their friendship, and their patience (particularly on my keyboard noise and questions)!

During my studies, there were few people who made the challenging times more bearable for me. I am grateful to Artur for being an excellent festival partner and his understanding in every aspect! I would like to thank Oscar for his career advises, after-work dinners, and amazing music he shared with me!

I am grateful to Francesco for being my office-room-coffee-lab-house-cooking-mate. He inspired me with his knowledge and intelligence in every aspect of life.

I am sincerely thankful to Tuğba for her support, patience, and being there for me when I needed it most.

...and to my family: Her kararımda yanımda ve her anımda bana destek olduğunuz için çok teşekkür ederim. Sizleri çok seviyorum.

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

I hereby declare, on oath, that I have written the present dissertation by my own and have not used other than the acknowledged resources and aids.

Hamburg, 10.05.2016