Master Thesis

Studies of the tW production process in p–p collisions at $\sqrt{s} = 13$ TeV at the CMS detector

---

Author
Víctor Rodríguez Bouza

Advisors
Sergio Sánchez Cruz
Francisco Javier Cuevas Maestro

Academic year 2017-2018

Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

UAB
Universitat Autònoma de Barcelona

Abstract

We present a method to measure for the first time in the CMS detector the differential cross section of the tW production process taking $36.9\,\mathrm{fb}^{-1}$ of data from 2016 observed at that experiment at LHC. This procedure enhances the signal extraction using MVA techniques such as BDT and maximum likelihood fits. The measurement is performed in a region has one jet that is b-tagged, one muon and one electron, and the variables chosen are the pseudorapidity of the system of both leptons and the jet, the p_T of the lepton with the highest one, the difference in the φ angle of both leptons, the invariant mass of the jet with the lepton with more p_T and the invariant mass of the jet with the other lepton. Observed distributions are in agreement with the theoretical predictions represented by the `Powheg` and `aMC@NLO` models, with further sensitivity required to give preference to one of the two. The final results are thus promising, though further study and work is expected in order to enhance the sensibility of the analysis.

Contents

Introduction

The main aim of this master thesis is to study the production of a top quark in association with a W boson in proton – proton collisions at the LHC accelerator using the data collected during 2016 by the CMS detector. Specifically, our target will be the measurement of the differential cross section of this process, depending on various variables. This goal implies a challenge as, even in the selected region, where the ratio of the presence of the tW process against its backgrounds is the highest, we face a large amount of events coming essentially of the production of a pair of top quarks (the $t\bar{t}$ process), and thus is complicate to have sensitivity to the signal process.

This work is a natural continuation of the inclusive cross section measurement of the same physical process done by the CMS Collaboration ([31]), that gave place to the beginnings of the analysis described in this document. The efforts that have been made for the inclusive measurement serve as basis for the differential cross section measurement: the analysis that is described in this master thesis. It has already been presented in the CMS single top group and its current developments are documented in an analysis note ([38]), before it continues to an official CMS publication.

The structure of this document is the following: in the first chapter (preceded by the abstract and this introduction) the theoretical and experimental resources that are needed to understand the procedures and the results of the analysis are presented. It is divided in three sections, encompassing the physical theoretical framework, a glimpse of the top physics inside the experimental high energy physics context, and the description of the experiment that provides us the data of the analysis.

Then, in the second chapter the full methodology of the analysis is explained in detail. First, the general mathematical and computational considerations are contemplated and afterwards the details of how each tool is implemented in our work. This is followed by the third chapter, where all the experimental results are presented.

Finally, the last chapter contains the conclusions of all the thesis, followed by the references which are used all over the document.

1 Theoretical and experimental prolegomena

Since the times of Thales of Miletus, considered the first contributor to Western philosophy[1] back in the 7th and 6th centuries B.C., humanity has always wondered about the nature of matter using the ways of logic. Inside the group that Thales founded, the pre-Socratics, atomists considered that the essence of all the entities that had the property of "being" was at the end in some indivisible elementary particles called atoms[2], which carried the characteristics of "being" defined by Parmenides.

Once the foundations for what we understand nowadays as science were settled during the Scientific Revolution, new discoveries between the Renaissance and the 20th century were made regarding the composition of matter. These achievements, such as the knowledge of the electron, or the quantum understanding of matter, ended up in the second half of the last century giving us what we know today as the standard model of particle physics, or simply, the SM. This enterprise had success also because of the experimental efforts in large particle accelerators and colliders.

In this chapter a very brief description of the SM is shown, detailing specifically the top quark physics in this theoretical framework. Subsequently, the experimental context where the data have been acquired is shortly explained.

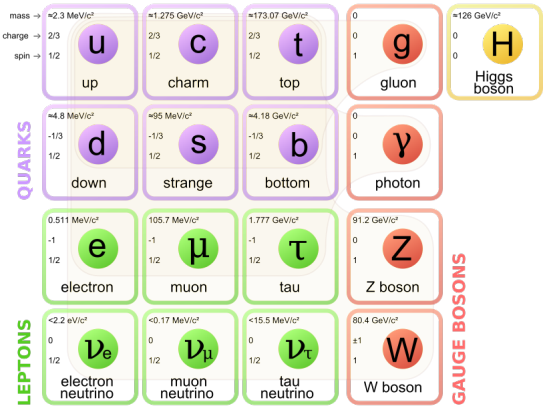


Figure 1.1: [67] Elementary particles that describe the standard model. The picture shows in violet the quarks, in green the leptons and in red the vectorial bosons. In yellow, the Higgs boson is displayed.

1.1 Theoretical framework: the standard model

The **standard model of particle physics** (more commonly, as we said, **SM**) is a scientific model based on quantum field theory (or QFT) that describes all the elementary constituents of matter discovered as well as the electromagnetic and nuclear weak and strong interactions between them. Its predictions have been thoroughly and continuously put to the test, and in each iteration with higher precision.

Within this framework **fermions** (states of spin $\frac{1}{2}$) and **bosons** (states of integer spin) conform all the possible particles, as we can see in Fig. 1.1. Any interaction between fermions is required to be mediated by the different bosons according to the processes that are derived from the Lagrangian of the SM. The different couplings or interactions between the fermions and the bosons are summarized in Fig. 1.2.

[1] And also the predecessor of the current science, though we cannot recognize such until the Scientific Revolution at the Renaissance. In a nutshell, before the works of Galileo one could say that (what we understand nowadays as) science was "mixed" with (what we understand nowadays as) philosophy.

[2] From the Greek of *ἄτομον*: indivisible, uncut, without parts.

As we said, the SM has been able to successfully make predictions since the last part of the 20th century, such as the last one: the Higgs boson. However, there are physical issues for which the SM does not offer an answer, or simply aspects of the Universe that are not taken into account in it. These are, for example, gravitation (which is not part of the SM), neutrino masses, dark matter, dark energy, the strong CP problem, the explanation of why there are only three generations, the origin of the SM parameters that must be obtained experimentally or the hierarchy problem.

Because of these topics, the scientific community has been proposing different enhancements or ideas to cope with them. These, and other propositions are usually grouped under the label "beyond standard model", or simply BSM. In this set of ideas one can find proposals like supersymmetry or SUSY, which offers a solution, for example, for the hierarchy problem. We also have suggestions as the grand unified theories or GUT whose main objective is to unify the interactions that operate in the SM in one only force. To do so, SM (which has a gauge group structure of $SU(2)_L \times SU(3)_C \times U(1)_Y$) would be embedded in a higher gauge group (for example, $SU(5)$), yielding at the end only one coupling. There are other proposals that aim to explain one or some of the issues that we listed before, such as the composite Higgs theories, axions or string theory, which hopes to explain in a single theory gravitation as well as the other three forces, and that can also support SUSY within it.

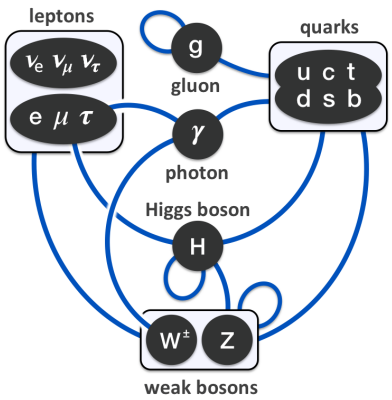


Figure 1.2: [66] Scheme of the SM particles and the couplings between them.

Main physical observables

As our experimental data for the analysis will be extracted from particle interactions (collisions), it is important for us to understand the observables that must be extracted from the physical process of interest. One, if not the most important of those observables, is the **cross section**. This scalar magnitude gives us an estimation of how probable is that an interaction occurs, given the context of it (i.e. the energy of the centre-of-mass frame, usually). It is measured usually in barns (b)[3], which is a unit of surface. Another interpretation of the cross section that might be given is that of the effective area which the particles that could interact must cross in order to do so. The most important reason to measure cross sections is that they are a link between measurements of events in large detectors and the fundamental physics that we aim to study. This is because they depend explicitly on the matrix element (ME) of the scattering matrix (or S-matrix) of the process, that depends itself on the interaction Hamiltonian of the quantum field theory that we want to put to the test.

If one desires to study the production rate of a process depending on some variable, one can put the cross section as a function of that variable (or variables) of the physical process (for example, if a muon participates in the interaction, its momentum could be one). In that case, we speak of the **differential cross section**, $d\sigma$, opposite to the **total cross section** obtained when integrating over all dependencies. Constraining the discussion to scattering processes in particle physics, the total cross section is defined as the number of interactions per unit of time and per unit of flux, that is

[3] $1\,\mathrm{b} = 10^{-28}\,\mathrm{m}^2$.

$$\sigma := \frac{\frac{dN_{\text{int.}}}{dt}}{\mathcal{L}} = \frac{\frac{dN_{\text{int.}}}{dt}}{\frac{dN_{\text{inc.}}}{dt \cdot dA}}, \tag{1.1}$$

where $N_{\text{int.}}$ represent the number of interactions, $N_{\text{inc.}}$ the number of incoming particles and A the area (with t being the time). The flux $\mathcal{L}$ is often defined as **instantaneous luminosity** in collider physics contexts, and it is commonly measured in c.g.s. units: $cm^{-2} \cdot s^{-1}$. A similar value is obtained through the time integration of this quantity, obtaining the **integrated luminosity**, L.

$$L := \int_{t_0}^{t_1} \mathcal{L}\, dt \tag{1.2}$$

This value gives us an estimation of how many interactions have taken place in a given time, thus for the scientific community L is commonly understood as a measure of the amount of data obtained. Integrating on time again, one can reach from Eq. (1.1) the following useful relation,

$$N_{\text{int.}} = L \cdot \sigma. \tag{1.3}$$

By definition, the instantaneous luminosity depends indirectly on the experimental setup, whatever the experiment is. If some collision data came from one collider, the instantaneous luminosity of that installation would depend on how well or bad the particle beams are collimated, how many particles are going to collide, and other experimental aspects. Eq. (1.3) encapsulates the separation between the underlying physics (represented by σ) and the experimental setup (represented by L) that combined explain the number of interactions ($N_{\text{int.}}$).

The collider from which the data of this analysis comes, the LHC[4] (described in the next section), runs $p - p$ collisions (and also others, but those are of no interest for this analysis). Protons are hadrons, and thus they are not elementary particles. Due to the energies at which the collisions take place, the components of the protons, usually called partons, are the ones that actually interact (in what it is called deep inelastic scattering). The total cross section in a $p - p$ collision can be interpreted depending on the cross sections of the interactions between partons, $\sigma(f_i f_j \to X)$ as

$$\sigma(pp \to X) = \sum_{ij} \int \text{PDF}(x_1, f_i)\text{PDF}(x_2, f_j)\sigma(f_i f_j \to X)\, dx_1 dx_2. \tag{1.4}$$

In the expression, the PDF functions describe the probability of finding the component i of one proton with a fraction x_n of the momentum of that very proton. Sums are needed to tackle the differences in colours and flavours of the partons. These PDF functions are called **parton distribution functions**, and they must be obtained experimentally, measuring them at a particular energy in the centre-of-mass frame ($\sqrt{s}$). This is unfortunate, as obviously not all particle physics experiments run at the same energy. However, this dependence with the energy scale of the measurement can be extracted using the SM formalism (specifically, quantum chromodynamics, QCD: the subpart of SM dedicated to the strong interaction) and thus the PDF

[4]Large Hadron Collider: see next section.

can be extrapolated to different $\sqrt{s}$.[5] The **coupling constant of the strong interaction**, α_S, plays a role in that dependence, being it also a function of the energy of the interaction, because all coupling constants from the three forces depend on it due to quantum corrections and the perturbation theory employed in QFT.

There is one last remark concerning this observable. Commonly, the process of calculating simulations using expressions such as Eq. (1.4) to obtain the cross section in contexts like $p-p$ collisions at the LHC is extremely difficult, due to the complexity itself of hadron collisions, and also because the production of coloured particles yields to very prolific final states. The approach that is used in the community to obtain these simulations is to separate the full physical process in two parts. The main physical process (called the "hard" process), whose information is obtained as previously commented (from the matrix element of the corresponding scattering matrix), and the so-called "soft" process, which aims to be essentially the remaining physical subprocesses (mainly, the posterior evolution of the final state products). The soft process is usually estimated through phenomenological models, being named the most important of all parton showers (PS). This is allowed thanks to the factorisation theorem as explained in the standard model course, but paying the cost of adding an extra parameter to the dependence of the cross section: the **factorization scale**, μ_F. Another factor that ends up affecting our results is the **renormalization scale**, usually written μ_R. The cross section is a priori not dependent on the renormalization scale, but only when all the infinite terms of the perturbation series are considered. As we cannot do that, because QCD is not renormalizable at low energies, we must calculate a limited number of terms, and thus, there will exist a dependence also in the renormalization scale.

The PDF are essential to perform this analysis, and it will be necessary to consider the uncertainties related with them, including the dependence on the values of α_S. In addition, other sources of uncertainties that will need to be taken into account in this analysis come from this factorisation between the hard process, modelled by the ME, and the PS (e.g. the matching between the two contributions, or the renormalisation and factorisation scales), the uncertainties related to the initial and final state radiations, or the different approaches to propagate through the PS the colour charge of the ME final state particles.

1.2 Top physics inside the SM

The top quark (t) is a member, as we saw in Fig. 1.1, of the third generation of fermions, along with the quark bottom (b), the tau (τ) lepton and its neutrino (ν_τ), and it was first observed in 1995 in the Tevatron (Fermilab) thanks to the work of the CDF ([12]) and D0 ([40]) collaborations. Though it is a member of the quarks subset of fermions, it has relevant differences with respect to the rest of them.

The most important one is its very high mass compared with the other quarks: $m_t = (173.1 \pm 0.6)\,\text{GeV}$[6], being the most massive object in the SM.[7] a value that is two orders of magnitude over its companion in the EW doublet, the quark b, $m_b = 4.18^{+0.04}_{-0.03}\,\text{GeV}$ and six over the mass of the lightest quark, the quark up: $m_u = 2.2^{+0.6}_{-0.4}\,\text{MeV}$. The implications of this act firstly on its very existence, as its mean lifetime due to its high mass is very low. Thus, differently than the rest of the quarks that later or sooner undergo the process of hadronisation, the top quark decays before it can hadronise. Actually, in 95.7% ([57]) of each cases, it decays

[5] Actually, the agreement of different measurements of the PDF with the extrapolations of others is considered to be a check of the QCD validity.

[6] In natural units ($\hbar = c = 1$) This convention will be used in the entire text.

[7] [57] is the source of all the masses of the paragraph.

in the same way, into a $W^{\pm}$ boson and a $b/\bar{b}$ quark. This is also reflected in the Cabibbo-Kobayashi-Maskawa (CKM) matrix V_{CKM} elements that account for the interactions of the top with its bottom, strange and down quark partners, being $V_{tb} >> V_{ts}, V_{td}$.

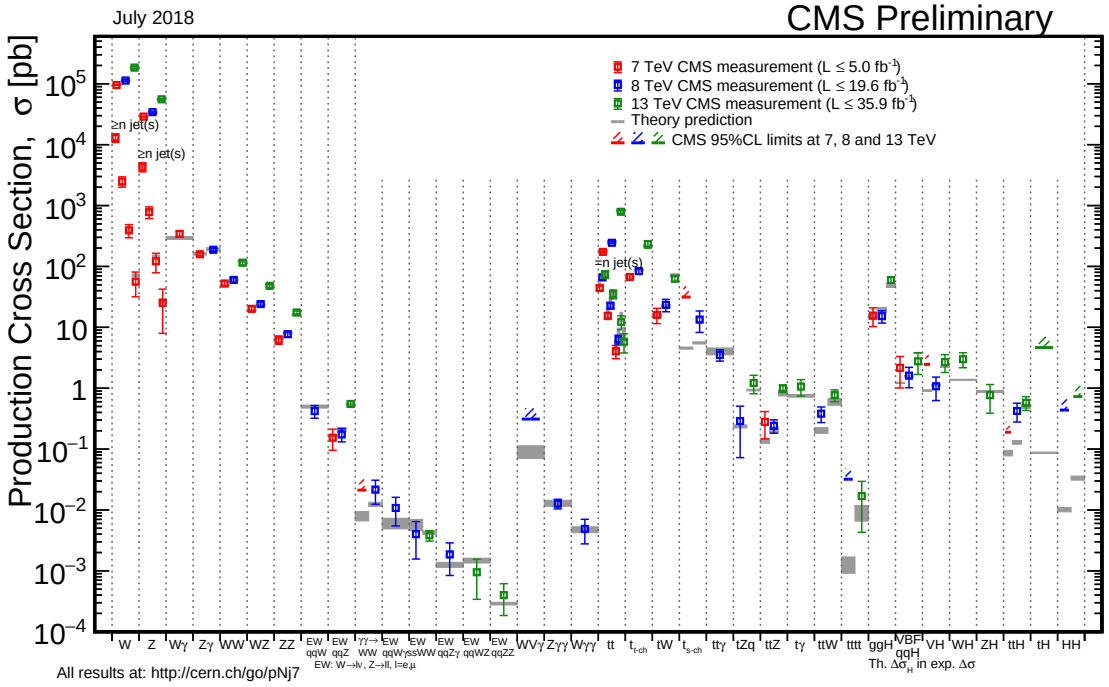


Figure 1.3: [36] Comparison of the cross sections of several processes that undergo in the LHC measured by the CMS Collaboration.

Inside SM, fermion masses are due to the spontaneously symmetry breaking (SSB) that its Lagrangian exhibit. As a consequence, they depend on the vacuum expectation value (v or "v.e.v.") that the Higgs boson acquires with the SSB, and the known as Yukawa couplings (usually represented as λ_f or Y_f, where f represents the fermion): these measure how intense is the interaction between the Higgs boson and the fermions. Then, when looking inside the Higgs potential, one finds terms that can be considered mass terms with coefficients that depend on v and these couplings. For the top case, the Lagrangian terms of interaction between the Higgs boson and the top quark are

$$-\frac{\lambda_t v}{\sqrt{2}} \cdot \bar{t}_L t_R - \frac{\lambda_t}{\sqrt{2}} \cdot H \bar{t}_L t_R + \text{h.c.}, \tag{1.5}$$

where "h.c." refers to the hermitian conjugate of the previous terms, v is the v.e.v. ([57]: $v \simeq 246\,\text{GeV}$), λ_t is the Yukawa coupling of the top quark, H is the Higgs boson field, $\bar{t}_L$ is the left-handed top quark field (member of a weak isospin doublet along the left-handed bottom field), and t_R is the right-handed top quark field (which is a weak isospin singlet). Identifying thus the top quark's mass, we can get the Yukawa coupling of the top. According to the value of the top mass from before, we can get a Yukawa coupling of order one. This makes the top the only quark that features such a coupling with the Higgs and therefore they are considered unique probes to understand the mechanisms of the electroweak SSB. In addition, their contributions to loops appear as the most important ones when considering quantum corrections to Higgs bosons procedures. Furthermore, they are also very useful to search for BSM physics. For example, some enhancements of the SM predict additional W' and Z' bosons that can decay to top quarks. Other BSM approaches, such as composite Higgs theories,

might predict a complex Higgs sector in an expanded standard model with more Higgs bosons that could couple preferentially to the top, due to its high mass.

In Tevatron, the discovery of the top was made with data from proton-antiproton collisions of 980 GeV per beam, thus yielding energies in the centre of mass of the order of $\sqrt{s} \sim 2\,\text{TeV}$. Nowadays, the LHC collider, with proton-proton collisions, is able to provide $\sqrt{s} = 13\,\text{TeV}$ and higher collision rates[8] than the Tevatron ones, getting the nickname of "top factory". This is evidenced with the comparisons of its cross sections for the main way of producing tops in both colliders: the pair production more known as $t\bar{t}$, of the order of $\sim 7\,\text{pb}$ ([12]) for the Tevatron (at the mentioned energies) and nowadays of two orders higher: precisely, of a predicted (at $\sqrt{s} = 13\,\text{TeV}$[9]; [39], [10], [55], [9]) $\sigma_{t\bar{t}} = 831.76^{+19.77}_{-29.20}(\text{scale}) \pm 35.06(\text{PDF},\alpha_S)\,\text{pb}$, and a measurement (in CMS, [30]) of $\sigma_{t\bar{t}} = 815 \pm 9(\text{stat}) \pm 38(\text{syst}) \pm 19(\text{lumi})\,\text{pb}$[10]. The Feynman diagrams associated to this process (either through gluon fusion or quark fusion) can be seen in Fig. 1.4. This is comparatively one of the most produced processes at $\sqrt{s} = 13\,\text{TeV}$ in the LHC , as Fig. 1.3 shows.

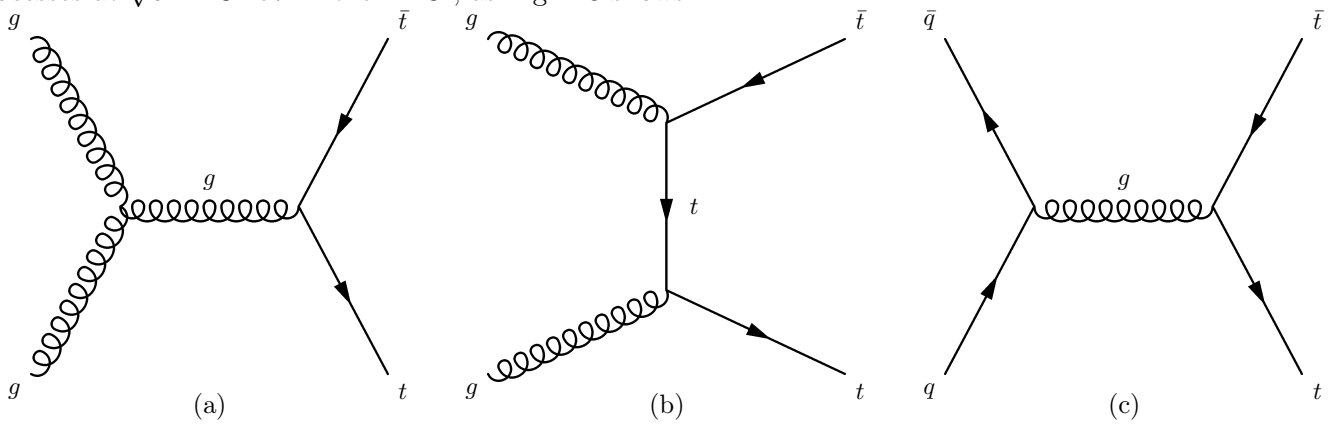


Figure 1.4: Leading order Feynman diagrams of the $t\bar{t}$ production processes for gluon fusion (a, b) and quark fusion (c). The most relevant one is gluon fusion channel, accounting for $\sim 80 - 90\%$ of the total $t\bar{t}$ production cross section in the LHC centre-of-mass energies ([57]).

The $t\bar{t}$ process is very important in a huge number of LHC analysis and also in this one (as it will be shown in next section), but the focus of this master thesis is not put upon it. When instead of a pair of top-antitop quarks only one ($t/\bar{t}$) is obtained, the process is considered to be a "single top" one. In this group, we have the t–channel, the s–channel, and the tW process, which has the particularity of yielding a W boson in addition of the top. In Fig. 1.5 the Feynman diagrams of the three single-top processes are shown.

From these three, the process with the higher predicted[11] cross section at $\sqrt{s} = 13\,\text{TeV}$ of the three is the t–channel, with $\sigma_{\text{t ch.}} = 217.0^{+6.6}_{-4.6}(\text{scale}) \pm 6.2(\text{PDF},\alpha_S)\,\text{pb}$, being the second the tW process $\sigma_{tW} = 71.7 \pm 1.8(\text{scale}) \pm 3.4(\text{PDF},\alpha_S)\,\text{pb}$, whereas the s–channel has the lowest one: $\sigma_{\text{s ch.}} = 10.32^{+0.29}_{-0.24}(\text{scale}) \pm 0.27(\text{PDF},\alpha_S)\,\text{pb}$ ([1], [49], [54], [51], [10], [55] for both). The two first processes have been observed both in CMS and ATLAS, with measurements for the first detector of ([29], [31] resp.) $\sigma_{\text{t ch.}} = 232 \pm 13(\text{stat}) \pm 12(\text{exp}) \pm 26(\text{theo}) \pm 6(\text{lumi})\,\text{pb}$ and $\sigma_{tW} = 63.1 \pm 1.8(\text{stat}) \pm 6.4(\text{syst}) \pm 2.1(\text{lumi})\,\text{pb}$. The s–channel has not been observed yet in the LHC, though yes in the Tevatron ([13]).

[8] Or instantaneous luminosities: see next section.

[9] And also at next-next-to-leading-order precision and assuming $m_t = 172.5\,\text{GeV}$. Uncertainties due to the factorization and renormalization scales, as well as the parton density functions and the running of the strong coupling constant are shown.)

[10] A remarkable fact is that last measurements, as this one shows, have surpassed the current precision of the theoretical predictions.

[11] At next-to-leading-order precision and assuming $m_t = 172.5\,\text{GeV}$. Uncertainties due to the factorization and renormalization scales, as well as the parton density functions and the running of the strong coupling constant are shown.

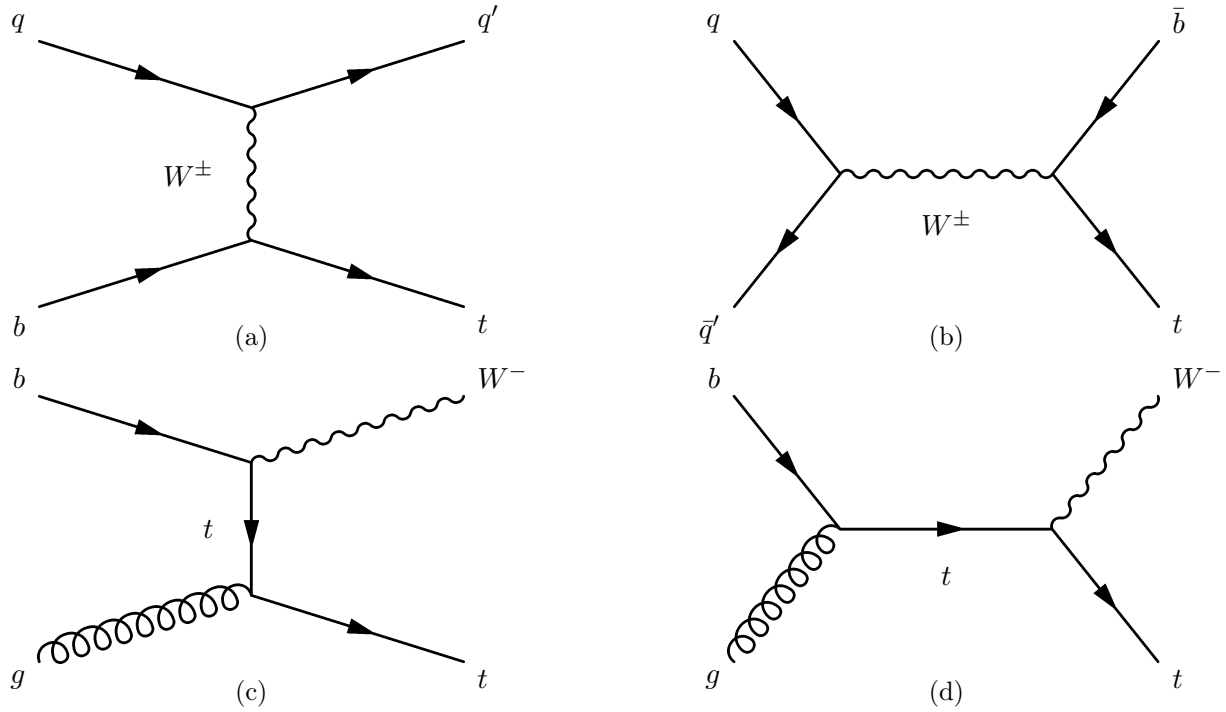


Figure 1.5: Leading order Feynman diagrams of the single-top processes: (a) and (b) are the t–channel and the s–channel respectively, while (c) and (d) show the tW case.

1.2.1 The tW process

This production mode of single tops is mainly characterised by the W boson that is obtained along with the top quark, where the former can be on its mass shell (i.e. $q^2 = -m_W^2$). In addition to the research interests of top physics that were listed in previous paragraphs, there are particular features that make the tW channel attractive, such as its sensitivity for new physics (e.g. [62], [11]) as well as its role as background process in other BSM searches. It is, in fact, the main background of the pair production of top quarks, the $t\bar{t}$ process, with which shares a remarkable feature.

This peculiarity is that, at next-to-leading-order, it interferes with the $t\bar{t}$ production, making it more difficult to precisely define the tW process when higher than leading-order approaches are considered. In Fig. 1.6 Feynman diagrams of the interference processes are shown. As it will be explained in the next chapter, it is necessary to obtain simulations of both processes to perform the analysis, and thus, it is problematic as usually simulations are done for each process separately. In consequence, a double counting issue arises that must be confronted.

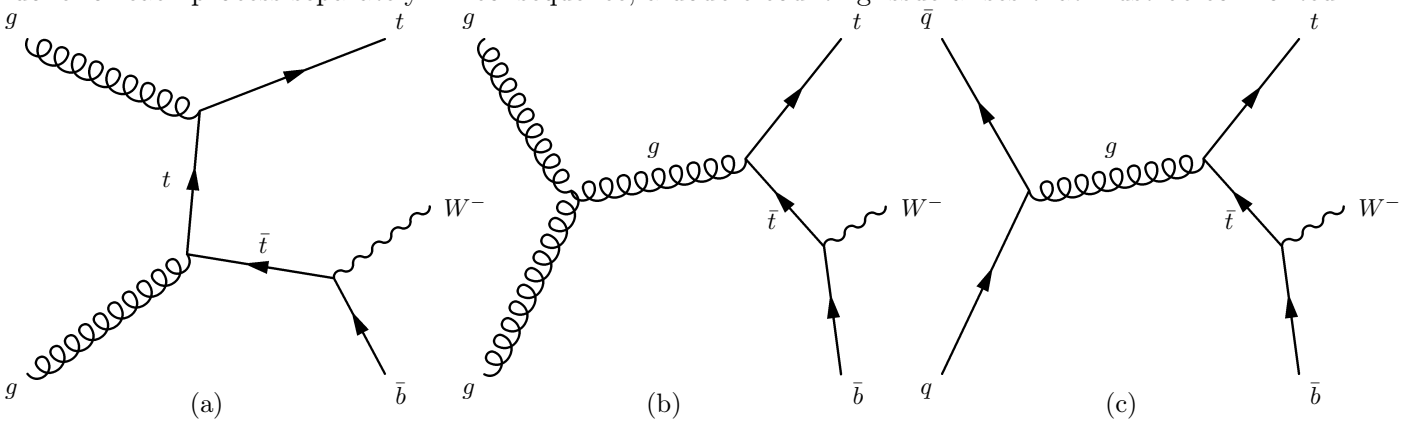


Figure 1.6: Examples of next-to-leading-order Feynman diagrams of the tW process that interfere with the $t\bar{t}$ cross section.

There are two main approaches considered by the community ([43]) to confront this problematic that help to "define" what process is tW at next-to-leading order. The first one, called **diagram removal** or DR removes from the amplitude calculation those contributions that come from diagrams where an internal top line can be on-shell (commonly called "double resonant"). This has an a priori important drawback: gauge invariance is lost. However, it has been seen that the result is independent from the gauge choice for covariant gauges, and that the dependence when taking a non-covariant gauge produces tiny variations, that can be absorbed easily by the statistic uncertainty of the simulation itself. The second approach is called **diagram subtraction** or DS. In it, a new gauge-invariant term is added at the cross section level that removes the contributions of those double resonant diagrams. This conserves gauge invariance, being the downsides here that special treatments must be done in order to obtain such artificial term which induce some small approximations.

In any case, the ideal solution of the issue is to obtain simulations of the combined process of tW and $t\bar{t}$ at NLO (that is, with final states $WWbb$). Unfortunately, during the time this analysis was done, they were still not available inside the CMS Collaboration. As the analysis upon which this is based on (the inclusive cross section measurement) as well as by convention in the community, we use for our tW simulations the DR method, while considering the difference with the DS approach as an uncertainty.

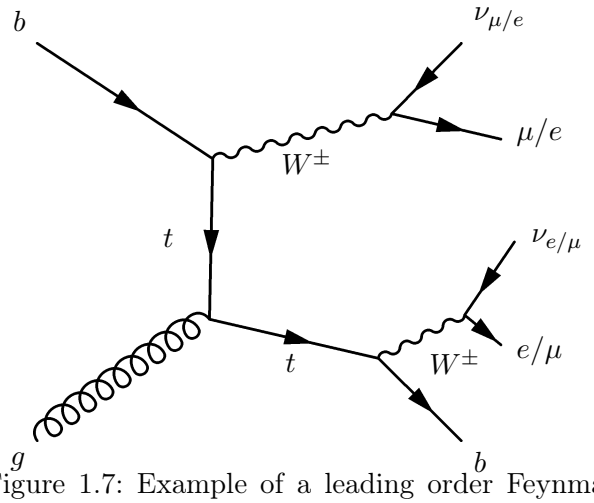


Figure 1.7: Example of a leading order Feynman diagram of the signal process ($e\mu$ channel).

The tW process is considered to be the signal process of this analysis due to its consideration as the aim of study. From the various final states that it can have depending on the decays of the top quark and the W boson, only the set of them where one electron and one muon are obtained are taken into account, i.e. what we call the $e\mu$ channel. The leading order Feynman diagram of this channel is depicted in Fig. 1.7.

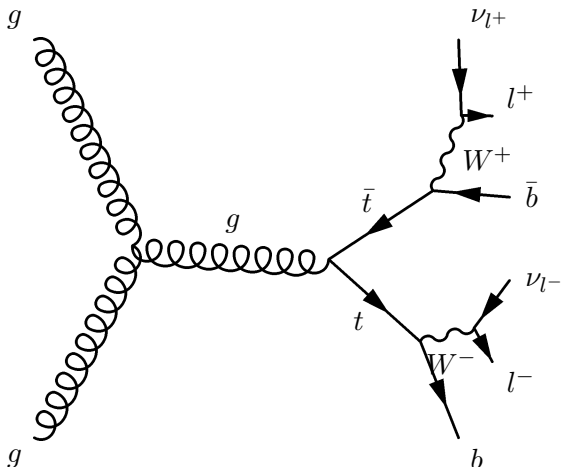


Figure 1.8: Example of a leading order Feynman diagram of the background process $t\bar{t}$ with a final state very similar to the signal one.

As we advanced in the abstract and introduction, the experimental data will come from $p-p$ collisions at $\sqrt{s} = 13\,\text{TeV}$ in the CMS detector at the LHC that will be described in next section. Unfortunately, this process will not be the only one that will happen: we will have to distinguish the final state events of our signal process from others that are not the object of study of this master thesis, commonly called background processes, which have the characteristic of presenting similar (called reducible background processes) or identical (called irreducible background processes) final states to the signal process. For the case of the tW in the $e\mu$ channel, the main background, by far, is the $t\bar{t}$ production, with the Drell-Yan (or DY) production of τ lepton pairs decaying leptonically the second most significative. An example of a $t\bar{t}$ with an

electron and a muon in it is shown in Fig. 1.8.

The choice of the $e\mu$ channel, instead of others with quarks in the final state, helps to select better signal events. The main reason for this is that when quarks are produced in the final state, they soon undergo the process of hadronisation (with the exception of the top quark, and with a small delay for the bottom and charm) due to the nature of the strong nuclear interaction. In this process, quarks conform hadrons and produce as a subproduct other particles: the set of all them is called a "jet", and they are usually worse identified than leptons such as electrons or muons in the LHC detectors. Thus, the choice of the $e\mu$ channel improves our signal event selection and background control (and thus the ratio between both signal and background).

Last results

The tW was not observed in the Tevatron, due to its low cross section in $p - \bar{p}$ collisions there. However, in the LHC, CMS ([23]) and ATLAS ([3]) got evidence for it with $\sqrt{s} = 7\,\text{TeV}$ and the same happened with the observation at $\sqrt{s} = 8\,\text{TeV}$ for CMS ([33]) and afterwards ATLAS ([7]). With $\sqrt{s} = 13\,\text{TeV}$ ATLAS published with an integrated luminosity[12] of $3.2\,\text{fb}^{-1}$ the measurement of the inclusive cross section ([5]), and at the end of 2017 the differential cross section depending on six variables ([4]). However, the most precise measurement of the inclusive cross section was published recently by CMS ([31]) with an integrated luminosity of $35.9\,\text{fb}^{-1}$, yielding the already mentioned value of $\sigma_{tW} = 63.1 \pm 1.8(\text{stat}) \pm 6.4(\text{syst}) \pm 2.1(\text{lumi})\,\text{pb}$. In addition, ATLAS has been able to submit recently a measurement of a differential cross section taking into consideration the interference of both tW and $t\bar{t}$ processes, as the ATLAS Collaboration has obtained Monte Carlo samples of the physical process with those $WWbb$ final states ([8]).

This last analysis (the inclusive cross section measurement by CMS) has encouraged the CMS single top group to aim to measure the differential cross section of the same process (tW), and as result, the analysis explained in this document was started. As it will be explained in the following sections, the quoted inclusive analysis provided the basis for the differential measurements that are the object of this master thesis.

1.3 The LHC accelerator and the CMS detector

1.3.1 The LHC at CERN

The 16th of December of 1994 the CERN[13] Council approved the construction of a proton–proton collider (though also lead–lead and lead–proton) in the old LEP[14] tunnel that was to bear the name of Large Hadron Collider or LHC ([15]). Its goal was to achieve energies at centre-of-mass frame of $\sqrt{s} = 14\,\text{TeV}$, never before seen in an accelerator. That set of energies could lead to significative advances in particle physics. Its construction budget raised to $\sim 3000\,\text{M€}$ and the works finished one decade later, in 2008.

Though its design energy at centre-of-mass frame was $\sqrt{s} = 14\,\text{TeV}$, the current value with which the installation is working is $\sqrt{s} = 13\,\text{TeV}$, achieved after working at 7 and 8 TeV. The main scientific achievement of the LHC

[12] See next section.
[13] European Council for Nuclear Research or, in French: Conseil Européen pour la Recherche Nucléaire.
[14] Large Electron Positron Collider or LEP: an old collider that was in the same place the LHC now occupies.

was the discovery of a Higgs boson-like particle in 2012. In the future, a huge upgrade of the LHC is planned, in what is called the high-luminosity LHC, or HL-LHC. This project ([44]) will enhance greatly the luminosity of the collider, as well as several other upgrades in each detector. With it the amount of data collected will be greatly increased, thus improving the chances of finding BSM physics in it.

([50], [14]) The LHC has circular shape (as seen in Fig. 1.9) with near 27 km of longitude and it is located near the city of Geneva, crossing the border between France and Switzerland, where CERN has its headquarters. It is in an average of 100 m underground, providing thus some isolation from cosmic rays as well as from other undesired perturbations such as vibrations caused by trains or traffic. Two tubes that are installed along the old LEP tunnel are the track through which the colliding particles travel, in opposite directions (one tube from another). These particles are injected in the LHC coming from other accelerators used as injectors that are still operating at CERN, such as the SPS[15], or the PS[16], where they are accelerated to an energy of 450 GeV. They are further energized inside the LHC.

Figure 1.9: [52] Map of the surroundings of Geneva over which a diagram of the LHC is shown.

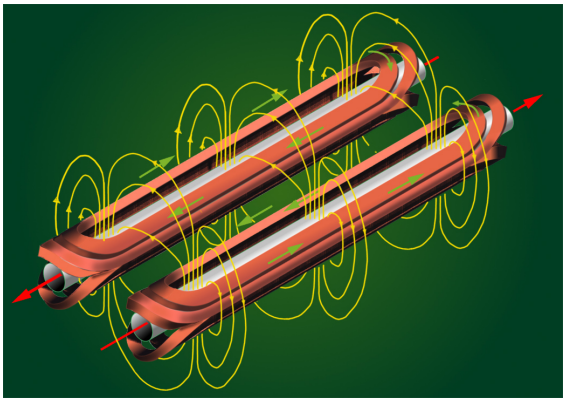
Both tubes cross each other in four collision points, where the main experiments of the LHC are located. These are huge detectors that are able to collect data from the interactions of the colliding particles. The four main experiments are the Compact Muon Solenoid or **CMS**, A Toroidal LHC Apparatus or **ATLAS**, A Large Ion Collider Experiment or **ALICE** and LHC-beauty, or **LHCb**. The two firsts, CMS & ATLAS, are meant to serve as general purpose detectors, while ALICE and the LHCb are more specialized. The CMS detector will be briefly described in the next section.

Particles do not reach the tubes chaotically, but in small packets or bunches, composed of $\sim 10^{11}$ hadrons and separated temporally by 25 ns. Once the particles arrive at the LHC, they still need to be even more accelerated, to increase their energies of 450 GeV up to, at this moment, 6.5 TeV per beam. There are three main devices to remark that help to keep the trajectory of the particles inside the tube, focus the beams, and accelerate the beams. The ones in charge of the first task are the more than 1200 **superconductor dipoles**: magnets cooled down to ~ 1.9 K that can offer a magnetic field of 8.33 T, though this could be increased up to 9 T.

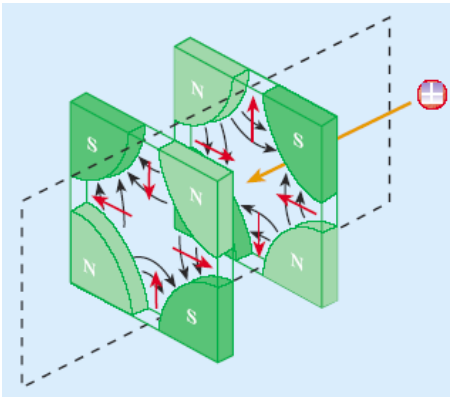
Magnetic multipoles are a set of magnets spatially arranged so that the magnetic field produced focuses the beams of the colliding particles that are distorted by reasons such as gravity or electromagnetic interactions. This allows to control the bunch dimensions, enhancing the number of interactions in each collision point and thus, luminosity. In Fig. 1.10 pictures of multipoles and the superconductor dipoles are shown.

The responsible for accelerating the colliding particles inside the LHC are the **radiofrequency cavities**. They are formed by superconductor magnets that work at 4.5 T that inducing an alternate electric field in the tubes, force the bunches to firstly be equally separated according to a specified frequency (essential in order to make the bunches from the different beams to collide), and secondly accelerate them to the desired energy.

[15] Super Proton Synchrotron.
[16] Proton Synchrotron.



(a) [48] Representation of the magnetic field created by the superconductor dipoles in both tubes of the LHC.



(b) [16] Representation of the magnetic field created by the magnetic cuadripole that suffer the charged particles travelling through the LHC.

Figure 1.10: Pictures of the superconductor dipoles and the multipoles.

Luminosity, pile-up and trigger

The LHC nominal instantaneous luminosity was $\mathcal{L} = 1 \cdot 10^{-34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. However, due to improvements done, it has been already surpassed, now having a value of $\sim 2 \cdot 10^{-34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. The luminosity that each detector measures, however, is not exactly the working luminosity of the LHC accelerator. This is due to the so-called dead time of each detector: lapses of time where no data is collected. In Fig. 1.11 the integrated luminosity recorded during the year 2016 at CMS are shown, as well as the delivered by the LHC.

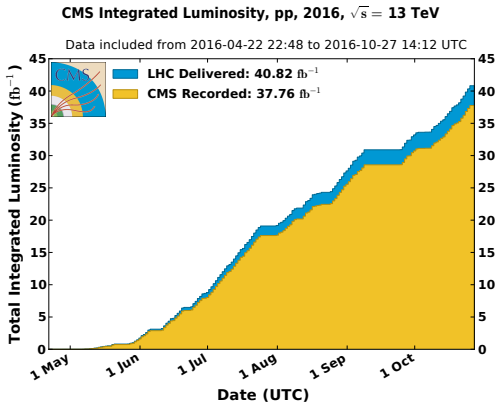


Figure 1.11: [22] Integrated delivered and collected (by CMS) luminosity during the $p-p$ collisions at $\sqrt{s} = 13\,\mathrm{TeV}$ at LHC in 2016.

The reason for this difference is due to the high amount of collisions that at each detector take place (order 10^{34} by squared centimetre and second), where lots of different physical processes take place. Detectors, simply, cannot withstand the enormous flux of information that recording all events would imply: the velocity at which data would have to be saved is $40\,\mathrm{TB} \cdot \mathrm{s}^{-1}$, and it does not exist any kind of technology to support that. What is inevitable done, is to select what events to record. With that intention, there are a collection of tools at hardware and software level that are activated whenever an interesting interaction happens, saving it (thus not all the events are recorded, explaining the differences in integrated luminosity). These procedures and tools are known simply as **trigger** and are specific of each detector. However, the triggers from the main detectors at LHC are similarly disposed in levels: the level 1 trigger is usually hardware-based and does the quickest preselection of events, whereas the superior levels are more based on conventional software and allow to categorise the events depending on different **trigger paths** that depend on different features of the events. Currently for the CMS ([59]) this superior levels are only one, called high level trigger or HLT.

There is another undesired consequence of the high luminosity that is *enjoyed* at LHC. If in one $p-p$ collision an interesting physical process happened that triggered the detector to record the event data, we would very probably not only see information from that specific interaction. As many others per second are likely to

occur, it is highly likely that we would get a "contaminated" event, that might have recorded information of one electron that comes from another interaction between two different protons from the two of the collision of interest. This "pollution" is commonly called **pile-up**.

1.3.2 The CMS detector

The Compact Muon Solenoid or CMS ([20], [37]) is, as we said, one of the main detectors at the LHC, and one of the two with a general purpose, along ATLAS. It is dimensionally speaking large and roughly with barrel or cylindrical shape, with 21.6 m of longitude and a diameter of 15 m, and it is made up of several subdetectors that are organised in cylindrical layers around the two pipes of the LHC that cross it through its axis. It is structurally divided (Fig. 1.12) in barrel (and this one, subdivided in five wheels) and endcaps (which would "close" the barrel). Its main subdetectors, which will be briefly described later in following paragraphs are the detector of trajectories and collision vertexes or **tracker**, the **electromagnetic calorimeter** or ECAL, the **hadronic calorimeter** or HCAL, the **superconductor solenoid** and the **muon system**. When particles cross the different subparts, they can be identified and characterised by the different responses they give in each subdetector, as Fig. 1.13 shows.

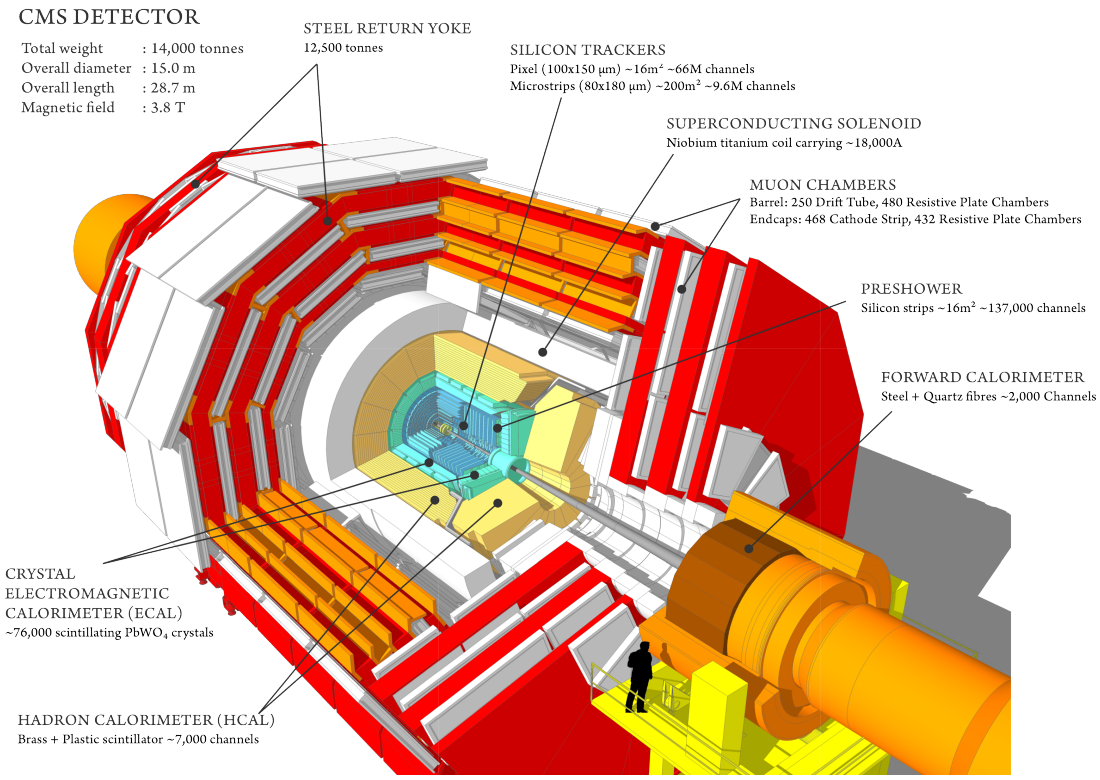


Figure 1.12: [17] Schema of the CMS detector as well as its subdetectors.

1.3.2.1 The coordinate system

The common euclidean coordinate system ($(x, y, z) \in \mathbf{R}^3$) is scarcely used in accelerator physics. The usual z axis is fixed through the pipe, with the positive semiaxis pointing to the Jura mountains, whereas $\varphi \in [-\pi, +\pi]$ and $\theta \in [-\pi/2, \pi/2]$ are used to define the xy plane (φ describing angles between the x and y axis and θ between the z and x ones). Being the origin of coordinates located in the collision point, any location in

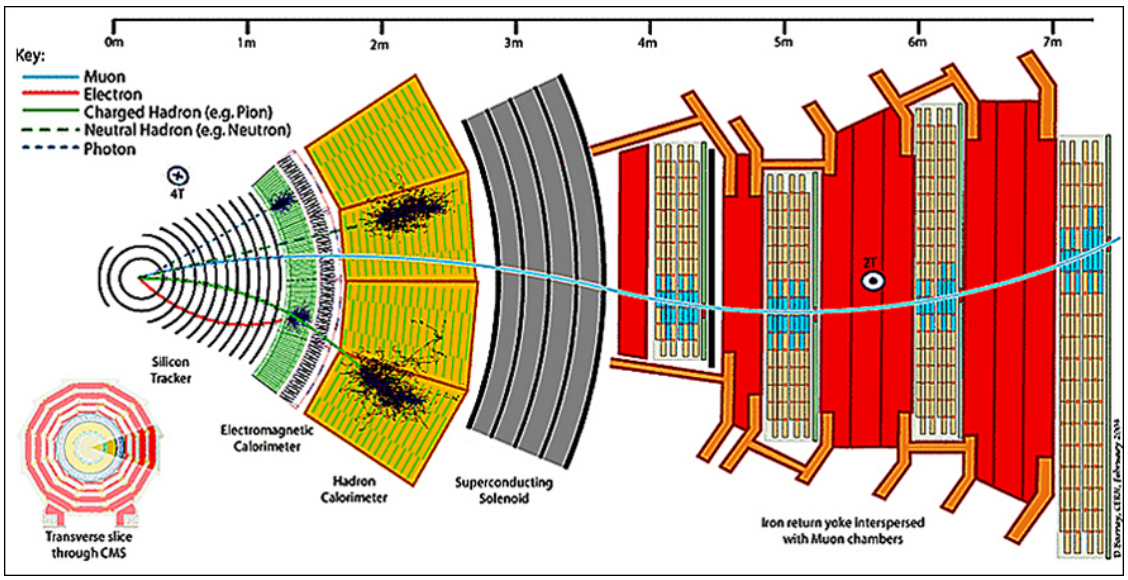


Figure 1.13: [64] Diagram of a circular sector of the detector, in which it can be seen how different particles go through it.

space can be defined with (r, θ, z), with r the distance from the origin to the location. However, if we are interested in the orientation of particles that come from the collision (for example, when we determine the four momentum of one final state particle of a process), we can precise that using (η, ϕ), where $\eta := -\ln(\tan\theta)$ is called **pseudorapidity**. It can be proven that this value for the direction of one particle can be written as

$$\eta = \frac{1}{2}\ln\frac{|\vec{p}| + p_z}{|\vec{p}| - p_z} \simeq \frac{1}{2}\ln\frac{E + p_z}{E - p_z} =: y, \tag{1.6}$$

where $\vec{p}$ is the momentum of the particle, p_z the absolute value of the z component of it, E the energy and y the **rapidity**. The pseudorapidity and rapidity are approximately equal as shown in that expression if $m \ll p$, a limit usually accepted in the conditions of experimental high energy physics.

1.3.2.2 Subdetectors

Beginning from its core, the first component of the CMS detector is the **tracker**, whose main mission is to obtain the momentum of charged particles that cross it by measuring the curvature of their trajectories. It also has the secondary mission of locating the interaction point of the interest process, more usually called "primary vertex". It is mainly made of silicon pixels, which offers granularity yielding a precision of $\pm 10\,\mu m$. It is cooled down using a gas system to -20^oC to reduce the aging and the effect of the radiation, as well as the heat due to the high number of connections. The pixels work as 65 millions of receptors that whenever a charged particle crosses them collect the electrons of the ionised silicon atoms and transform them into a signal. A similar functioning happens for the outer part of it, made of a barrel of silicon strips.

The next subdetector is the **electromagnetic calorimeter** (or ECAL), responsible for measuring the energy of the electrons and photons. It is made of a solid scintillator, lead tungstate, that is disposed in form of crystals, giving the advantage of celerity, as scintillation is a quick and known process. Whenever an electron or a photon enters the ECAL, it will soon interact and trigger an electromagnetic shower (a sequence of processes due to the interaction of an electron or photon with matter) that will produce scintillation, a process in which photons are

released as a consequence of the interaction of the incident particle with matter. Photons are then collected and its signal enhanced in a photomultiplier that also transforms it into an electric signal. There are of the order o 80000 of these crystals that rise the weight of the electromagnetic calorimeter to the 100 tn. A brief mention deserves the so called preshower of the ECAL, whose aim is to provide a better differentiation in trajectories for particles like neutral pions that might have a high velocity in the z direction after being produced and thus their most probable decay, two photons, too: these two photons would be usually interpreted as one, in these cases. The preshower, located in the endcaps, has a higher granularity that grants a better capability of identifying both of them. It is made of lead and silicon.

After the ECAL, the **hadronic calorimeter** (or HCAL) measures the energies of hadrons, i.e., particles made up of quarks, such as kaons or pions. It is a sampling calorimeter, made of successive layers of an absorbent material, and a fluorescent scintillator. The procedure for collecting the light of each scintillator is the same as in the tracker or the ECAL, but now the energy of different layers of scintillating material (that is said to form a "tower") is collected using optical fibres, and then the energy of the hadron is said to be the sum of all the light that the hadronic shower (analogous as the electronic, but more complicated and with more type of processes) produces whenever a hadron enters the HCAL. It is hermetic, in order to not let any kind of particle (except muons and obviously neutrinos) that might have not been stopped yet to go beyond it. This will ensure later that we can obtain information from particles that we cannot directly detect, as will be explained in the next chapter.

Between the HCAL and the muon system the **superconducting solenoid** provides a large magnetic field of 4.2 T. It is actually the largest superconducting magnet ever built, with 12500 tn of weight, and its function is essential. Thanks to it, the momentum (through curvature) of charged particles can be obtained, and we can also differentiate between positive and negative charged particles. To guide the field lines outside the solenoid, though still inside the detector, a structure scaffold-like called the **return yoke**, has been constructed through all the muon system. It is made of iron and helps define the magnetic field lines.

Finally, the **muon system** has been constructed in between the return yoke structure. It is the most outer subdetector due to the fact that muons are expected to be able to exit the hadron calorimeter, due to their comparative long mean lifetime and the speeds at which they will surge from the collision point. Its main function is to identify the particles as effectively muons, and to measure their momenta by tracking their trajectories, an effort for which their information is crosschecked with the one from the tracker, allowing usually to obtain a good muon identification. There are three elements that conform the muon system. The most important and numerous are the drift tubes (DT). These are small cavities filled of gas in which an electric field is present between an anode (a wire through all the tube) and a cathode. Whenever a muon crosses the gas, it will ionise some gas atoms and thus generate electrons that will drift to the cathode, triggering a signal. The DT are located in the barrel and dispose in several layers through the return yoke structure. Cathode strip chambers (CSC), instead, are used mostly in the endcaps. They are chambers also filled of gas, and with a similar working method as the DT, but their anodes and cathodes are disposed orthogonally, allowing to measure both in only one CSC, which is an advantage considering that in the endcaps there is no return yoke and the magnetic field is less uniform, thus identifying particles with higher precision is more necessary. The third component of the muon chambers are the resistive plate chambers (RPC), whose working method is also similar to the other two, but far more faster. They are distributed in the barrel and endcaps and their information is used for the trigger procedures.

2 Methodology

The research in experimental particle physics is done with the absolutely necessary help of computational tools. The main reason being the huge amount of data that is usually analysed in the field. Though not all particle physics analysis are the same, and they do not follow identical procedures, there are general similarities between them. In this chapter, we present a description of the methodology of the analysis, which is explained in the first section. Afterwards, the implementation section shows how these procedures are exactly used in our analysis.

Before that, a brief view of the experimental workflow is presented. The aim of this master thesis is to measure the differential cross section of the tW process, as explained in the abstract, introduction and first chapter. To do so, we first apply a selection procedure to $35.9\,\mathrm{fb}^{-1}$ of integrated luminosity of data obtained in the CMS detector during the year 2016 in order to get a set of events with similar features to what we expect from the tW processes. In addition, this selection must be done also to simulations, in order to perform afterwards the procedure of signal extraction. This is necessary because, although our selection criteria might be very well chosen, our detector is not perfect, and mistakes identifying objects, confusions with other collisions from pile-up can happen, as well as the presence of events from irreducible backgrounds (which have identical final states as the signal process). Thus, is almost certain that the selected group of events from data will not be pure. As a consequence, a procedure for signal extraction is needed.

The main procedures for these steps are essentially a continuation of the work done with the publication by CMS ([31]) of the measurement of the inclusive cross section of the same process, tW. However, the analysis is obviously not the same. When considering total cross sections, once the signal extraction is done, the calculation of the final value is direct. Unfortunately, when considering the differential one, this is not possible. The differential cross section depends on the values of variables of the process, as it has been said, but the values measured do not necessarily coincide with the real ones, as any detector is not perfect, and variations can happen. Consequently, a method is used to try to eliminate the effect that the detector has in the chosen variables, called unfolding. All these procedures will be explained in the following section

2.1 Description of the experimental tools

2.1.1 Generation

The processes related to the production of the simulations needed in particle physics are identified with the word **generation**. These simulations belong to the so-called "Monte Carlo" type (or simply MC), composed of methods based on (pseudo)random numbers. As explained in the previous chapter, the collisions in high energy physics yield to a very complex picture, with multitude of physical processes happening at the same time, and

thus the generation of these simulated samples of events is very complicated. There are various generators, each one with its own different physics model, that simulate the hard process, such as `Madgraph`, `Powheg` or `aMC@NLO`. Afterwards, other programs such as `Pythia` or `Herwig` treat the soft process part, that is essentially, the parton shower. Then, the `Geant4` software allows to simulate how the response of the particles will be when they cross the detector. Finally, they are treated as if they were data, considering even which trigger path would make that event, if it was an actual one, be recorded (or not), and also the "hits" (signals obtained in some part of one subdetector) in each subdetector that particles would have made.

When these steps are complete, we obtain a set of simulated events that is comparable to data and thus, the next step can be done: the reconstruction.

2.1.2 Event reconstruction and identification of objects

The raw data of the detector (or the simulated data samples obtained as previously explained) does not tell if there were two muons or not in the event, or which momenta they had then. The process of extract that information from raw data or simulated data is called **reconstruction**. Their foundations are set upon the **particle flow** algorithm or PF ([34], [42]). Its function is to, taking the raw data, combine it to remake the physical objects of the event (i.e. electrons, muons, charged hadrons...). The algorithm builds up some elements (tracks and clusters) that are used afterwards to recognize those objects.

Using information from the tracker and the muon system, the **tracks** or trajectories of particles are reconstructed using an algorithm with few iterations (four or five). The second element that is used are the so-called **clusters** in the calorimeters, obtained from the energy depositions that particles deposit in both the ECAL and the HCAL. Afterwards, these two elements are combined to compose the objects that are candidate to be muons, electrons, etc., reconstructing its four–momenta after the collision. The first object to be reconstructed is the primary vertex. Afterwards muons, as they require hits in the muon system that other signatures from particles do not. Then, electrons and charged hadrons, using information from the tracker, ECAL and HCAL. Once the electrons have been identified, the remaining clusters in the ECAL must be from the photons, and the ones that remain in the HCAL, must be from the neutral hadrons.

As the requirements and context of all analysis are very different, and the reconstruction is the same for all of them, the objects that the PF algorithm yields are not defined in a very restrictive way. This is done so that each analysis can afterwards impose their conditions to identify their objects: these conditions are put upon variables of the candidate objects that the PF algorithms give. There are, however, some general recommendations about which inside the CMS Collaboration from the called physical object groups (POG), that are specifically dedicated to study how we can clearly reconstruct and identify the different particles we measure in the detectors.

Here follows a brief description of how each object is reconstructed and the variables that are used in this analysis to identify them.

Primary vertex As we already said, the primary vertex (PV) is the point where the two protons that collided to yield the interaction of interest and triggered the recording of the event happened. It is reconstructed projecting the trajectories from the tracker and the muon system into the collision point. As pile-up is always present, the PV is chosen between the many that appear as the one that have the highest sum

of the transverse momentum, p_T, from its reconstructed physical objects. The impact parameters of the particles in the different dimensions (d_x, d_y, d_z) are usually used later as criteria for the identification of other objects, and measured in centimetres; they are sometimes given respect to the transverse plane (d_{xy}).

Muons Given that muons leave a trace in the tracking detector and are able to transverse the detector until the muon chambers, they are reconstructed combining the information of the tracker and the muon system. Once this is done, various variables are used to correctly identify muons. Two variables that are used also for other objects are the transverse momentum, p_T, and the pseudorapidity, η. Constraints on this variables are put because in high energy collisions we expect that often the products of the main interaction move in the transverse plane, as they come from an interaction in almost perfectly opposite directions of particles with similar momentum (we never know exactly the momentum of quarks inside the protons). The constraints in η are also understandable, as low values of eta imply particles pointing to central regions of the detector whereas high values are common in particles that point to the endcaps, or more generally, the "forward" regions, where more undesired signals coming from uninteresting processes are expected. Thus, demanding low η and high p_T we get objects in the central region, which is generally speaking more "clean" of background hits. In addition, for $\eta > 2.4$ CMS cannot detect muons.
In this analysis other variables are used to correctly identify muons. Quality criteria are imposed in the global fit to the muon track measured both in the tracking and muon systems. Additionally minimal requirement of muon hits in each subsystem are applied Another example of variables used for the identification is the so-called isolation. This variable is used because of the process of hadronisation described in the first chapter: one of its subproducts can clearly be leptons, and we might misidentify one of those leptons as if it was coming from the final state of the physical process of interest. To correctly identify the leptons that come from W or Z bosons (the majority of the ones that interest us, usually called in the community jargon "prompt") that come from the main process from the other leptons (idem "fake" or "non-prompt" leptons), the isolation variable I is defined for a lepton l as $I(l) := \sum_{\Delta(R_i)<\Delta(R_{\text{max.}})} |\vec{p}_T(i)|$, where R is the distance in the (ϕ, η) space defined as $\Delta(R) := \sqrt{(\eta_i - \eta_j)^2 + (\varphi_i - \varphi_j)^2}$ and where the sum is done over all the i particles that surround the lepton l in a cone of a maximum value of ΔR. Usually, a cut is imposed on the relative isolation, defined easily as $I_{\text{rel.}}(l) := \frac{I(l)}{|\vec{p}_T(l)|}$. A particularisation of the general isolation for muons, called the relative muon isolation (RMI), is used in our analysis.

Electrons The reconstruction of these objects is done in a similar way as the one from muons, but without, obviously, the tracks from the muon system: only the trajectories derived from tracker data and the clusters from ECAL are considered. However, it is important to take into account that electrons will produce more Bremsstrahlung radiation due to their low mass, which is a relevant factor for the track reconstruction.
Regarding the variables used for the identification, apart from the transverse momentum, pseudorapidity, or a relative electron isolation (REI), a combined classification done by the electron and photon POG of the CMS is used. This classification is done in four subcategories, commonly called working points: veto, loose, medium and tight. These classes, that are ordered from the less exigent to the most ones, depend on conditions imposed over different variables on which we will not deepen.

Quarks and hadrons: jets As explained in the first chapter, when quarks appear in the final state of our process, they hadronise (with the exception of the top quark, that decays before hadronising) yielding at the end a set of particles, with one or more hadrons as well as (e.g.) muons or electrons, that we usually call jet (see Fig. 2.1 (a)). Each particle is reconstructed following its corresponding guidelines: the hadrons are reconstructed taking the information from the HCAL clusters and the tracker when they are charged, when they are not they appear as HCAL clusters to which a track cannot be linked.

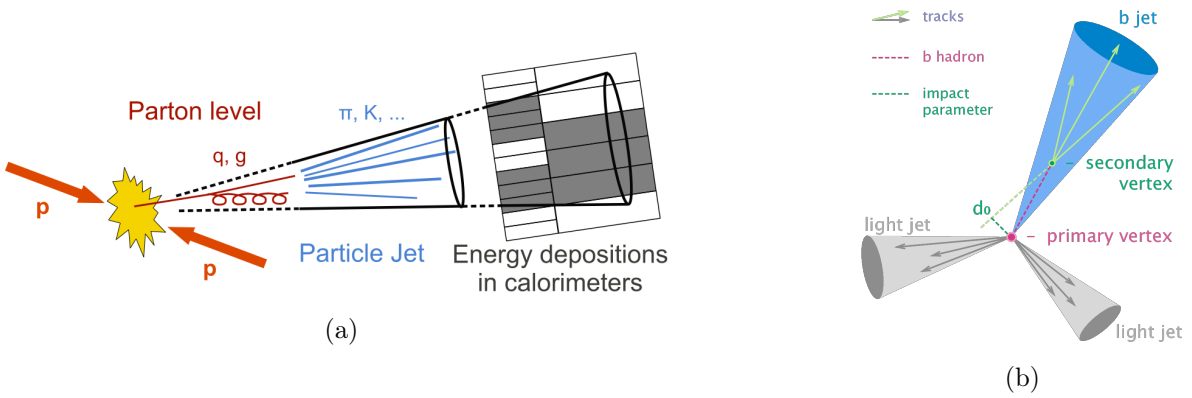


Figure 2.1: Diagrams of the production of a jet ([18]) and a b-jet ([65]).

Jets are complex objects whose reconstruction begin from the information of the PF algorithm that afterwards is combined by other mechanism: the anti-k_t algorithm ([53]). Using the transverse momenta of the candidate particles given by the PF algorithm the anti-k_t one reconstructs each jet making it possible to obtain at the end the four–momentum of the quark from the main process that originated the jet. Unfortunately, it is generally not possible to identify the flavor of the quark (i.e. to say whether it is a strange, charm, down... quark), except from some cases.[1] The bottom quark is another special situation: it does not have a mass so large as the top quark, but it is still the second one when considering all quarks. Thus, when a bottom quark arises in the final state of one process, it hadronises producing initially a hadron: a B meson. However, due to the mass of the bottom quark, this meson does not have a mean lifetime so high as other hadrons, and consequently decays soon after being produced, inside of the detector (and actually, inside the pipe of the LHC). Thus, a situation like the one depicted in Fig. 2.1 (b), where a secondary vertex (were the hadron produced from the b quark decays) can be seen. The precision of the CMS tracker allows to identify this secondary vertex and the impact parameter between it and the PV and with it we can identify the b quarks.

To do this the CSV and CSVv2 algorithms ([25]) have been developed. They take information from tracks, impact parameters, etc. and return for each reconstructed jet a value (called b–tagging discriminant) between 0 and 1 that describes the probability that one jet is a "b–jet" (a jet coming from a b quark) or not. This whole procedure to try to identify these b–jets is called b–tagging.

Regarding the identification of all jets, usually variables such as η, p_T are common, as well as the b-tagging discriminant. in order to avoid leptons being reconstructed as jets, only those jets that are separated a fixed distance $\Delta(R)$ (in the (η, ϕ) space) of identified leptons are taken into account. This is called in the jargon "cleaning".

[1] We here ignore the case of the top quark because as we already stated in the first chapter, it decays so quickly that it does not hadronise. Thus, a "top jet" is not produced.

Taus ([35], [57]) Tau leptons cannot be reconstructed and identified as its colleagues electrons and muons. The reason is that it is the only lepton with the sufficient mass ($m_\tau = 1.777\,\text{GeV}$) to be able to decay to hadrons. Actually, it is the main decay channel, with a branching ratio (in percentage) of roughly the 65% , whereas the decay to muons ($\tau \to \mu\,\nu_\mu\,\nu_\tau$) and electrons ($\tau \to e\,\nu_e\,\nu_\tau$) have a $\approx 17.5\%$ each. Taus that decay to electrons and muons are hard to be identified, and thus usually they are not considered. An algorithm has been developed in order to identify those tau leptons that decay hadronically that is called the "hadron-plus-strips" algorithm, which takes into account several kinematic variables.
In this analysis, taus that hadronically decay are not considered, though as it has been already said, those that decay to electrons or muons pass the selection because we are unable to identify if they are truly a final state electron or muon, or come from a tau decay.

Other features of the event There are other variables that can be defined for each event that are useful. One example of these is the **missing transverse energy**, or E_T^{miss}. It is defined as the modulus of the opposite vector that is obtained from summing all the transverse momenta of all the objects reconstructed by the PF algorithm, i.e.

$$E_T^{\text{miss}} := \left| -\sum_i \vec{p}_T(i) \right|. \tag{2.1}$$

It is important to note that, although the E_T^{miss} is considered as an "energy", by its own name, formally it is not, as it would only be that presumed (transverse) energy if and only if all the particles whose momentum is summed had zero mass.
In the final states of some processes, particles such as neutrinos appear. Because they only interact weakly, they are very hard to detect, and even more in a context with the order of 10^{34} interactions per second and squared centimetre. The E_T^{miss}, stands out here because by definition, if in an interaction, one neutrino is produced, it almost surely will not been directly detected, but the missing transverse energy allows us to assign a transverse momentum to it due to momentum conservation. Consequently, E_T^{miss} is a way to try to study particles that are not directly detected, although it has limitations: e.g. if instead of one neutrino, two were produced, then the E_T^{miss} would ideally correspond to the modulus of the sum of both transverse momenta, and they are inseparable. In this analysis E_T^{miss} is used to take part in a multivariate analysis that is done, as will be explained later in the document, and to veto some events that have unusual and unexplainable values of E_T^{miss}.

2.1.3 Mathematical resources: BDT, fits, and unfolding

Nowadays the work in the field of particle physics is in general complex, as one can be analysing physical processes that have a low cross section, or that have associated backgrounds that are overwhelmingly more present than the process of interest. In such a context, it is crucial to obtain improvements that might enhance your signal against your background, for example. To do so, it is now common the use of multivariate analysis (or MVA), that combine several characteristics (variables) of the events or physical objects at the same time to discriminate events. Its advantage is that although those variables by their own might not be a very good discriminator, the scan of the multidimensional space of all of them at the same time can be better, although it is not easy to understand a hyperspace of several dimensions.

A **boosted decision tree** (or BDT) is a clear example of a MVA algorithm. They can be understood as an

evolution of the "traditional" tool of the decision trees (see Fig. 2.2), which are used for the aim of classification of events. The main idea behind boosting is to construct a powerful learner out of an ensemble of weak learners. In the case of a boosted decision tree, an ensemble of shallow trees is trained. The training of the trees is performed sequentially, and the training of each tree depends on the false positives of the previously trained trees. There are different boosting algorithms, though their main idea is the same: after each iteration of the algorithm, improve the classification of the events by varying the selection criteria. At the end, all BDT are able to classify events by associating each one with a value called discriminator, which encompasses all the classification of the several variables. This value usually goes from -1 to $+1$ and the idea is that the BDT is usually trained so that events that have values near $+1$ tend to be more signal–like, whereas those near -1 are more background–like.

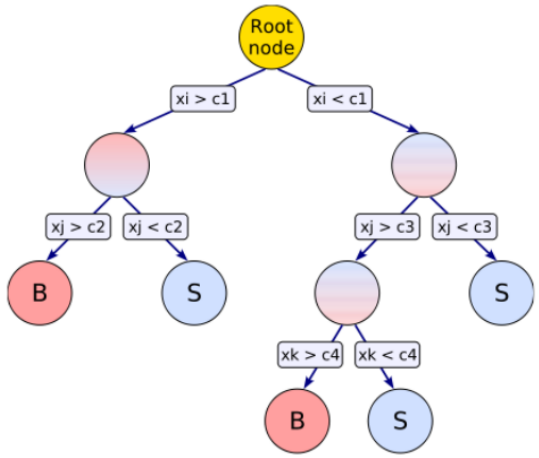


Figure 2.2: Example of a decision tree. The root node is the beginning point, which contains all the set of events.

A crucial part of almost any particle physics analysis is to extract the signal from your observations. As it has been already said, the data we collect do not have only events related to the signal process: they almost certainly have events coming from other processes (background), although one could take a lot of care when doing its event selection. Thus, if one physical observable, e.g. one cross section, must be calculated, the amount of those observed events that belong to the signal process must be known, as we always will have background. This procedure is known as signal extraction. In order to perform this, one could just take the difference between the data and the background processes modelled by MC simulations. However, other procedure that can be done are fits to the shape of distributions through maximising the likelihood. This is commonly called a **maximum likelihood fit**.

The likelihood function, denoted as $\mathcal{L}$, is an application that parametrises the feasibility (or plausibility) of a given model or a set of parameters according to some observation: the higher the values of the likelihood for some given arguments, more likely that set of parameters have yielded to the observed data. Thus, a likelihood for a given variable or distribution can be constructed depending on our observations. Afterwards, using as inputs our MC simulations into the likelihood, we can use a numerical procedure in which beginning from those inputs, we maximise the likelihood. At the end, when this maximisation is done, the amount of events corresponding to the signal process can be extracted.

Numerically, the maximisation of the likelihood is done by minimising the $-\log(\mathcal{L})$ (commonly called log–likelihood), which is a better computational approach[2]. The likelihood is constructed based on a binned histogram, as follows.

$$\mathcal{L}\left(\vec{r}, \vec{s}(\vec{\theta}), \vec{b}(\vec{\theta}), \vec{\theta}\right) := \prod_{i=1}^{N_{\text{bins}}} \text{Pois}\left(n_i | r_i, s_i(\vec{\theta}), b_i(\vec{\theta})\right) \cdot \prod_{j=1}^{N_{\text{uncs.}}} e^{\frac{\theta_j^2}{2}} \tag{2.2}$$

[2]This has also statistical advantages, as $-2\log(\mathcal{L})$ features interesting asymptotic propierties.

Here, $\vec{n} = (n_1, \ldots, n_{N_{\text{bins}}})$ represent the observed values (the data from the detector) in each bin i. The arguments of the likelihood are each r_i, which is a parameter that scales the amount of signal in the bin i which is exactly s_i, whereas b_i represents the amount of background in it. The vector $\vec{\theta}$ is the last input of this function and it is a collection of several parameters which are in the likelihood to represent other information regarding our analysis: essentially, uncertainties (e.g. the uncertainty in the luminosity). The $\text{Pois}(n|\nu)$ represents a Poisson probability density function, that is defined as $\text{Pois}(n|\nu) := \frac{e^{-\nu}n^{\nu}}{n!}$. In our case, the arguments are:

$$\begin{aligned}
&\text{Pois}\left(n_i|r_i, s_i(\vec{\theta}), b_i(\vec{\theta})\right) = \tfrac{1}{n_i!}\left(r\cdot s_i(\vec{\theta}) + b(\vec{\theta})\right)^{n_i} e^{-\left(r\cdot s_i(\vec{\theta})+b(\vec{\theta})\right)} \\
&s_i(\vec{\theta}) = s_i^T \prod_{k=1}^{N^S_{\text{norm.}}} q_k^{\theta_k} \prod_{m=1}^{N^S_{\text{shape}}} p_m(\theta_m) \\
&b_i(\vec{\theta}) = b_i^T \prod_{l=1}^{N^B_{\text{norm.}}} q_l^{\theta_l} \prod_{m=1}^{N^B_{\text{shape}}} p_n(\theta_n)
\end{aligned} \tag{2.3}$$

In these expressions, s_i^T and b_i^T represent the total amount of events in the bin i of signal and background respectively, $N^S_{\text{norm.}}$ and $N^B_{\text{norm.}}$ show the amount of normalisation uncertainties[3] in signal and background (resp.) and the values $q_k^{\theta_k}$ and $q_l^{\theta_l}$ are factors that represent those uncertainties. E.g. if one of them were of the 20%, then $q = 1.20$ and thus $q^{\theta} = 1.20^{\theta}$. The factors $p_m(\theta_m)$ and $p_n(\theta_n)$ represent how each uncertainty that affects the shape of the distribution influence the total amount of signal and/or background. They are proportional to a quadratic polynomial that depends on the corresponding θ whose coefficients are determined before the algorithm begins to iterate through a vertical interpolation in each bin i taking into account the varied shapes of the distribution for each uncertainty (one time when a $+1\sigma$ uncertainty is propagated into one distribution and another when is the -1σ). In a nutshell, we can separate two groups of uncertainties that are represented by those parameters θ: the normalisation and the shape ones, and each of them has an additional Gaussian term as appears at the end of Eq. 2.2.

These Gaussian terms, that conform the second factor of eq. (2.2), make our likelihood partly Bayesian, as it express a degree of belief of the probability of those θ parameters[4]. This term is used to introduce the effects of the uncertainties thanks to the θ parameters associated to them. When the minimisation algorithm iterates, new values of all parameters θ and r_i are taken and with them s_i and b_i are derived.

The parameter r_i is our main aim in this procedure, as afterwards we can just multiply it by s_i (for each bin) obtaining the signal (what, after all, we wanted). Thus it is called the **parameter of interest** (POI). The remaining parameters of the fit, although necessary in order to construct a realistic statistical model, are "secondary" and not of our interest, and thus they are called **nuisance parameters**, or more commonly,

[3] I.e. uncertainties whose effect is only in the total amount of events of a given process, not of its distribution. Uncertainties in luminosity, or the normalisation of the MC samples (see next section) are the ones this analysis takes as "normalisation" here.

[4] Likelihoods are *in principle* built as $p\,(\text{Data}|\text{Model})$ (taking it as a function of the model's variables), but what we have called before likelihood, $\mathcal{L}$ adds information about our model, effectively a term like $p\,(\text{Model})$. This is something impossible in the frame of frequentist statistics (to express knowledge a priori about our model; likelihood functions exist in frequentist statistics) applied to physics (as "the model" or "the parameters" are supposed to be unknown, and thus we cannot know its probability), though in the Bayesian ones such assumption is not crazy at all and the Bayes theorem itself gives a proxy to its understanding:

$$p\,(\text{Model}|\text{Data}) \propto p\,(\text{Data}|\text{Model}) \cdot p\,(\text{Model}) \sim \mathcal{L}.$$

This method is necessary in order to assess our uncertainties, what makes our $\mathcal{L}$ a mixed frequentist–Bayesian object. We will not enter in the details of this: we can state, however, that it is a common procedure in the field.

"nuisances". As we after the signal extraction will still have to perform the unfolding, we must look after a way to propagate the uncertainties. Fortunately, the procedures of maximising maximum likelihoods can be transformed into a minimisation of χ^2 parameters, that can be expressed as a function of a covariance matrix of the fit that carries the uncertainties. This matrix can be calculated as the Hessian of the $-2\log(\mathcal{L})$ and allows to take into account the correlation that all the parameters of interest (i.e. the r_i) have between them.

Once the amount of signal of our analysis is extracted, only the last stage remains: the **unfolding**. The variables of the particles we store, such as the transverse momentum of one muon, do not reflect exactly the real features of the particles we measure. The reason is that our detector affect the measure itself, as its components interact with the particles, and consequently our measurements are smeared (e.g. measuring a muon of 20 GeV as one of 35 GeV). Other possibility is that our detector, although the particle do passes through it, does not detect it. And on top of these effects, one must take into account that each measure has its own uncertainty and that as we are seeing these distributions of the variables from histograms, we depend also in the amount of data that we have.

When one tries to measure a total cross section, the effects of changing the value of the variable are not relevant, because the relevant information is the total amount of signal that we have. The effects of not "seeing" the particles are on the other hand taken into account because they affect the total amount of signal. Unfortunately, these effects are relevant when one wants to measure differential cross sections (as we do), because it is relevant to know the amount of events that one have with (e.g.) an electron transverse momentum of $20-30$ GeV and not of $30-50$ GeV. The procedure of removing those effects is called unfolding.

The problem can be easily presented as follows. Let us set the physical values of some variable in some bin i as μ_i and the measured and stored values as ν_i. One can parametrise the effects of the detector in a matrix R called the response matrix. Thus:

$$\vec{\nu} = R\vec{\mu}, \quad \nu_i = R_{ij}\mu_j \tag{2.4}$$

These matrices are a feature of our detector and our methods of reconstructing events, that here are crucial. They are obtained through simulations that only contain signal events whose characteristics are those that we expect that our *actual* data, from the LHC and CMS with our selection, have. In other words, these Monte Carlo samples must contain signal events with the same **fiducial phase space** (e.g. if we demand that leptons must have $p_T > 20$ GeV, then in the simulations we cannot have a different thing than this). Afterwards, what we do is the same procedure as with usual simulations or Monte Carlo: we reconstruct the objects in the events, but now also saving the information of the physical process just after simulating it. At the end, we have both the information of the fiducial phase space that corresponds to the final state of our physical process, and the information that we have after reconstructing it. With this the elements of the response matrix are defined as follows, for a determined binned variable of our events,

$$R_{ij} = \frac{n_{ij}}{n_i}, \tag{2.5}$$

where n_{ij} are the number of reconstructed events whose value of the measured variable fall in the bin j that had the simulated value of the variable in the bin i, and n_i are the number of events whose simulated value of the variable fell in the bin i. The number of bins (and their limits) in the reconstructed (or "folded") space does not have to be the same as in the generated (or "unfolded") space: actually, it can be seen that if the number of bins of the reconstructed space is higher than the one in the unfolded, the problem is ill–posed (infinite solutions will exist) in our approach to it (explained in the following paragraphs): to remove this limitation, and the bias in the proportion of bins in the unfolded and folded space, the relation between both must be of 1:2 (unfolded:folded)[5].

Starting from the problem already presented, it is trivial to guess a very direct solution: invert the response matrix so that one can obtain the values of your variable in the unfolded space. Unfortunately, inverting matrices can be a very challenging numerical problem depending, of course, on the matrix itself: the more diagonal the matrix is, more easy numerically is to be inverted. There is a way to enhance the "diagonality" of R: the choice of binning in the folded and unfolded space. From the previous definitions, two quantities can be defined that will be useful later: the **stability** of a unfolded space bin i and the **purity** of a folded space bin j. They are defined respectively as follows,

$$s_i := \frac{\sum_{j=1}^{N_{\text{bins}}^{\text{fol.}}} n_{ij}}{n_i} \qquad p_j := \frac{\sum_{i=1}^{N_{\text{bins}}^{\text{unf.}}} n_{ij}}{n_j^R}, \tag{2.6}$$

where n_{ij} and n_i are the same mentioned before and where n_j^R is the amount of simulated events in the (folded) bin j. Essentially, stabilities give us a notion of, on one hand, the amount of simulated events that we end up reconstructing and selecting as signal, but also of how many of them stay in the same bin and are not measured elsewhere. Purities are an estimation of the amount of reconstructed events in one bin j related with the signal process over the total number of reconstructed events in that bin. The relevant point is that the maximisation of them through the choice of binning in the folded and unfolded spaces enhances the diagonality of the response matrix, thus making the unfolding problem easier.

The common procedure to perform the unfolding, however, is not that of directly inverting the matrix. It can be shown that the problem can be rewritten as finding the values $\vec{\mu}$ that minimise a χ^2 expression such as

$$\chi^2_R(\vec{\mu}) = (R\vec{\mu} - \vec{\nu})^T V^{-1} (R\vec{\mu} - \vec{\nu}). \tag{2.7}$$

In this expression, V^{-1} represents the covariance matrix that ultimately encodes the uncertainties of our final measurements. At the end, we face again other minimisation problem. This allows to implement a way of dealing with problems that might arise when the response matrix is not very diagonal (a feature that, as already said, is undesirable) that is called **regularisation**. In practice, this characteristic implies to add a new term to the previous χ^2 that modifies effectively the minimisation: the difference between the various regularisation approaches is how this new term is defined. If one were to do such thing, the first conclusion would be that the final result is affected, as all regularisation adds an artificial bias in the whole procedure. However, such a thing might be preferred when the alternative is the impossibility of unfolding at all.

[5] If not, the degrees of freedom would not be equal to the number of free parameters.

In general, all regularisations are modulated by a parameter τ that allows us to get a final χ^2 expression for the minimisation as

$$\chi^2_{\text{unf.}}(\vec{\mu}, \lambda) = \chi^2_R(\vec{\mu}) + \tau \cdot \chi^2_{\text{reg.}}(\vec{\mu}) + \lambda \sum_i (R\vec{\mu} - \vec{\nu})_i \tag{2.8}$$

Here, we have added another term (the last) that is necessary to account for some problems that can arise when bins with a low amount of events are present. If that were to happen, the count of those events would follow a Poisson distribution effectively, not a Gaussian, and for the χ^2 minimisation approach it is necessary that a Gaussian distribution is followed in all bins. To correct for those possible divergences, that last term is added, which helps to take into account the total amount of events (the normalisation). This is term is denoted as an area constraint.

The choice of the parameter τ is crucial, as it determines in the minimisation what gets "more minimised" and thus, indirectly, the bias. Our interest is that the relevant minimisation is done in $\chi^2_R(\vec{\mu})$, the term of the "actual" unfolding, and that the regularisation term affects only the necessary so that we could forget about our problems. One of the ways of doing so is the called L-curve method ([47]).

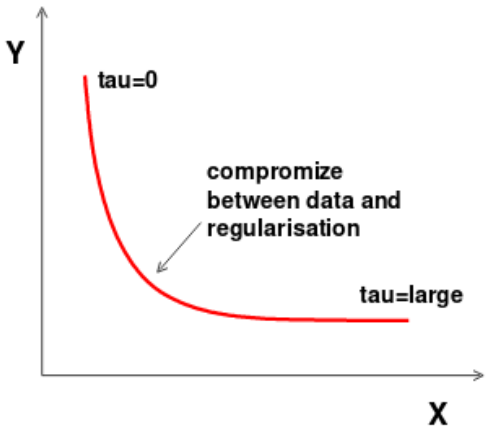


Figure 2.3: Schema of the shape of a L-Curve. the values to the right of the curve are equivalent to a high minimisation of the term $\chi^2_{\text{reg.}}(\vec{\mu})$, which is represented in the y axis. The opposite happens if we move to the left side of the graph, where those lower τ force a higher minimisation in the term $\chi^2_R(\vec{\mu})$.

In it, a plot is made of the term $\chi^2_R(\vec{\mu})$ in the x axis vs. $\chi^2_{\text{reg.}}(\vec{\mu})$, and a scan of values is done varying the τ values. The result is usually a graph with the shape of an "L", as seen in Fig. 2.3. Once the L-curve is established, the value of the tau parameter is chosen as the corresponding to the maximum curvature point, which represents the best compromise between minimising the interesting term and the regularisation term.

2.2 Implementation

The entire analysis has been coded mainly in `C++`, `Python` as well as `Bash` , along with `ROOT` (and `PyROOT`): a set of libraries built upon `C++` used essentially by the community of particle physics due to the several handy functions and classes that it posses. The framework (i.e. the structural code that articulates any analysis done within it) used ([45]) has been developed by the Experimental High Energy Physics Group of the University of Oviedo on the basis of PAF (**P**ROOF **A**nalysis **F**ramework, [46], [41]). This is a software in `C++` and `ROOT` that allows to structure easily and orderly, in a modular way, any usual particle physics analysis.

2.2.1 Data and Monte Carlo samples

The data used in this analysis consists on $35.9\,\text{fb}^{-1}$ recorded during 2016 in the CMS detector at 13 TeV of energy in the centre of masses. The raw data, unfortunately, is still huge and after it is collected, a procedure starts where different **skims** are done on them, based on pre-selections. This allows to reduce the size of the

Process	Decay mode	Generator	σ (pb)	Classification
tW^-		Powheg	35.6	tW
tW^-	$\#(\ell) \geq 1$	Powheg	19.47	tW
$\bar{t}W^+$		Powheg	35.6	tW
$\bar{t}W^+$	$\#(\ell) \geq 1$	Powheg	19.47	tW
$t\bar{t}$		Powheg	831.76	$t\bar{t}$
$t\bar{t}$	$t\bar{t} \to l + l' + \nu_l + \nu_{l'}$	Powheg	88.29	$t\bar{t}$
DY + *jets*	$\gamma \to l + \bar{l}$ ($m_{ll} \in [5, 50)$ GeV)	MadGraph	22635.09	DY
DY + *jets*	$\gamma \to l + \bar{l}$ ($m_{ll} = 50$ GeV)	MadGraph	6025.2	DY
W + *jets*	$W \to l + \nu_l$	MadGraph	61526.7	Non-W|Z
$t\bar{t}$	$\#(\ell) \geq 1$	Powheg	451.67	Non-W|Z
ZZ		MadGraph	16.523	$VV + t\bar{t}V$
WW		MadGraph	115	$VV + t\bar{t}V$
WZ	$WZ \to l + l' + l'' + \nu_{l''}$	Powheg	4.4297	$VV + t\bar{t}V$
$t\bar{t} + \gamma$		aMC@NLO	3.697	$VV + t\bar{t}V$
$t\bar{t}W$	$W \to l + \nu_l$	aMC@NLO	0.2043	$VV + t\bar{t}V$
	$W \to q + q'$	aMC@NLO	0.4062	$VV + t\bar{t}V$
$t\bar{t}Z$	$W^+W^- \to l + l' + \nu_l + \nu_{l'}$	aMC@NLO	0.2529	$VV + t\bar{t}V$
	$Z \to q + \bar{q}$	aMC@NLO	0.5297	$VV + t\bar{t}V$

Table 2.1: MonteCarlo simulation samples used in this analysis.

data, as well as to reformat them, ending up with files that are manageable and with which actual analysis can be done. The intermediate formats, such as AOD or miniAOD ([58]) give us the format of our actual data, which are called HeppyTrees or n-tuples ([24]).

The Monte Carlo simulations used in this analysis are written in the table 2.1. All the samples have been categorised in five groups, that allow to visualize better them in the results. The main process, tW, is one of them, as well as the main background, $t\bar{t}$. The other groups consist in the Drell-Yan processes, where Z or γ bosons take part, and then the $VV + t\bar{t}V$ as well as a more mixed Non-W|Z category.

When obtaining these simulations, a very large number of events of each one is produced. When a comparison with data is wanted, a global reweight of them is necessary. This is done using a weight depending on its theoretically predicted cross section value (and essentially Eq. (1.3)). The uncertainties that this prediction carries (due to the PDF and α_S running) are called of MC normalisation and are propagated in the final result of the analysis. For the sample groups that are considered, a 50% of uncertainty is taken for the $VV + t\bar{t}V$, DY and Non-W|Z groups, whereas a 6% is chosen for the $t\bar{t}$ (from [31]). In addition of this uncertainty, the statistical one linked to the amount of generated events is also propagated to the final result.

To estimate uncertainties related with the modelling more Monte Carlo samples are used: when the simulations are produced, some parameters must be set, and its uncertainty must be taken into account because they depend on the model that is used to generate our samples. To take into account these uncertainties that depend on the model, new simulations that vary those parameters (or features of them, such as the colour reconnection model) are done, and afterwards the difference with the "nominal" simulations are used as uncertainties. The simulations dedicated to estimate uncertainties are collected in the table 2.2.

The uncertainties are taken from the signal process (tW) and from the main background process ($t\bar{t}$), whose contributions are the most relevant: in the first case trivially and in the second case because of the large cross section of this process. For each of them, except in the cases for the colour reconnection and diagram

Process	Decay mode	Generator	σ (pb)	Uncertainty estimated
tW^-	$\#(\ell) \geq 1$	Powheg	19.47	ISR
tW^-	$\#(\ell) \geq 1$	Powheg	19.47	FSR
tW^-	$\#(\ell) \geq 1$	Powheg	19.47	ME scale
tW^-	$\#(\ell) \geq 1$	Powheg	19.47	PS scale
tW^-	$\#(\ell) \geq 1$	Powheg	19.47	DS
tW^+	$\#(\ell) \geq 1$	Powheg	19.47	ISR
tW^+	$\#(\ell) \geq 1$	Powheg	19.47	FSR
tW^+	$\#(\ell) \geq 1$	Powheg	19.47	ME scale
tW^+	$\#(\ell) \geq 1$	Powheg	19.47	PS scale
tW^+	$\#(\ell) \geq 1$	Powheg	19.47	DS
$t\bar{t}$		Powheg	831.76	UE
$t\bar{t}$	$t\bar{t} \to l + l' + \nu_l + \nu_{l'}$	Powheg	88.29	UE
$t\bar{t}$		Powheg	831.76	ISR
$t\bar{t}$		Powheg	831.76	FSR
$t\bar{t}$		Powheg	831.76	Matching ME/PS
$t\bar{t}$	$t\bar{t} \to l + l' + \nu_l + \nu_{l'}$	Powheg	88.29	Matching ME/PS
$t\bar{t}$		Powheg	831.76	Colour r. model 1
$t\bar{t}$	$t\bar{t} \to l + l' + \nu_l + \nu_{l'}$	Powheg	88.29	Colour r. model 1
$t\bar{t}$		Powheg	831.76	Colour r. model 2
$t\bar{t}$	$t\bar{t} \to l + l' + \nu_l + \nu_{l'}$	Powheg	88.29	Colour r. model 2
$t\bar{t}$		Powheg	831.76	Colour r. model 3
$t\bar{t}$	$t\bar{t} \to l + l' + \nu_l + \nu_{l'}$	Powheg	88.29	Colour r. model 3
$t\bar{t}$		Powheg	831.76	Colour r. model 4

Table 2.2: Monte Carlo simulation samples used in this analysis to estimate modelling uncertainties.

subtraction cases, there are two variations ("up" and "down"), which effectively are two samples each. The different modelling uncertainties that appear in the table are the following.

ISR ([61]) The initial state radiation consist on the possible emissions by the initial state that come from the protons of the collision.

FSR ([61]) The final state radiation is analogous to the ISR, but with the final state.

ME scale These samples vary the renormalization scale used in the matrix element.

PS scale As the previous ones, these take into account deviations of the factorisation scale between the parton shower and the matrix element.

DS This sample is the $tW^-/\bar{t}W^+$ process but instead of using the diagram removal schema explained in previous sections, the diagram subtraction model is used. The difference between them are taken into account as a source of a modelling uncertainty.

UE ([61], [27]) These samples vary the parameters related to the underlying event, i.e. the other interactions that might take place between the two protons that give lead to the main collision.

Matching ME/PS ([61], [27]) The link between the matrix element final states and the parton shower evolution depends on the modelling too.

Colour reconnection models ([61], [27], [19], [2]) Inside the particle shower, the modelling of how the colour charge evolves through it is a complex feature and there are different models of it. We take into consideration four of them (apart from the one used by our nominal samples). At the end, the largest uncertainty of the four models for each bin and for each side (either an upper or lower deviation) is taken as the uncertainty of the colour reconnection.

Finally, and in order to perform the unfolding procedure as described before, all the $tW^-/\bar{t}W^+$ nominal and uncertainties samples have been reprocessed to obtain exactly the objects at particle level in our correct fiducial phase space through (as described in [32]) thus obtaining a better representation of the distributions of each simulated particle. These dedicated samples are then used to compute the response matrices necessary to unfold the selected variables.

2.2.2 Trigger selection

The trigger requirements that we impose over the events essentially consist on demanding one electron or one muon, or one electron and muon. In addition, activity in the calorimeters or tracker is required in some of the trigger HLT "paths" (or sets of requirements).

2.2.3 Object identification

As already stated, after the PF algorithm gives us its candidates, a more constrained identification is done using variables such as the described in the previous section. The criteria for the selection of electrons are the following.

- $p_T > 20$ GeV.
- $|\eta| < 2.4$ and $1.4442 \nless |\eta| \nless 1.5660$.
- Cut–based tight ID recommended by the Electron and Gamma CMS POG based on different trace characteristics and other variables.
- Selection in the relative electron isolation in a cone of $\Delta(R) < 0.3$ on different values depending on its pseudorapidity.

For muons, the criteria are:

- $p_T > 20$ GeV.
- $|\eta| < 2.4$.
- Cut–based tight ID recommended by the Muon CMS POG based on different trace characteristics and other variables.
- Selection in the relative muon isolation in a cone of $\Delta(R) < 0.15$ on different values depending on its pseudorapidity.

Regarding jets, we consider firstly the PF candidates with the anti-k_T algorithm with an opening angle of 0.4. In addition, the following selection criteria are imposed upon them.

- $p_T > 30$ GeV.
- $|\eta| < 2.4$.
- "Loose" jet identification recommendation set up by the Jet and MET CMS POG based on different cuts in various variables depending e.g. on characteristics of the cluster.

After identifying one jet, the Combined Secondary Vertex tagger (CSVv2) is used in order to check if it could or not be a jet originated from a b quark. This analysis uses the "medium" working point (with a cut in the value given of 0.8484).

When the data is reconstructed, there are events where the amount of missing transverse energy is known to shelter problems. These are then identified and tagged. In this analysis we filter those events.

2.2.4 Event selection

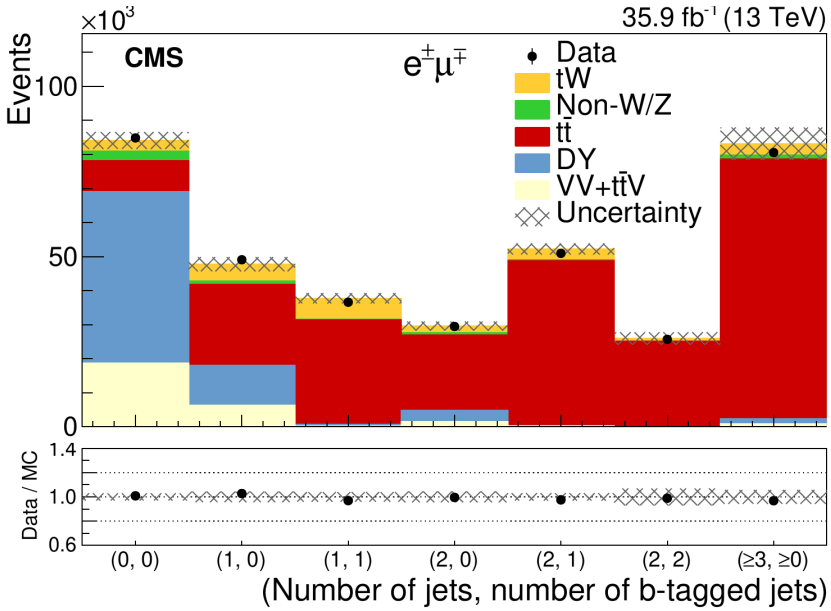


Figure 2.4: ([31]) Comparison between different regions with one electron and one muon. As can be seen in the graph, having one jet b–tagged gives the highest signal to background ratio.

Once the physical objects are reconstructed and identified for one event, we impose more requirements upon itself so that we are more confident that the event is actually one where its main physical process is that of our signal (i.e. tW). These selection criteria are the following, based essentially in restricting ourselves to the region with an electron and a muon and only one jet that must be b-tagged (see Fig. 2.4 for comparison with other regions).

- We require to have one electron and one muon with the identification that has been described in previous paragraphs (also taking into account electrons and muons coming from tau decays).
- The lepton with more p_T, denoted *leading* lepton, must have $p_T(\ell_1) > 25$ GeV.
- Both leptons must have opposite charge.

- To reduce contamination from low mass resonances processes the invariant mass of the lepton pair must fulfil $m(\ell_1, \ell_2) > 20$ GeV.
- Exactly one jet, that must be identified as a b–jet.

2.2.5 Uncertainties

During this section various of the **systematic** uncertainties that are taken into account have already been commented, especially for the **modelling** ones, having spoken of the DS, PS & ME scales, ME/PS matching, UE, colour reconnection models, ISR, FSR and the normalisation of the MC. There is one more to be added in this list:

PDF ([56]) The uncertainties in the parton distribution function (PDF) affect the theoretical prediction of the cross–section as wel as the prediction of the distributions. In this analysis these uncertainties are considered by obtaining distributions depending on variations in this uncertainties.

Apart from these modelling uncertainties, there is the subset inside the systematics of the **experimental** ones, which are the following.

B-tagging ([26]) The b-tagging method is validated with data, doing comparisons with events where a b-jet is clearly present and with others where although a jet is present, there is expected that it does not come from a B hadron. Afterwards, and to account for the difference, quantities called scale factors are derived comparing simulations and data. These quantities are values that affect the efficiency of b-tagging, helping the MC to agree with the data. However, as each scale factor has its own uncertainty, it must be propagated to the entire analysis by varying each one of them by its own uncertainty.

Mistagging ([26]) A similar procedure, although with different quantities and factors, is followed for the efficiency of detecting a jet that do not comes from a b quark (i.e. one that comes from an up, down, charm, or strange quark). We take into account its uncertainties.

Electron and muon identification efficiencies As well as is done with the b-tagging procedure, something similar is done with the electron and muon identification. Its efficiency is obtained by comparing data with simulations, obtaining them. Scale factors are also derived here that carry an uncertainty propagated afterwards in the analysis. They are obtained also in a 2D space in η and the p_T. For the case of muons, and following the recommendations of the CMS Muon POG, additional uncertainties of 1% for the cut–based ID and 0.5% for the relative isolation are added quadratically.

Trigger efficiency In analogy with the previous cases, the trigger efficiencies are derived by comparisons with accredited data and afterwards its uncertainties are propagated in our analysis.

Pile-up ([6]) The number of collisions per event is modelled depending on the cross section of the p–p collisions. The effect of these "extra" events has thus an uncertainty, which is taken varying this cross section $\pm 4.6\%$.

Jet energy scale (JES) and jet energy resolution (JER) ([28]) When jets are reconstructed, the value of each jet energy has obviously an uncertainty, which is divided in two parts: its scale (i.e. the uncertainty in traslations of the whole distribution) and its resolution (i.e. the wide of the distribution itself). Both

uncertainties have several sources, and both are sampled in a 2D space depending on η and the p_T of the jet.

Luminosity ([21]) The value of the integrated luminosity is essential for the correct normalisation. Its uncertainty is of a 2.5%.

The main idea to propagate this uncertainties is to obtain different distributions of the same observable (e.g. the transverse momentum of the jet), but with one feature varied (e.g. for the luminosity uncs., varying the nominal integrated luminosity a 2.5% up). Afterwards, the final uncertainties are taken as the difference between those varied distributions and the nominal ones, being the total uncertainty the quadratic sum of all those differences, although in some steps of this analysis a slight different approach is followed (see next subsection).

On the other hand, statistic uncertainties are considered both for the data as well as the Monte Carlo samples. There are other especially linked to the unfolding part that will be discussed in the next subsection.

2.2.6 Signal extraction and unfolding

Once the collider data and the Monte Carlo simulations have passed the selection criteria, they are up to the next step: the signal extraction. This is done, as explained, through a maximum likelihood (ML) fit. The distribution that we have chosen is the one of the BDT discriminant (implemented with the Toolkit for multivariate data analysis, `TMVA`, [63]), as it has a good discriminating power. The BDT has been trained to discriminate tW against $t\bar{t}$ with samples independent of those used in the rest of the analysis. To construct this BDT we use as input variables the following. They are obtained after applying the selection and identification criteria described in previous sections but using the simulated events of the mentioned samples.

- Number of jets that fulfil all the conditions already mentioned to be identified except from its p_T, that must be in $20\,\text{GeV} < p_t < 30\,\text{GeV}$. This is done to take into account jets with a relevant p_T (higher than 20 GeV) that are not identified as such in our analysis.
- Highest p_T of those jets. If there is no one, it is fixed to zero.
- Number of the jets with the energy range described in the first item that are also b-tagged.
- p_T of the vectorial sum of the four momenta of both leptons of the selected event, the jet and the E_T^{miss}, which is called the p_T of the system: p_T^{sys}
- Scalar sum of the p_T of the four momenta of both leptons, the jet and the E_T^{miss}, denoted as H_T.
- Ratio between p_T^{sys} and H_T.
- p_T of the jet.
- m_{sys}: the invariant mass of the combination of both leptons, the jet and the E_T^{miss}.
- Centrality (ratio between the p_T and $|\vec{p}|$) of the system of the jet and the two leptons.
- Ratio of the scalar sum of the p_T of both leptons over the H_T of the full system.
- Vectorial sum of the p_T of the jet and the leptons.

We have chosen the following variables to be unfolded and to present in this document preliminary results. We chose them to show relevant kinematic information of the physical objects of the events and to offer a global glimpse of the results. They are the following.

- $p_T(\ell_1)$: the transverse momentum of the lepton in the event with the highest one.
- $p_Z(\ell_1, \ell_2, j)$: the momentum in the Z axis (the one of the pipe of the LHC) of the system formed by the two leptons and the jet.
- $\Delta\varphi(\ell_1, \ell_2)$: the difference in the φ angle between the two leptons.
- $m(\ell_1, j)$: the invariant mass of the system of the jet (j) and the lepton with the highest p_T (ℓ_1).

With these distribution chosen to be unfolded (i.e. with the dependencies selected to measure the differential cross section), we can continue the procedure. For each observable of those we obtain the distribution of the BDT discriminant for each of the bins of the variable (bins that belong to the folded space) with a fixed number of bins for all of them. The limits of each bin are chosen so that the amount of the $t\bar{t}$ background group is equal in all of them: this way we expect to slightly enhance the differences in signal between all the bins (that must exist by construction of the BDT: from less signal in the negative values, to more in the positive). This is done for each of the bins of the variable folded space: once all the distributions are done, the fit is performed to all of them at the same time. At the end, we obtained different values of the POI (r) for each bin of the variable's folded space, that allowed us to extract the signal.

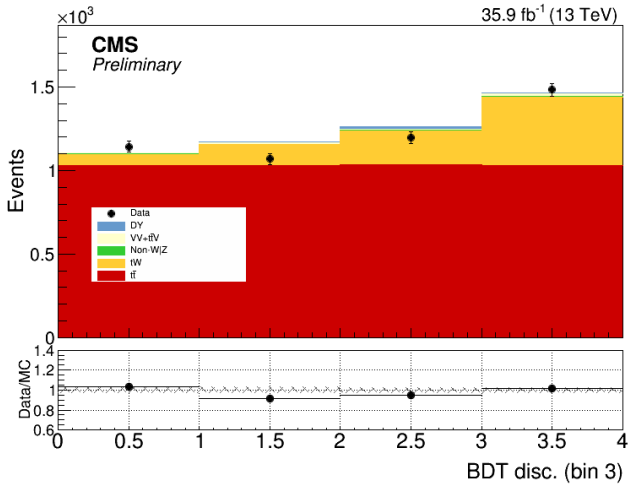


Figure 2.5: This is an example of the BDT distribution extracted from one particular bin of one variable. In this case, this is the third bin of the distribution of the transverse momentum of the lepton with the highest one: from this bin, the BDT distribution has been extracted as commented before, choosing four bins and an equal content of the $t\bar{t}$ process (the X axis shows arbitrary values: they are not the values of the discriminant). The uncertainties shown here are only of statistical nature.

Apart from the r parameters (one for each bin), all uncertainties are introduced in the fit, except from the modelling ones (listed in the previous subsection). The modelling uncertainties are propagated by taking its varied samples and performing the fit to them, as if it were the nominal ones. It is said that the modelling uncertainties are **externalised** (a dedicated fit is performed with the varied distribution to asses the uncertainty), whereas the others are parametrized with nuisance parameters of the fit.

During the development of this analysis we faced an obstacle, as we observed that when the number of bins of the BDT discriminant distribution was changed, large and random variations in the predominance of the various modelling uncertainties in the "post-fit" results appeared. A careful check in those BDT discriminant distributions convinced us that the problem was one of the amount of simulated events in the Monte Carlo samples, as the low quantity of them made that the post-fit values had unnatural large or small signal amount because of statistical fluctuations in the bins. This forced us to reduce largely the number of them, until we noticed that there were not any kind

of strange phenomena due to statistical fluctuations. The highest number of bins that fulfilled this conditions was four, and it was the chosen one. In Fig. 2.5 one of those distributions is shown.

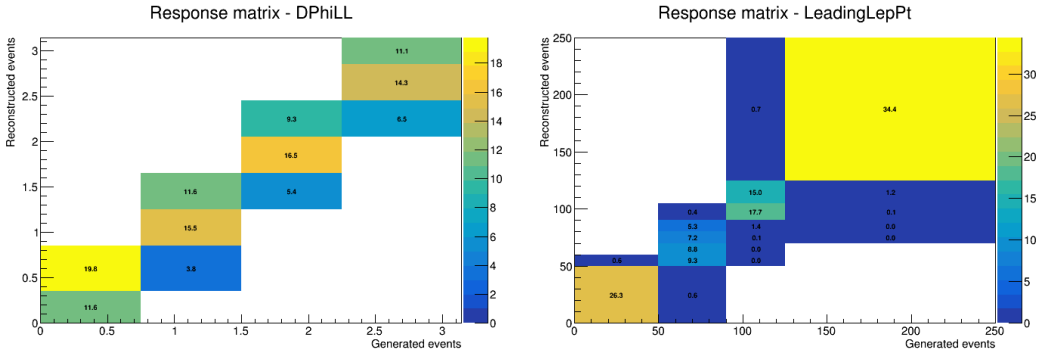
After the signal extraction procedure, we perform the unfolding on our variable. To do so, we use the `TUnfold` library ([60]), that implements the whole procedure in a comfortable way. The response matrices are calculated previously from dedicated samples as explained in the previous section, and can be seen in Fig. 2.6. As explained in the previous section, the binning in both folded and unfolded spaces was optimised by checking the purities and stabilities of each bin. In figure 2.7 the plots of both for the chosen distribution appear.

For the variables chosen to be unfolded, no regularisation was needed to perform the unfolding. This was estimated calculating the condition number[6] of each response matrix, that gave us low values (of order ~ 1). The numbers for the nominal response matrices are shown also in Fig. 2.6. We also performed scans in the L-curve using Tykhonov regularisation to confirm this, under the assumption that if no regularisation was needed, then the values of the τ parameter should be low, as they were for all the considered variables. The L-Curves of all the variables are plotted in Fig. 2.8, and the optimum (by the described criteria) tau parameter value is written in each of them.

The profiled uncertainties from the signal extraction are propagated to this new fit by importing the covariance matrix from that fit, whereas the externalised uncertainties (the modelling ones) are propagated by performing the unfolding for each of the variations of the distributions. There are also other uncertainties related to the unfolding procedure itself. In order to estimate the uncertainties in the response matrices themselves, we also calculate the response matrices when varying each distribution by the source of uncertainty (e.g. we also calculate a response matrix with the varied distributions of the jet energy scale). The uncertainties due to the statistics of the sample that is used to obtain the response matrix are taken into consideration in the procedure and added to the final uncertainty. To asses that our model does not bias us deeply, we compare the final results with the generation values from simulations of different software: in our case, we use tW samples of the `aMC@NLO` generator as well as our nominal `Powheg`. As no regularisation is imposed, no uncertainties concerning it are needed.

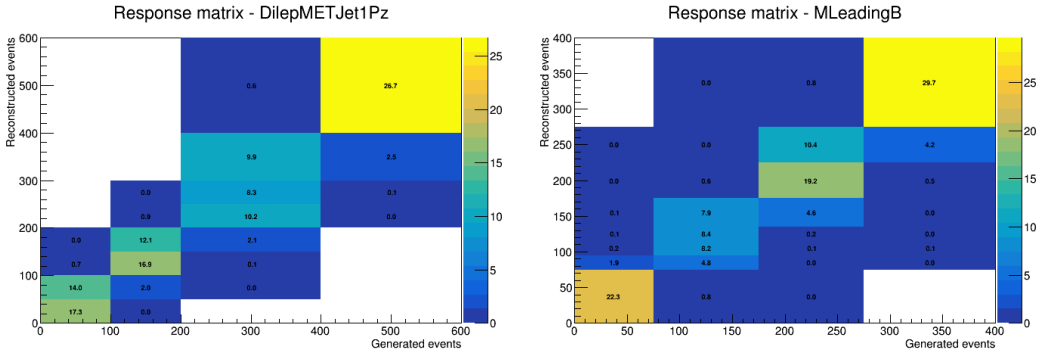
At the end, the uncertainties that came from the signal extraction, in addition of those of the unfolding procedure itself, are grouped and joint in the covariance matrix of the fit, which will be shown in the results plots as one group called "Fit". The other uncertainties, that were externalised (the modelling ones) are presented and propagated asymmetrically for each bin of the unfolded space of the variable. The total uncertainty in each of them is obtained by the quadratic sum of all of them.

[6]The condition number of a matrix is a mathematical concept that expresses how much a vector transformed from that matrix (e.g. $\vec{u} = A\vec{v}$) would be changed if slight variations in $\vec{v}$ were made. The definition depends on the norm used, and we will not go in deep here. Values of the condition number of $\sim 10^0$ for response matrices are considered low and it is expected that those unfolding could be made without regularisation, whereas values higher, $\sim 10^n$, might be in need of it.



(a) $\Delta\varphi(\ell_1, \ell_2)$. Condition number: 1.4.

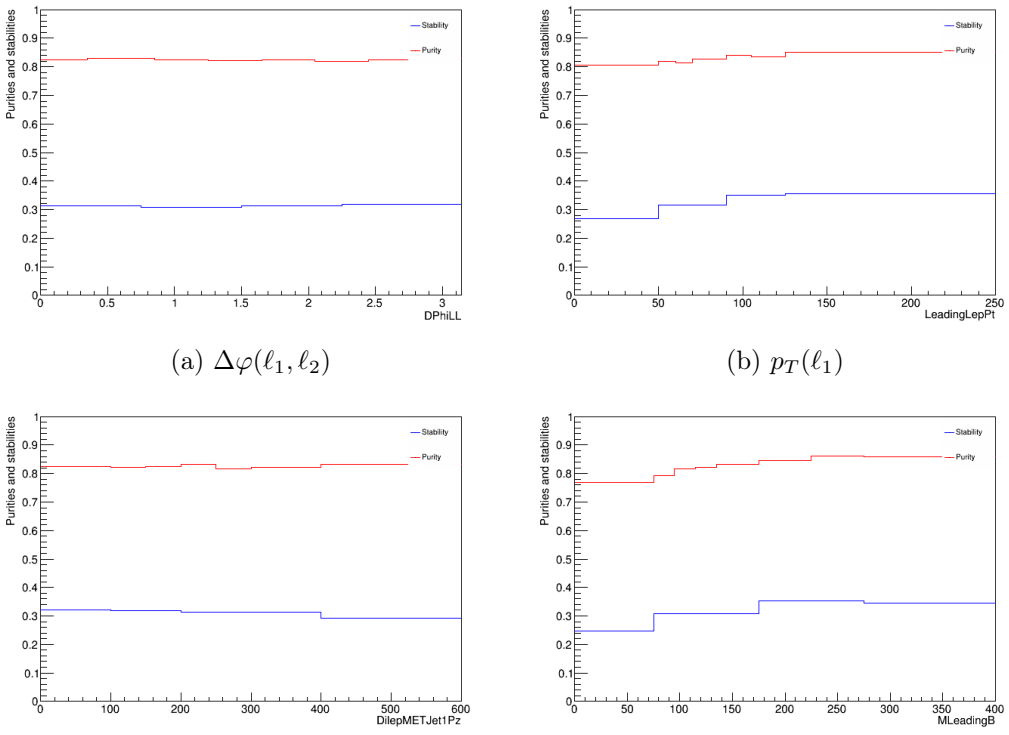
(b) $p_T(\ell_1)$. Condition number: 2.2.



(c) $p_Z(\ell_1, \ell_2, j)$. Condition number: 1.7.

(d) $m(\ell_1, j)$. Condition number: 2.1.

Figure 2.6: Response matrices of the variables chosen to be unfolded (of its nominal values): the reconstructed events axis is the folded space axis and the generated events axis, the unfolded. The condition numbers of each one are shown, and are all of them of order ~ 1.



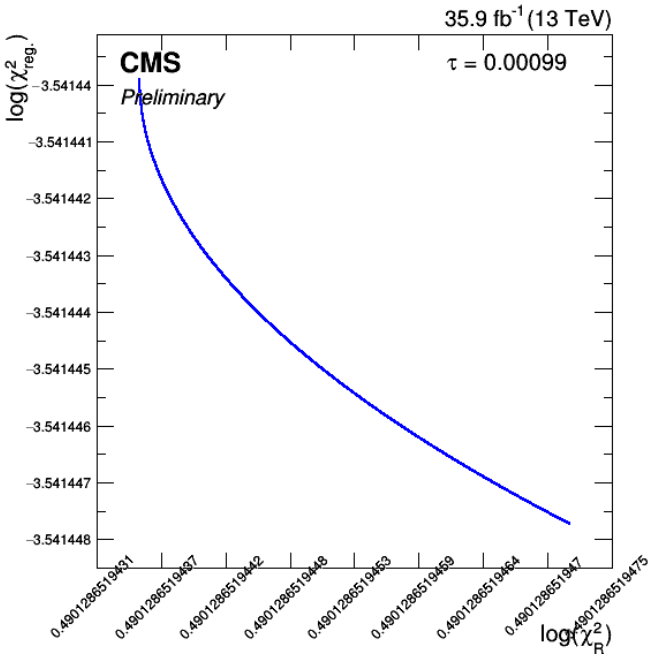
(a) $\Delta\varphi(\ell_1, \ell_2)$

(b) $p_T(\ell_1)$

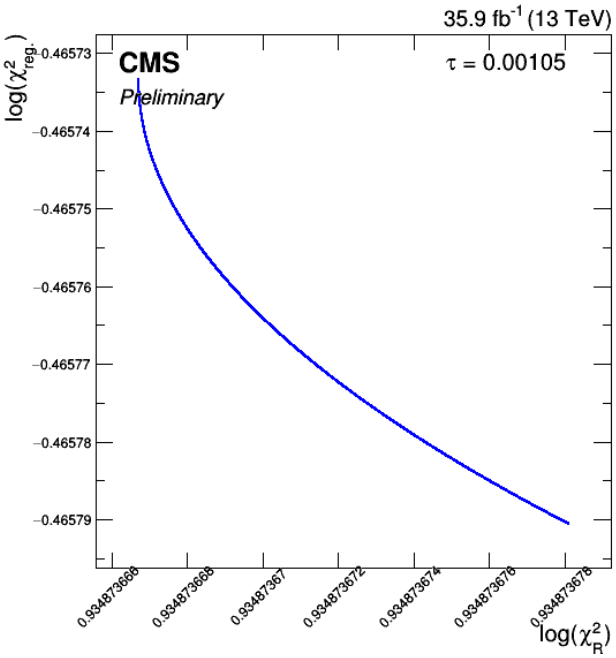
(c) $p_Z(\ell_1, \ell_2, j)$

(d) $m(\ell_1, j)$

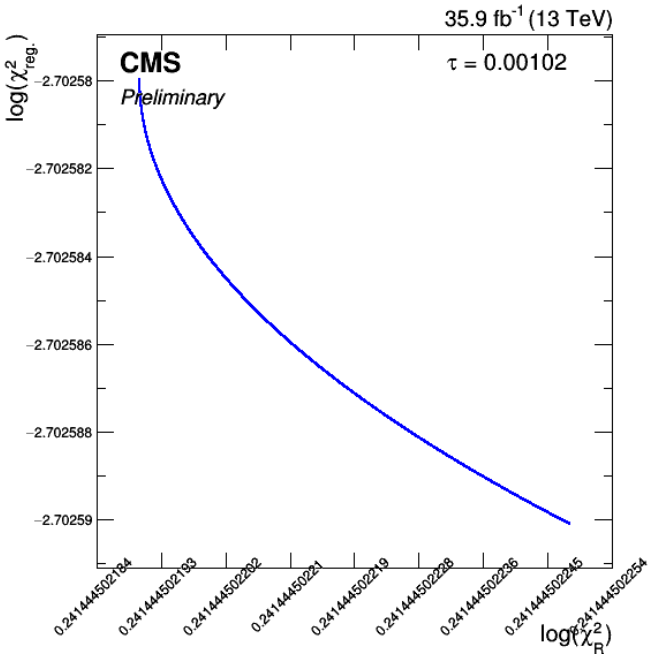
Figure 2.7: Graphs showing the purities and stabilities of each bin of the folded and unfolded (respectively) spaces for all the distributions chosen to be unfolded. It can be seen a relatively low stability, that is related with a low reconstruction efficiency for the events in our fiducial region.



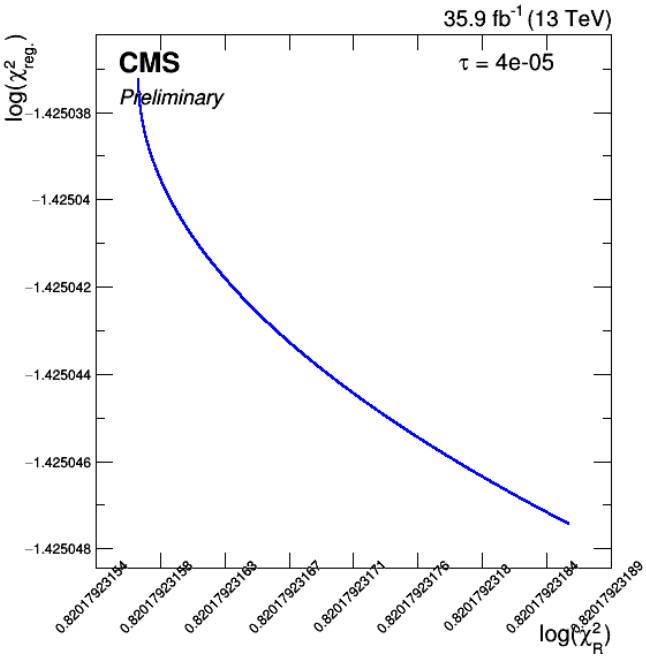
(a) $\Delta\varphi(\ell_1, \ell_2)$



(b) $p_T(\ell_1)$



(c) $p_Z(\ell_1, \ell_2, j)$



(d) $m(\ell_1, j)$

Figure 2.8: L-Curves of the variables chosen to be unfolded (of its nominal values). Te optimum value of the tau parameter (the one corresponding to the point of maximum curvature) is also presented.

3 Experimental results

Here follow the results of the preliminary analysis. The differential cross section has been measured depending on different physical features of the events, which are $p_T(\ell_1)$, $p_Z(\ell_1, \ell_2, j)$, $\Delta\varphi(\ell_1, \ell_2)$, $m(\ell_1, j)$ and $m(\ell_2, j)$, as we can recall from the previous section.

In the figure 3.1 the distributions with the chosen binning of each variable are shown. It can be seen an acceptable agreement between the observed data and the Monte Carlo simulations, with deviations appearing essentially in the tails of the distributions, and especially in the graph of the transverse momentum of the lepton with the highest p_T, which is a known effect of the simulations. The amount of observed data and simulated events after our selection are 36606 and 37736 events, respectively. The uncertainties that are shown are the total (i.e. the quadratic sum of all the uncertainties). It can be clearly seen that the main background is, by far, the pair production of top quarks.

The minimisation procedure and the maximum likelihood fit did not give any kind of numerical problem. In figure 3.2 the distributions of the variables after the signal extraction are shown, as well as a graph where a comparison between some of all the uncertainty sources is presented (in addition to the total one), as described in the figure caption. These uncertainties are presented asymmetrically in each bin. In addition, and as a check, the comparison with the information of two Monte Carlo generators is shown. It can be seen that the results fit with both generators within the uncertainties ($\pm 1\,\sigma$), or near them, with the exception of the p_T of the lepton with highest p_T In some distributions, such as $\Delta\varphi(\ell_1, \ell_2)$, a phenomenon of small statistical fluctuations can be seen in the relative uncertainties. Overall, although there is agreement with the Monte Carlo generators, the global uncertainties are moderately large. The main sources are in general common for all, and are some that we can expect: the final and/or initial state radiation or the grouped item "fit", which encompasses uncertainties such as the jet energy scale or jet energy resolution. This can be understood as our selection of events chooses strictly one jet that also must be b-tagged: thus, uncertainties related with the jets, that are usually higher than for the leptons, or the criteria for us to identify jets are expected to be predominant.

Regarding the last step, the unfolding, we found that no regularisation at all was needed as explained in the previous section. After the unfolding procedure, we obtain the results of figure 3.3. Again, the check after comparing with the particle level information from the Monte Carlo simulations is relatively good, with the exception of the transverse momentum of ℓ_1. The uncertainties, asymmetrically presented again, are overall large, even reaching more than the 100% of the nominal value in some cases. the main sources of the are more or less the same as for the post–fit results in the folded space, although a bit larger as expected after performing the unfolding.

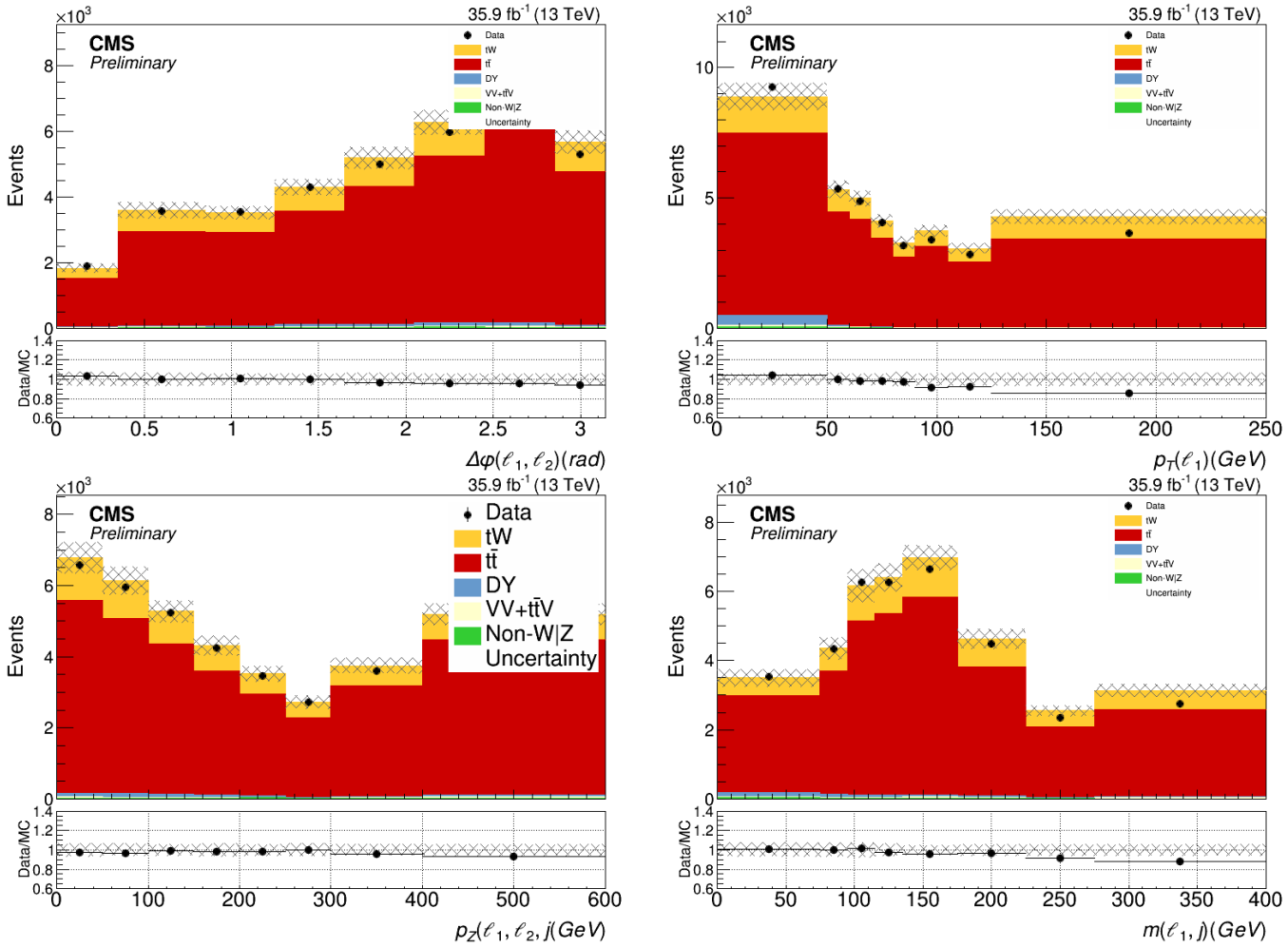


Figure 3.1: Distribution of the chosen variables to be unfolded. The agreement between data and Monte Carlo is fairly good, taking into account known deviations such as the one from $p_T(\ell_1)$. The binning shown is the selected for the unfolding (i.e. the folded space binning).

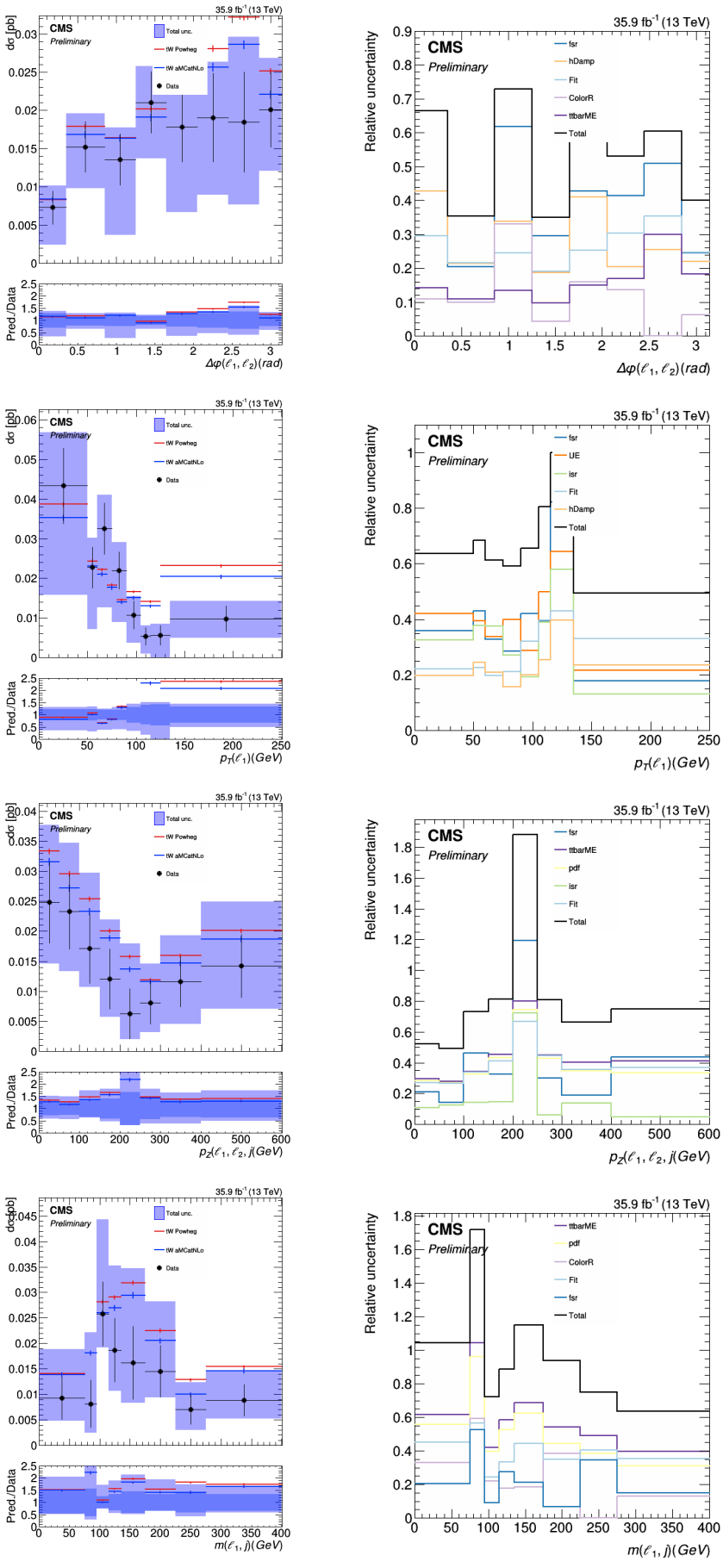


Figure 3.2: Distributions of the variables chosen to be unfolded after the maximum likelihood fit (the signal extraction), already expressed as a differential cross section, but in the folded space. For each variable the first plot shows the results themselves, with the total uncertainty. In the ratio plot, the group of uncertainties "fit" (that represent those uncertainties carried during the ML fit) are shown, as well as the total. The other graph represents the five source of uncertainty ordered by the maximum uncertainty (taking both variations) in all the bins of the variable, in addition of the total uncertainty. The uncertainties shown are asymmetrical.

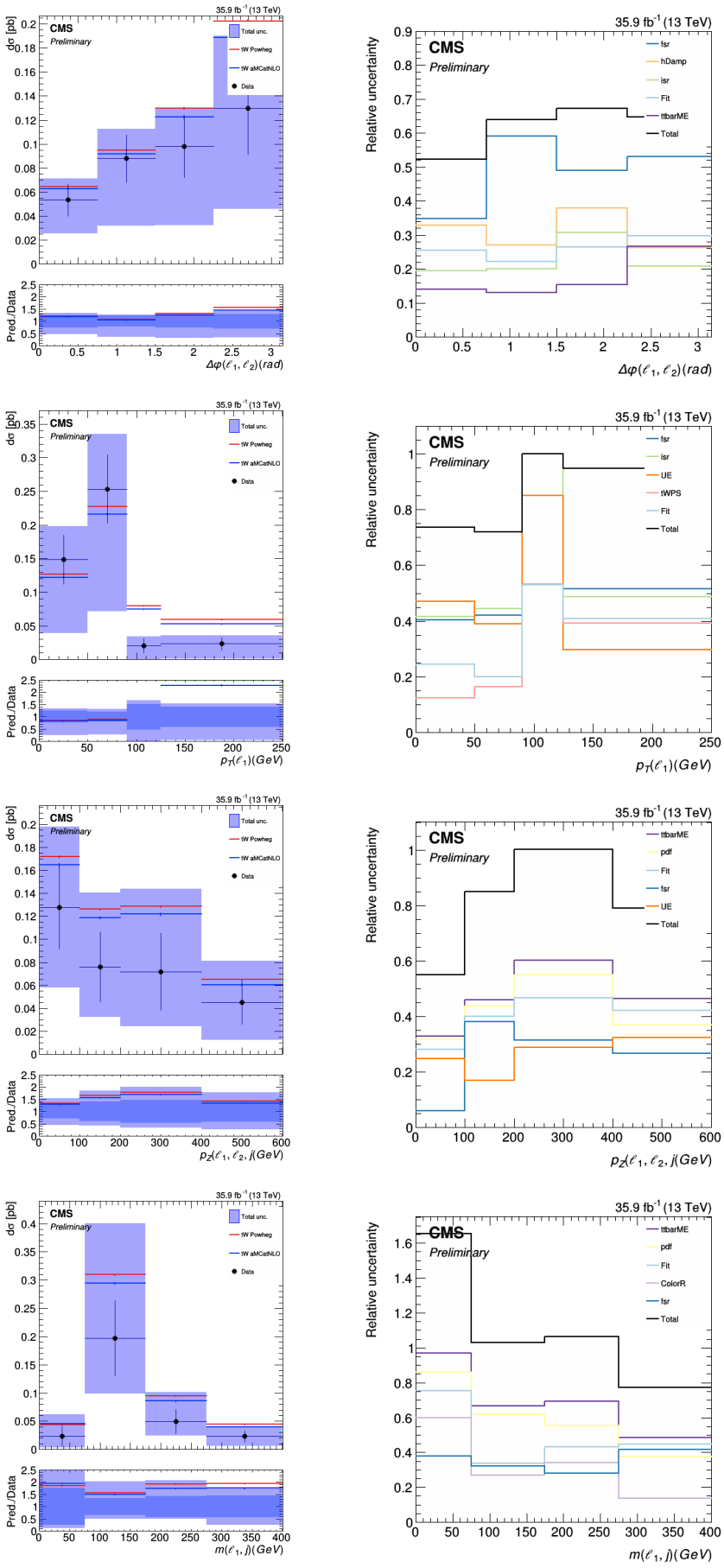


Figure 3.3: Final measurements of the differential cross section depending on the different chosen variables. As with the post–fit results, for each distribution the first plot shows the results themselves, with the total uncertainty. In the ratio plot, the group of uncertainties "fit" (that represent those uncertainties carried during the fit and also the unfolding) are shown, as well as the total. The other graph represents the same as in the post–fit results. The uncertainties shown are asymmetrical.

4 Conclusions

In this document we have presented the preliminary results of a measurement of the differential cross section of the physical process where a top quark is produced in association with a W boson, taking data of the CMS detector at LHC from 2016 at $\sqrt{s} = 13\,\text{TeV}$. This analysis is done in a region with one electron, one muon and one jet that must be b-tagged. The differential cross section has been measured depending on the transverse momentum of the lepton in the event with the highest one, the momentum in the Z axis (the one of the pipe of the LHC) of the system formed by the two leptons and the jet, the difference in the φ angle between the two leptons, the invariant mass of the system of the jet (j) and the lepton with the highest p_T (ℓ_1) and the invariant mass of the system of the jet (j) and the lepton with the lowest p_T (ℓ_2).

The complex workflow of the analysis, that to tackle the dominating $t\bar{t}$ background in the measurement region exploits the use of multivariate techniques as well as maximum likelihood fits to extract the signal, has been proven to perform correctly. In addition, the closure checks done by comparing with the particle level information from simulations, and other tests performed to it were successful. The final results in the selected variables are found to be in agreement with predictions, specifically with the two generators (`Powheg` and `aMC@NLO`) that we have considered. The main sources of uncertainties are those concerning the characteristics of the jets, as well as those related with the predominant background: the pair production of top quarks. The most relevant drawback of the current results is that the global uncertainties are large, as can be seen in Fig. 3.3. This also makes us unable to give preference both physic models used to compare with the results.

These achievements are a good foundation to improve the method trying to understand better our fiducial region and enhancing it if needed, and also interpret correctly our uncertainties.

5 Bibliography

[1] M. Aliev et al. '– HATHOR – HAdronic Top and Heavy quarks crOss section calculatoR'. In: *Comput.Phys.Commu 1046,2011* (2010). DOI: 10.1016/j.cpc.2010.12.040. arXiv: 1007.1327.

[2] Spyros Argyropoulos and Torbjörn Sjöstrand. 'Effects of color reconnection on $t\bar{t}$ final states at the LHC'. In: *Journal of High Energy Physics 11* (2014). DOI: 10.1007/JHEP11(2014)043. arXiv: 1407.6653.

[3] ATLAS Collaboration. 'Evidence for the associated production of a W boson and a top quark in ATLAS at $\sqrt{s} = 7\,\text{TeV}$'. In: *Phys.Lett. B716 (2012) 142-159* (2012). DOI: 10.1016/j.physletb.2012.08.011. arXiv: 1205.5764.

[4] ATLAS Collaboration. 'Measurement of differential cross-sections of a single top quark produced in association with a W boson at $\sqrt{s} = 13$ TeV with ATLAS'. In: *Eur. Phys. J. C 78 (2018) 186* (2017). DOI: 10.1140/epjc/s10052-018-5649-8. arXiv: 1712.01602.

[5] ATLAS Collaboration. 'Measurement of the cross-section for producing a W boson in association with a single top quark in pp collisions at $\sqrt{s} = 13$ TeV with ATLAS'. In: *JHEP 01 (2018) 63* (2016). DOI: 10.1007/JHEP01(2018)063. arXiv: 1612.07231.

[6] ATLAS Collaboration. 'Measurement of the Inelastic Proton-Proton Cross Section at $\sqrt{s} = 13$ TeV with the ATLAS Detector at the LHC'. In: *Phys. Rev. Lett. 117, 182002 (2016)* (2016). DOI: 10.1103/PhysRevLett.117.182002. arXiv: 1606.02625.

[7] ATLAS Collaboration. 'Measurement of the production cross-section of a single top quark in association with a W boson at 8 TeV with the ATLAS experiment'. In: *JHEP01(2016)064* (2015). DOI: 10.1007/JHEP01(2016)064. arXiv: 1510.03752.

[8] ATLAS Collaboration. 'Probing the quantum interference between singly and doubly resonant top-quark production in $p - p$ collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector'. In: (2018). arXiv: 1806.04667.

[9] Richard D. Ball et al. 'Parton distributions with LHC data'. In: *Nuclear Physics B* (2012). DOI: 10.1016/j.nuclphysb.2012.10.003. arXiv: 1207.1303.

[10] Michiel Botje et al. 'The PDF4LHC Working Group Interim Recommendations'. In: (2011). arXiv: 1101.0538.

[11] Qing-Hong Cao, Jose Wudka and C. -P. Yuan. 'Search for New Physics via Single Top Production at the LHC'. In: *Phys.Lett.B658:50-56,2007* (2007). DOI: 10.1016/j.physletb.2007.10.057. arXiv: 0704.2809.

[12] CDF Collaboration. 'Observation of Top Quark Production in $p - \bar{p}$ Collisions'. In: *Phys. Rev. Lett. 74* (1995), pp. 2626–2631. DOI: 10.1103/PhysRevLett.74.2626. arXiv: hep-ex/9503002.

[13] CDF Collaboration and D0 Collaboration. 'Observation of s-channel production of single top quarks at the Tevatron'. In: *Phys. Rev. Lett. 112, 231803 (2014)* (2014). DOI: 10.1103/PhysRevLett.112.231803. arXiv: 1402.5126.

[14] CERN. *"Taking a closer look at LHC"*. URL: http://lhc-closer.es/taking_a_closer_look_at_lhc/1.home (visited on 31st May 2018).

[15] CERN. *CERN Council Gives Go-ahead for Large Hadron Collider*. 1994. URL: https://press.cern/press-releases/1994/12/cern-council-gives-go-ahead-large-hadron-collider (visited on 29th May 2018).

[16] CERN. *Schema of a LHC multipole*. URL: http://www.lhc-closer.es/webapp/files/1435504123_b887b3b5c6aab9b0259320ea21935bbd.png (visited on 30th May 2018).

[17] CERN and CMS Collaboration. *Diagram of the CMS detector*. URL: https://cms-docdb.cern.ch/cgi-bin/PublicDocDB/RetrieveFile?docid=11514&version=1&filename=cms_120918_03.png (visited on 31st May 2018).

[18] CERN and CMS Collaboration. *Jet diagram production*. URL: http://cms.web.cern.ch/sites/cms.web.cern.ch/files/styles/large/public/field/image/Sketch_PartonParticleCaloJet.png?itok=SbSUp7_9 (visited on 1st June 2018).

[19] Jesper R. Christiansen and Peter Z. Skands. 'String Formation Beyond Leading Colour'. In: *Journal of High Energy Physics 8* (2015). DOI: 10.1007/JHEP08(2015)003. arXiv: 1505.01681.

[20] CMS Collaboration. '"The CMS Experiment at the CERN LHC"'. In: (2008). DOI: 10.1088/1748-0221/3/08/S080048.

[21] CMS Collaboration. 'CMS Luminosity Measurements at 13 TeV - Winter 2017 update'. In: *CMS Physics Analysis Summary CMS-PAS-LUMI-17-001* (2017).

[22] CMS Collaboration. *CMS luminosity measurements for the 2016 data taking period*. URL: https://cms-service-lumi.web.cern.ch/cms-service-lumi/publicplots/int_lumi_per_day_cumulative_pp_2016.pdf (visited on 1st Aug. 2018).

[23] CMS Collaboration. 'Evidence for associated production of a single top quark and W boson in $p-p$ collisions at $\sqrt{s} = 7\,\text{TeV}$'. In: *Phys. Rev. Lett. 110 (2013) 022003* (2012). DOI: 10.1103/PhysRevLett.110.022003. arXiv: 1209.3489.

[24] CMS Collaboration. *GitHub repository "CMGTools"*. URL: https://github.com/CERN-PH-CMG/cmg-cmssw (visited on 25th June 2018).

[25] CMS Collaboration. 'Identification of b quark jets at the CMS Experiment in the LHC Run 2'. In: *CERN Document Server, CMS-PAS-BTV-15-001* (2016).

[26] CMS Collaboration. 'Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV'. In: *JINST 13 (2018) P05011* (2017). DOI: 10.1088/1748-0221/13/05/P05011. arXiv: 1712.07158.

[27] CMS Collaboration. 'Investigations of the impact of the parton shower tuning in Pythia 8 in the modelling of $t\bar{t}$ at $\sqrt{s} = 8$ TeV and 13 TeV'. In: *CMS Physics Analysis Summary CMS-PAS-TOP-16-021* (2016).

[28] CMS Collaboration. 'Jet algorithms performance in 13 TeV data'. In: *CMS-PAS-JME-16-003* (2017).

[29] CMS Collaboration. 'Measurement of the t-channel single-top quark production cross section at 13 TeV with the CMS detector'. In: (2016). arXiv: 1611.08443.

[30] CMS Collaboration. 'Measurement of the $t\bar{t}$ production cross section using events in the $e\mu$ final state in $p-p$ collisions at $\sqrt{s} = 13\,\text{TeV}$'. In: *EPJC* (2016). DOI: 10.1140/epjc/s10052-017-4718-8. arXiv: 1611.04040.

[31] CMS Collaboration. 'Measurement of the production cross section for single top quarks in association with W bosons in proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$'. In: (2018). arXiv: 1805.07399.

[32] CMS Collaboration. *Object definitions for top quark analyses at the particle level.* Tech. rep. CMS-NOTE-2017-004. CERN-CMS-NOTE-2017-004. Geneva: CERN, June 2017. URL: https://cds.cern.ch/record/2267573.

[33] CMS Collaboration. 'Observation of the associated production of a single top quark and a W boson in $p-p$ collisions at $\sqrt{s} = 8\,\text{TeV}$'. In: *Phys. Rev. Lett. 112 (2014) 231802* (2014). DOI: 10.1103/PhysRevLett.112.231802. arXiv: 1401.2942.

[34] CMS Collaboration. 'Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET'. In: *CMS Physics Analysis Summaries* (2009).

[35] CMS Collaboration. 'Reconstruction and identification of tau lepton decays to hadrons and tau neutrino at CMS'. In: *JINST 11 (2016) P01019* (2015). DOI: 10.1088/1748-0221/11/01/P01019. arXiv: 1510.07488.

[36] CMS Collaboration. *Summary of CMS cross section measurements.* URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsCombined (visited on 5th May 2018).

[37] CMS Collaboration. *The CMS experiment at the CERN LHC.* URL: http://cms.web.cern.ch/news/what-cms (visited on 31st May 2018).

[38] Javier Cuevas Maestro et al. 'Differential cross-section in single top tW-channel'. In: *CMS internal note* (2018). DOI: CMSAN-2018/173.

[39] Michal Czakon and Alexander Mitov. 'Top++: a program for the calculation of the top-pair cross-section at hadron colliders'. In: *Computer Physics Communications 185 (2014) 2930* (2011). DOI: 10.1016/j.cpc.2014.06.021. arXiv: 1112.5675.

[40] D0 Collaboration. 'Search for High Mass Top Quark Production in $p - \bar{p}$ Collisions at $\sqrt{s} = 1.8\,\text{TeV}$'. In: *Phys. Rev. Lett. 74* (1994), pp. 2422–2426. DOI: 10.1103/PhysRevLett.74.2422. arXiv: hep-ex/9411001.

[41] J. Delgado et al. 'PROOF Analysis Framework (PAF)'. In: *J. Phys.: Conf. Ser.* (2015). DOI: 664032009.

[42] F. Beaudette (CMS Collaboration). 'The CMS Particle Flow Algorithm'. In: (2013). arXiv: hep-ex/1401.8155.

[43] S. Frixione et al. 'Single-top hadroproduction in association with a W boson'. In: *JHEP0807:029,2008* (2008). DOI: 10.1088/1126-6708/2008/07/029. arXiv: 0805.3067.

[44] G. Apollinari, O. Brüning, L. Rossi (CERN). *High Luminosity LHC Project Description.* URL: https://cds.cern.ch/record/1974419/?ln=es (visited on 28th May 2018).

[45] Grupo Experimental de Altas Energías de la Universidad de Oviedo. *"AnalysisPAF" repository.* URL: https://github.com/Oviedo-PAF/AnalysisPAF (visited on 29th June 2018).

[46] Grupo Experimental de Altas Energías de la Universidad de Oviedo. *PAF webpage.* URL: http://www.hep.uniovi.es/PAF/ (visited on 29th June 2018).

[47] Christian Hansen. 'Analysis of discrete ill–posed problems by means of the L-Curve'. In: *SIAm Review Vol.34, No.4, pp. 561-580* (1992).

[48] Jean-Luc Caron, CERN. *Magnetic field induced by the LHC dipole's superconducting coils.* URL: https://cds.cern.ch/record/841511/?ln=es (visited on 30th May 2018).

[49] P. Kant et al. 'HATHOR for single top-quark production: Updated predictions and uncertainty estimates for single top-quark production in hadronic collisions'. In: *Comput.Phys.Commun. 191 (2015) 74-89* (2014). DOI: 10.1016/j.cpc.2015.02.001. arXiv: 1406.4403.

[50] L. R. Evans, P. Bryant. *LHC Machine*. URL: https://cds.cern.ch/record/1129806/?ln=es (visited on 31st May 2018).

[51] Hung-Liang Lai et al. 'New parton distributions for collider physics'. In: *Phys. Rev. D* (2010). DOI: 10.1103/PhysRevD.82.074024. arXiv: 1007.2241.

[52] *LHC map over the surroundings of Geneva*. URL: https://alexeinstein.files.wordpress.com/2014/10/lhc.jpg (visited on 30th May 2018).

[53] G. Soyez M. Cacciari G. P. Salam. 'The anti-k_t jet clustering algorithm'. In: (2008). arXiv: hep-ph/0802.1189.

[54] A. D. Martin et al. 'Parton distributions for the LHC'. In: *Eur. Phys. J. C* (2009). DOI: 10.1140/epjc/s10052-009-1072-5. arXiv: 0901.0002.

[55] A. D. Martin et al. 'Uncertainties on α_S in global PDF analyses and implications for predicted hadronic cross sections'. In: *Eur.Phys.J.C64:653-680,2009* (2009). DOI: 10.1140/epjc/s10052-009-1164-2. arXiv: 0905.3531.

[56] NNPDF Collaboration et al. 'Parton distributions for the LHC Run II'. In: *Journal of High Energy Physics 4* (2014). DOI: 10.1007/JHEP04(2015)040. arXiv: 1410.8849.

[57] Particle Data Group, C. Patrignani et al. 'Review of Particle Physics'. In: *Chin. Phys. C* 40 (2017), p. 100001.

[58] G. Petrucciani, A. Rizzi and C. Vuosalo. 'Mini-AOD: A New Analysis Data Format for CMS'. In: (2017). arXiv: 1702.04685.

[59] S. Cittolin, A. Rácz, S. Paris (CERN, CMS Collaboration). *CMS The TriDAS Project : Technical Design Report, Volume 2: Data Acquisition and High-Level Trigger*. URL: https://cdsweb.cern.ch/record/578006? (visited on 28th May 2018).

[60] Stefan Schmitt. 'TUnfold: an algorithm for correcting migration effects in high energy physics'. In: (2012). DOI: 10.1088/1748-0221/7/10/T10003. arXiv: 1205.6201.

[61] Peter Skands, Stefano Carrazza and Juan Rojo. 'Tuning PYTHIA 8.1: the Monash 2013 Tune'. In: *EPJC* (2014). DOI: 10.1140/epjc/s10052-014-3024-y. arXiv: 1404.5630.

[62] Tim M. P. Tait and C. -P. Yuan. 'Single Top Production as a Window to Physics Beyond the Standard Model'. In: *Phys.Rev.D63:014018,2000* (2000). DOI: 10.1103/PhysRevD.63.014018. arXiv: hep-ph/0007298.

[63] Jan Therhaag and TMVA Core Developer Team. 'TMVA - Toolkit for multivariate data analysis'. In: *AIP Conference Proceedings 1504, 1013* (2012). DOI: 10.1063/1.4771869.

[64] Unknown. *Diagram of a CMS section and of the pass of particles through itself*. URL: http://www.particlecentral.com/images/cms_slice.jpg (visited on 31st May 2018).

[65] Wikimedia. *B-jet diagram production*. URL: https://upload.wikimedia.org/wikipedia/commons/b/ba/B-tagging_diagram.png (visited on 17th July 2018).

[66] Wikimedia Commons. *Summary of interactions between particles described by the Standard Model.* URL: https://en.wikipedia.org/wiki/Standard_Model#/media/File:Standard_Model_of_Elementary_Particles.svg (visited on 23rd Apr. 2018).

[67] Wikimedia Commons. *The Standard Model of elementary particles (more schematic depiction) with the three generations of matter, gauge bosons in the fourth column and the Higgs boson in the fifth.* URL: https://en.wikipedia.org/wiki/Standard_Model#/media/File:Standard_Model_of_Elementary_Particles.svg (visited on 22nd Apr. 2018).