Università degli Studi di Milano - Bicocca

FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI
Corso di Laurea Magistrale in Fisica

Search for tZq in dilepton final states with Machine Learning techniques

Master thesis in cooperation with the Deutsches Elektronen-Synchrotron (DESY)

Candidato:
Federica Colombina
Matricola 802301

Relatore:
Prof. Pietro Govoni

Correlatore:
Dr.ssa Abideh Jafari

Anno Accademico 2019-2020

Abstract

In this work, events in which a single top quark is produced in association with a Z-boson are investigated. Particular focus is on final states in which two leptons originate from the Z boson and the top quark decays hadronically. The analysis uses data collected with the Compact Muon Solenoid (CMS) detector between 2016 and 2018, corresponding to an integrated luminosity of 137 fb^{-1}; data are produced from proton-proton collisions at the Large Hadron Collider (LHC), with a center of mass energy equal to 13 TeV. The discrimination between signal and backgrounds is optimized by testing different methods for the event reconstruction and with the implementation of a Deep Neural Network (DNN). In the last part of the analysis, a statistical fit is performed to the DNN output for the signal extraction, reaching an expected significance of the tZq signal over background of 2.4σ.

Contents

1 CMS detector at the LHC **6**
- 1.1 The accelerator complex at CERN . . . 6
- 1.2 Compact Muon Solenoid, CMS . . . 8
  - 1.2.1 Silicon tracker . . . 8
  - 1.2.2 ECAL and HCAL . . . 9
  - 1.2.3 Muon chambers . . . 10
  - 1.2.4 Triggering and data acquisition . . . 10
  - 1.2.5 Object and event reconstruction . . . 11

2 Preliminary notions on top quark physics **14**
- 2.1 The Standard Model . . . 14
- 2.2 Top quark in the SM . . . 15
- 2.3 tZq production and search . . . 17

3 Multivariate Analysis, MVA **20**
- 3.1 Introduction . . . 20
- 3.2 Neural Networks . . . 21
- 3.3 Learning process . . . 21
- 3.4 Hyperparameter optimization . . . 24
- 3.5 Applications in high energy physics . . . 25

4 Analysis of tZq in the dilepton final state **28**
- 4.1 MC samples and pre-selection . . . 28
- 4.2 Event reconstruction . . . 32
  - 4.2.1 Hadronic top quark reconstruction . . . 32
  - 4.2.2 Recoil jet reconstruction . . . 33
- 4.3 Signal-background discrimination with a DNN . . . 35
  - 4.3.1 Input variables . . . 35
  - 4.3.2 Hyperparameter optimization . . . 37
- 4.4 Data-MC agreement . . . 40
- 4.5 Statistical fit of the DNN output . . . 41

5 Conclusions **46**

Introduction

This work investigates the Z-boson associated single top quark production, tZq events, with exactly two leptons in the final state. The analysis uses data collected with the Compact Muon Solenoid (CMS) detector from 2016 to 2018, corresponding to an integrated luminosity of 137 fb^{-1}; data are produced from proton-proton collisions at the Large Hadron Collider (LHC), with a center of mass energy equal to 13 TeV. The two leptons are assumed to originate from the Z boson, and the top quark decays hadronically. So far, the process has not been observed in this final state, since the identification of signal events is challenging due to the overwhelming $t\bar{t}$ and Drell-Yan (DY) backgrounds. The Z-boson associated top quark productions, pair ($t\bar{t}Z$) or single (tZq), provide a unique opportunity to directly measure the top-Z couplings. Existing cross section measurements of $t\bar{t}Z$ and tZq consider events with at least three isolated leptons, in which two of them originate from the Z boson decay and one lepton appears in the decay of a top quark. The study the dilepton channel is supposed to bring complementary information, exploring a different phase space of the data. This channel has more statistics compared with the one having three leptons in the final state, meaning that it can bring additional sensitivity to future combinations of all measured channels, which would help to improve the understanding of top-Z coupling and production process. In the current tZq cross section measurement, $t\bar{t}Z$ dilepton contributes as an irreducible background with a comparable cross section to that of signal.

The tZq event configuration is exploited to achieve a good separation between signal and the irreducible backgrounds. Different methods have been tested for the event reconstruction, focusing on the top quark and the recoil jet, in order to achieve the highest reconstrucition efficiency. The outcome of the event reconstruction is used, in addition to other kinematic variables, in a Deep Neural Network (DNN) to further discriminate signal from backgrounds: the use of jet kinematics that are associated with the top quark and the recoil jet improves the performance upon cases where the event reconstruction information is not used.

The last part of the analysis aims to perform a statistical fit to the DNN output for the signal extraction. The fit is performed simultaneously on three DNN output distributions. In order to have orthogonal categories, only events for which the node value is maximum among all nodes have been included in each distribution. Contributions from

DY and $t\bar{t}$ are treated as freely floating rate parameters, in order to be determined from data without any prior constraint.

A description of the CERN (European Organization for Nuclear Research) accelerator complex and the CMS detector is illustrated in Chapter 1. Theoretical notions on top quark physics and Machine Learning (ML) algorithms are reported in Chapter 2 and 3 respectively. The different steps of the analysis are all illustrated in Chapter 4, while Chapter 5 contains a summary of the conclusions that have been achieved in this work.

Chapter 1

CMS detector at the LHC

Particle accelerators make use of electromagnetic fields to accelerate charged particles to velocities close to the speed of light. In a circular collider, particles circulate in opposite directions and are forced to head-on collisions, in which a high amount of energy is released. As a result, new particles are produced and their properties will be measured by surrounding detectors. The largest accelerator currently operating is the Large Hadron Collider (LHC) [1, 2], located at the European Organization for Nuclear Research (CERN) on the border between Switzerland and France.

1.1 The accelerator complex at CERN

The CERN accelerator complex consists of a chain of particle accelerators that increase the energy of particle beams before bringing them into the LHC for collisions. Hydrogen atoms coming from a bottle of compressed gas are subjected to an electric field, in order to remove the electrons from the atoms and produce protons, which are injected into the first accelerator of the chain, Linear accelerator 2 (Linac 2): here protons reach the energy of 50 MeV. The beams are then delivered to a succession of three synchrotons, where the protons are accelerated to 450 GeV (details are reported in Table 1.1).

Table 1.1: *Synchrotron names and energies achieved; the synchrotrons are listed in the same order they occur in the accelerator complex.*

Synchrotron name	Energy achieved
Proton Synchrotron Booster (PSB)	1.4 GeV
Proton Synchrotron (PS)	25 GeV
Super Proton Synchrotron (SPS)	450 GeV
Large Hadron Collider (LHC)	6.5 TeV

The protons are finally injected into (LHC): two particle beams travel in separate beam pipes, circulating in opposite directions, and are held in cyclic orbits by a magnetic

field. The instantaneous luminosity is defined as the coefficient of proportionality between the number of events occurred in a certain time ($\frac{dN}{dt}$) and the cross section σ:

$$\frac{dN}{dt} = \sigma\mathcal{L}. \tag{1.1}$$

As a result, the total number of event is proportional to the time-integrated luminosity:

$$N = \sigma L \tag{1.2}$$

where $L = \int \mathcal{L}dt$ is the integrated luminosity, corresponding to the total amount of collisions produced.

Detectors are located at four collision points around the LHC ring: ALICE (A Large Ion Collider Experiment), LHCb (Large Hadron Collider beauty), CMS (Compact Muon Solenoid) and ATLAS (A Toroidal LHC ApparatuS). At these points, collisions generate a large amount of energy, enough to produce new and heavy particles. Collisions occur once every 25 nanoseconds, with a total centre-of-mass energy of 13 TeV (in the years 2015-2018), and up to 40 simultaneous interactions take place at each crossing. The CERN accelerator complex is illustrated in Figure 1.1.

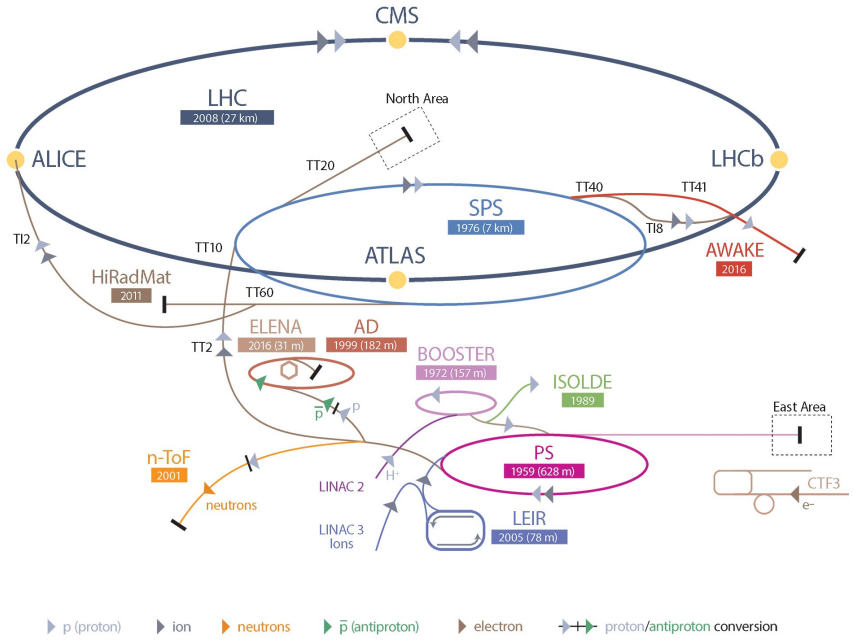


Figure 1.1: *CERN accelerator complex.*

1.2 Compact Muon Solenoid, CMS

The stable decay products of the particles produced at the interaction vertex are detected by means of experimental apparatuses built around the interaction points. In this way, the particle momenta and energies are measured and the collision can be reconstructed. The Compact Muon Solenoid (CMS) [3], schematically shown in Figure 1.2, is one of the detectors installed along the LHC ring.

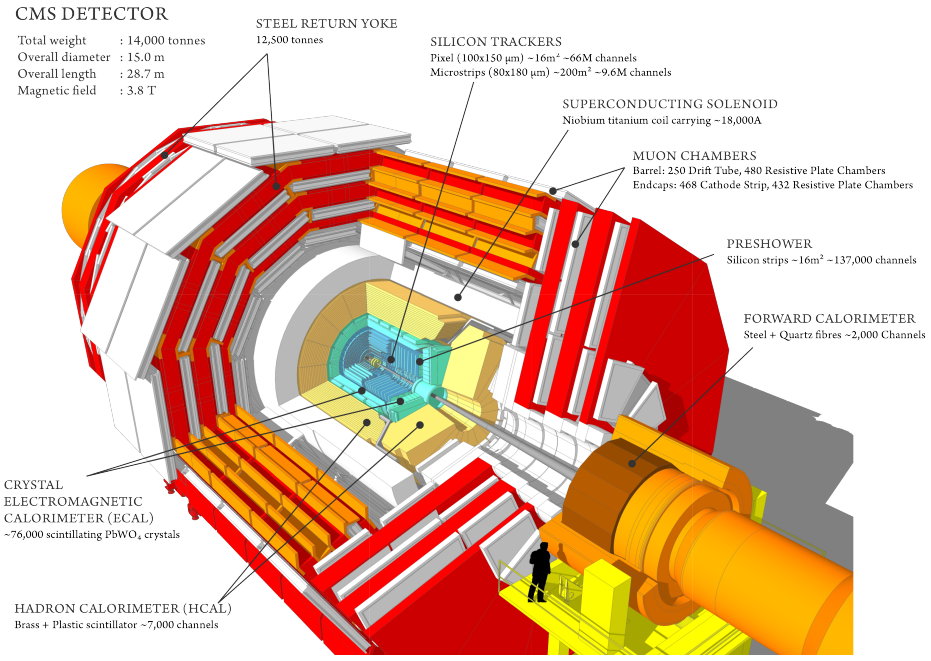


Figure 1.2: *CMS detector and its components, arranged as several concentric cylinders.*

The main feature of CMS is a solenoid magnet, made of superconducting wires that are able to produce a 3.8 Tesla magnetic field which bends charged particles trajectories. The core of the detector consists of a silicon tracker. The two CMS calorimeters, ECAL (Electromagnetic Calorimeter) and HCAL (Hadron Calorimeter), are located inside the magnet as well, surrounding the tracker. In the end, muons are detected in the outer layer of the experiment with muon chambers.

1.2.1 Silicon tracker

The tracker [4] is able to reconstruct the path of the particles by saving their positions at a limited number of points; it covers a region of $|\eta| < 2.5$ where the pseudorapidity, η, is given by

$$\eta = -\ln \tan \frac{\theta}{2} \tag{1.3}$$

and θ is the angle between the particle three-momentum and the direction of the beam. A precise measurement of trajectories of charged particles is essential to identify primary and secondary vertices of the collisions. Once the path has been recorded, its curvature gives a measure of the momentum, while the bending direction identifies the charge of the particles.

A large amount of paritcles is produced at each bunch crossing, about 1000 on average: a high granularity is needed to distinguish them from each other and to reconstruct the corresponding trajectories. Furthermore, the response time must be really short, due to the high frequence in which collisions occur and the detecotr components should be able to resist the high amount of radiation in the core of the detector. These requirements lead to a design entirely based on silicon detector technology: three barrel layers of pixel detectors surround the beam axis in the innermost part of the tracker, from a radius of 4.4 cm to 10.2 cm, and ten layers of silicon micro-strip detectors are placed outward up to a radius of 1.1 m; finally, at the endcaps, some disks of strip detectors are installed. As charged particles travel through these modules, electron-hole pairs are produced, which in turn generate an electric current which il be amplified and detected in the read-out electronics.

1.2.2 ECAL and HCAL

The Electromagnetic Calorimeter (ECAL) [5] measures the energy of electrons, positrons and photons; the other particles that interact electromagnetically may leave tracks, but they are not fully absorbed. The ECAL surrounds the silicon tracker and covers the region $|\eta| < 3$. It is made of crystals of lead tungstate $PbWO_4$, which is both absorber and scintillator material, and produces light in proportion to the primary electron or photon energies. When electromagnetically interacting particles start to interact with the calorimeter, they emit photons through bremsstrahlung, which in turn create electron-positron pairs (pair production) resulting in an electromagnetic shower. Atoms in the calorimeter are then excited and subsequently emit photons; this signal is amplified with photomultipliers in order to measure the energy that has been deposited in the calorimeter.

Hadrons are stopped in the Hadron Calorimeter (HCAL) [6], where their energies are estimated from scintillation light as well. Here absorbing and scintillating layers alternate, allowing to measure forward jets up to $|\eta| \simeq 5.2$. When hadronic particles pass through the absorber materials, interactions occur and many secondary particles are produced; these can in turn interact in the following absorber layers resulting in a shower of particles. As the shower develops, particle pass through the scintillating layers and their energy is measured.

1.2.3 Muon chambers

Since muons can penetrate several metres of absorber materials (like iron or barrel) without interacting, they are not stopped by any of CMS's calorimeters and are detected in the outermost part of the experiment, the muon chambers [7]. These detectors cover a pseudorapidity region of $|\eta| < 2.4$ without any gaps; only neutrinos, which cannot be directly detected by CMS, and muons reach this part of the detector. Muon identification and position determination is done with gaseous detectors: as they pass through the gas, atoms are ionized and resulting electrons and ions produce an electric current; combining these measurements with the ones from the tracker allows to reconstruct the path of the particle. This gives a measurement of particle's momentum, since it can be determined by measuring the curvature of the track due to the magnetic field.

1.2.4 Triggering and data acquisition

Proton-proton collisions generate approximately 70 terabytes of data every second and the current existing technology does not allow to store such a high amount of information for offline analysis: it exceeds the data-taking bandwidth limits. As a consequence, the acquisition rate must be reduced. For this purpose, CMS developed a two-level trigger system [8] that applies selection criteria in order to identify those events which are of possible physic interest.

The Level-1 (L1) trigger processes information from calorimeters and muon chambers within 3.8 μs after a collision, reducing the event rate down to 100 kHz. A low spatial and energetic granularity is used in this first part of the selection, since decisions must be taken in a short time. The selected events are then passed to the high-level-trigger (HLT), which restricts the output rate between 1 and 2 kHz. HLT algorithms have access to data from the whole CMS detector, with full granularity and resolution: this online object reconstrucition that is quite similar to the one used in the offline analysis.

The output rate of the L1 trigger and the HLT can be regulated with the application of a factor f called "prescale", that reduces the trigger rate by accepting just one event every f selected candidates.

At the beginning of a fill of LHC the instantaneous luminosity $\mathcal{L}$ is equal to 2×10^{34} $\mathrm{cm}^{-2}\mathrm{s}^{-1}$ and it decreases with time (as shown in the first plot of Figure 1.3) because of the effects of the repeated crossings of bunches, where protons are lost in the collisions and the beam-to-beam interactions spread the particles distributions. The two beams deteriorate and prescale and threshold values may be changed in order to select more interesting events while keeping a constant output rate (see Eq. 1.2). The second plot on Fig. 1.3 shows the HLT output rate as a function of time: it does not decrease as quickly as the luminosity function, since prescales and cuts are changed at the red lines, increasing the number of accepted events. Only unprescaled trigger paths must be used for physics analyses, otherwise an appreciable fraction of interesting events would be lost.

The role of the CMS trigger system is to identify the interesting collision events before starting the offline analysis, which means a search strategy must be defined and trigger paths must be chosen accordingly.

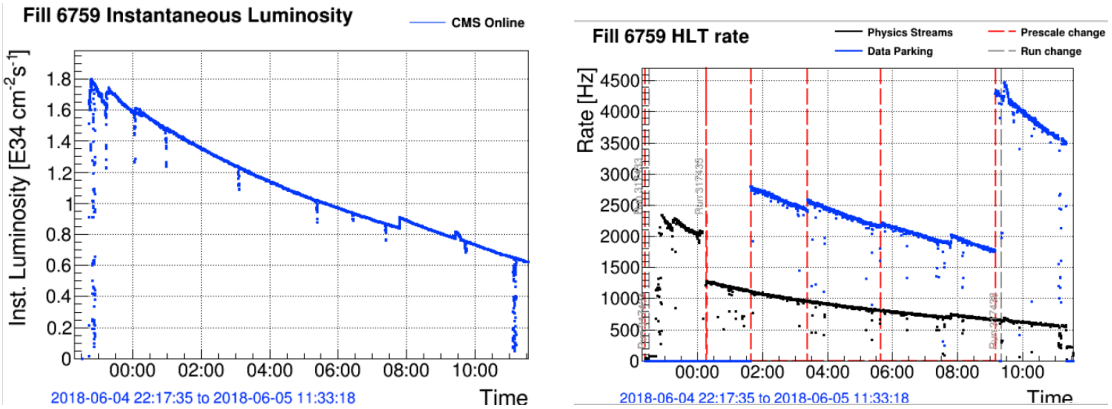


Figure 1.3: *On the left: instantaneous luminosity vs time. On the right: the black line is the output rate, the vertical bars denote prescale or cut changes.*

1.2.5 Object and event reconstruction

After the reconstruction of particle tracks and interaction vertices, the identification and reconstruction of particles that are produced in the event is done. This process might not be perfectly described in simulations: this could lead to data-MC discrepancies after the reconstruction. Several steps can be done to improve the agreement, like fixing the discrepancies with reweighting processes or systematic uncertainties.

The measured jet energy usually does not correspond to the real parton energy, from which the jet originates. This might be due to different reasons, including pileup and underlying events. Charged particles from pileup can be removed reasonably well with the reconstruction of partucle tracks, but this cannot be done with neutral particles. Moreover, inaccuracies in the simulation aggravate the discrepancies. With jet energy corrections (JECs) [9], these differences are fixed with some variation for data and simulation.

The pileup distribution is simulated before the data taking period. In order to correct most disagreements caused by the different pileup interactions, simulated events are multiplied by a weight that ensures a better agreement between data and simulations:

$$w_{\mathrm{PU}} = \frac{n_{\mathrm{data}}}{n_{\mathrm{MC}}} \tag{1.4}$$

where n_{data} and n_{MC} are the number of data and simulated events.

Also the trigger efficiency is different between data and simulation. In order to reduce this discrepancy, simulations are corrected with a multiplication factor called *scale factor*, which is given by:

$$sf_{\text{trigger}} = \frac{\epsilon_{\text{data}}(p_{\text{T}}, \eta, ...)}{\epsilon_{\text{MC}}(p_{\text{T}}, \eta, ...)} \tag{1.5}$$

where ϵ_{data} and ϵ_{MC} refer to data and MC efficiencies. In a similar way, also lepton and b-tag scale factors are needed whenever a selection on these objects is made, since the efficiencies are different in data and simulations.

Chapter 2

Preliminary notions on top quark physics

2.1 The Standard Model

The electroweak unification theory [10] is based on the local gauge invriance of the Lagrangian under SU(2) $\otimes$ U(1) and includes the electromagnetic and weak interactions. Electroweak theory and QCD (Quantum Chromodynamics) together form the Standard Model of fundamental interactions (SM), a quantum field theory that predicts the existence of gluons, photons and three more vector bosons, two charged, W^+, W^-, and one neutral, Z^0; the mass of the former is approximately equal to 80 GeV and the mass of the latter is about 90 GeV.

The theory is able to estimate branching ratios of vector bosons in all the decay channels, although it does not predict their masses. So far, these predictions have all been verified experimentally. The SM was proven to be renormalisable in 1972, which means the infinities that appear in the calculations of physical objects, cross-sections and decay rates can be eliminated by a mathematically correct procedure. This is done by writing the Lagrangian of all the interactions without mass terms, as if their mediators had zero mass; but while photon and gluons are massless, W and Z are not. Masses are generated by the spontaneous breaking of the local gauge symmetry, with a process called the Brout-Englert-Higgs mechanism (BEH). This mechanism does not destroy the renormalisability of the model and predicts the existence of a scalar boson known as Higgs boson (H) with a mass of approximately 125 GeV.

A summary of the elementary particles predicted by the SM is shown in Figure 2.1. However, inspite of its huge success in describing all particle physics observations so far, the SM seems to be incomplete: for example, it does not provide explanations for gravitation or dark matter and it implies that neutrinos have no mass, which contradicts experimental data. This means that more theoretical developments are needed.

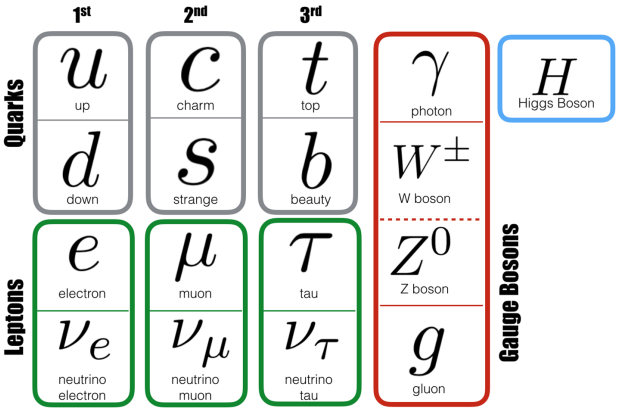


Figure 2.1: *Elementary particles of the SM.*

2.2 Top quark in the SM

The top quark is the most massive elementary particle in the SM. Even before its discovery, which happened in 1995 at the Tevatron collider at Fermilab [11, 12], there were already indirect measurements of its mass, since it appears in radiative corrections of electroweak processes, which led to the result

$$m_{\text{top}} \simeq 173\,\text{GeV}, \tag{2.1}$$

much bigger than the mass of the other quarks. The top quark mainly decays in a W boson and a b quark, with an amplitude that is approximately equal to 1.5 GeV and a lifetime of 5×10^{-25} s; its decay happens before the hadronization process: this means no hadron has the top as valence quark, since it cannot form bound states. The high value of its mass implies a large coupling to the Higgs boson, which is proportional to m_{top}, as well as an important role in electroweak corrections: measuring the mass of the top quark allows to verify the theory of the SM, as well as looking for deviations from it.

There are two different categories of processes that can lead to the production of top quarks: top quark pair production and single top quark production. The first one has higher cross sections and it includes both gluon (Figures 2.2a and 2.3b) and quark (Figure 2.2c) initial states. The single top quark production is shown in Figures 2.3a and 2.3b: it can be produced either by an intermediate W boson that decays into a top and an antibottom quark (s-channel) or by a bottom quark exchanging a W boson with an up or down quark (t-channel). Production of single top quarks, though more rare than pair production, offers the opportunity to study the top quark produced via the electroweak interaction.

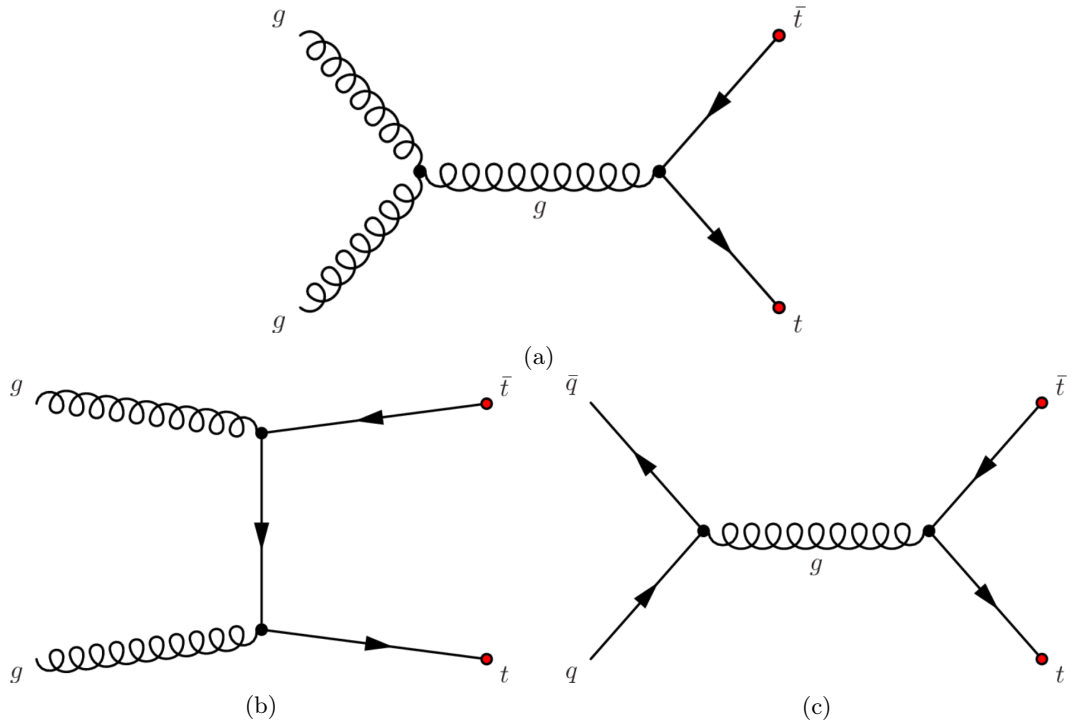


Figure 2.2: *Diagrams showing the production of top quark pairs $t\bar{t}$ in different channels.*

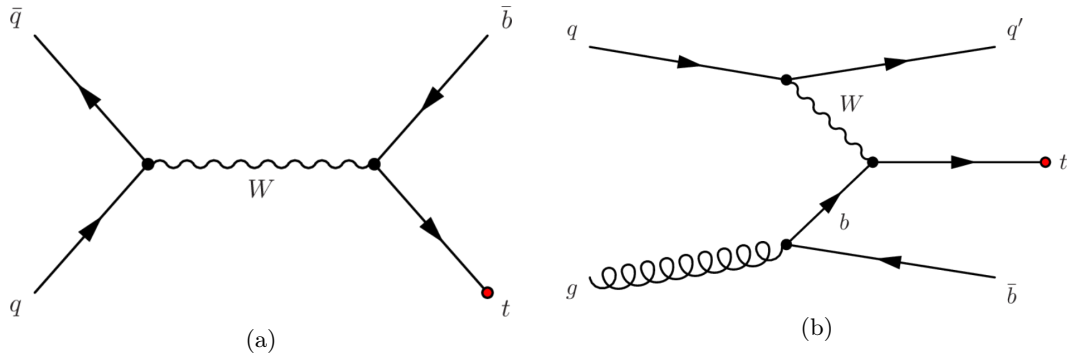


Figure 2.3: *Production of single top quark in (a) s-channel and (b) t-channel.*

2.3 tZq production and search

The dominant top quark production channels of pair production ($\sigma \simeq 830$ pb) and t-channel single production ($\sigma \simeq 220$ pb) have been measured with high precision, while many rare top quark production modes still needs to be studied further. These processes include, for example, the associated production of top quarks with heavy electroweak vector bosons, which allows a direct measurements of the electroweak couplings. A good summary of recent results about measurements of rare top quark production modes from the ATLAS and CMS Collaborations can be found in reference [13].

This analysis investigates the top quark-Z boson coupling. With data collected at the LHC from 2016 to 2018, direct measurements of single top quarks produced together with a Z boson became possible. This signature is commonly referred to as tZq, since the top quark and Z boson are produced together with a light quark. Top-Z couplings can also be studied considering top pair production, with $t\bar{t}Z$ events. The tZq process is particularly interesting for electroweak studies, since it allows direct measurements of the top-Z couplings (Figure 2.4a) as well as the WWZ triboson coupling (Figure 2.4b). Furthermore, there are measurements and searches of rare processes that have large backgrounds from tZq, like the boson production in association with a single top and a light quark (tHq); this provides an additional motivation for the study of the tZq process.

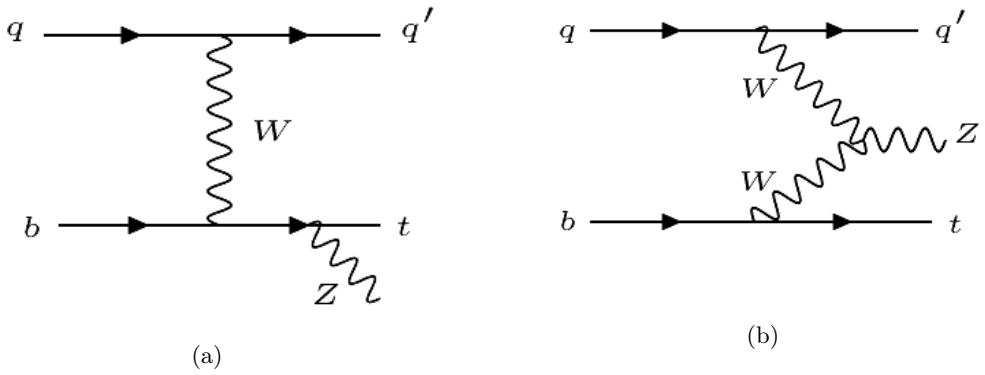


Figure 2.4: *Examples of diagrams for tZq production.*

The CMS and ATLAS Collaborations have carried out searches for tZq events with three leptons in the final state, two originating from the Z boson and the other one coming from the decay of the top quark. The extremely small production rate of tZq events and the large backgrounds made the observation of tZq really challenging. Machine Learning algorithms (see Section 3.1) were used to identify the leptons from the top quark and Z boson decays. Observation of the process above 5σ significance finally became possible in 2018 [14] (CMS) and 2019 [15] (ATLAS). .

Furthermore, tZq production is particularly sensitive to effects from Physics-Beyond-

the-SM (BSM): in the SM, interactions between quarks and Z bosons are not supposed to alter the flavor of the quarks involved; on the contrary, this is supposed to happen in some models concerning physics BSM.

Chapter 3

Multivariate Analysis, MVA

3.1 Introduction

A Multivariate Analysis (MVA) is a set of statistical techniques used to analyze data with multiple and possibly correlated variables; most Machine Learning (ML) methods [16] lie within this group. With the increasing amount and complexity of events in high energy physics, MVA is becoming increasingly important for analyzing data collected at LHC. The algorithms are usually categorized as supervised or unsupervised, depending on the type of learning used for the training process.

- **supervised learning**: the training procedure is done with a dataset of some observations having their corresponding labels/classes; the algorithm learns from the labeled dataset and produces a function to finally make predictions about some new unseen observations that can be given to the model.
- **Unsupervised learning**: the learning process is directly done with a dataset without the need of having the corresponding labels/classes; this type of learning aims to find a function to describe a hidden structure from unlabeled data.
- **Reinforcement learning**: algorithms that use the estimated errors as rewards or penalties: if the error is big, then the penalty is high and the reward low, and viceversa.

Supervised models can be further grouped into regression and classification cases:

- **classification** consists in identifying a decision boundary between objects of distinct classes;
- **regression** aims at estimating relationships among (usually continuous) variables.

Early ML applications in particle physics often used decision trees: a tree like-model that takes a set of features which will be used to split input data recursively. It can be used for both classification and regression problems; in the case of classification, each leaf

represents a decision assigning a data item to a class. In high energy physics, the most widely used are Boosted Decision Trees (BDT), where many trees are combined together in order to obtain a stronger classifier. More specifically, the learning process starts with unweighted events and a single decision tree; if a training event is misclassified, then the weight of that event is increased (boosted). A second tree is built using the new weights and again misclassified events have their weights boosted. The procedure is repeated again for more trees (typically 1000 or 2000).

Artificial Neural Networks (ANN or just NN) usually allow to create a more accurate and powerful classifier compared to BDTs, even though the structure is more complex and the training generally requires more computing resources.

3.2 Neural Networks

Neural Networks can be compared to biological brains in a simplified way: neurons and synapses correspond to connected layers of nodes, as shown in Figure 3.1. A node receives a set of input signals $(x_1, x_2, ..., x_n)$ from the previous layer and a weight $w_i (i = 1, .., n)$ is associated to each input signal. The *net* signal can be calculated as

$$net = \sum_{i=1}^{n} w_i x_i \tag{3.1}$$

and the output signal is calculated using an *activation function* which depends on the net signal and a threshold value b called *bias*. The bias can be considered as an additional input unit with the value $x_i + 1 = -1$ and the weight $w_i + 1 = b$; in this case the net signal is calculated as:

$$net = \sum_{i=1}^{n+1} w_i x_i \tag{3.2}$$

and the activation function becomes dependent only on this net signal.

A Neural Network is composed by three types of layers (see Figure 3.2):

- input layer, composed by nodes containing initial data;
- hidden layers, where all the computation is done;
- output layer, which produces results for the given inputs.

An ANN with multiple hidden layers is called Deep Neural Network (DNN).

3.3 Learning process

The learning process is the technique through which the weights of the network are determined. This is done by adjusting the weights until certain criteria are satisfied. For

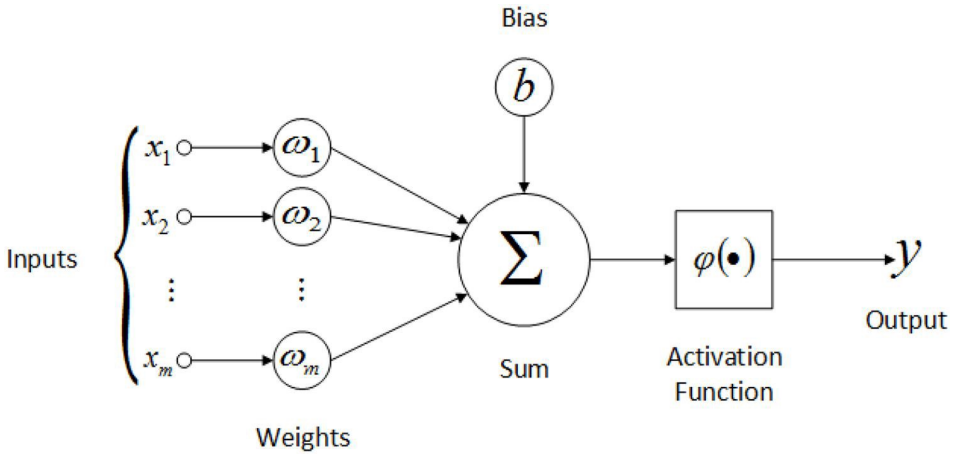


Figure 3.1: *Scheme illustrating how a Neural Network works: a node receives a set of input signals* $(x_1, x_2, ..., x_n)$ *from the previous layer and a weight* $w_i (i = 1, .., n)$ *is associated to each input signal with a weighted sum. The output signal is calculated using an activation function which depends on the net signal and a threshold value* b *called bias.*

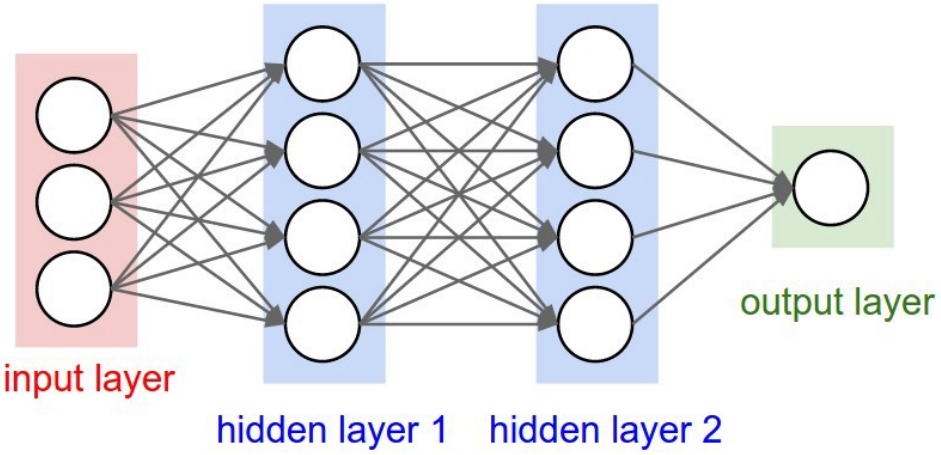


Figure 3.2: *Scheme with the three types of layers that compose the ANN: input, hidden and output layers.*

supervised learning, this is usually done with the Gradient Descendent (GD) rule, which aims to find the absolute minimum of a *loss function* $F(x)$: this method is based on the observation that if $F(x)$ is differentiable in a neighborhood of a point a, then this function decreases fastest if one goes from a in the direction of $-\nabla F(a)$:

$$a_{n+1} = a_n - \eta \nabla F(a_n) \tag{3.3}$$

where η is the learning rate, a parameter which controls the size of each step toward the minimum of $F(x)$.

The search of the minimum in GD algorithms can be accelerated by using the Momentum, a method that adds a fraction γ of a_n to the value a_{n+1}:

$$a_{n+1} = a_n + \gamma \Delta a_n - \eta \nabla F(a_n). \tag{3.4}$$

One of the disadvantages of these optimizers is that the learning rate is constant for each cycle. Some optimization algorithms, like the one that will be used for this analysis (Adam [17]), change the learning rate for each parameter and at every time step t, updating its value depending on the mean and the variance of the gradients; this property allows to identify the minimum with higher precision (Figure 3.3).

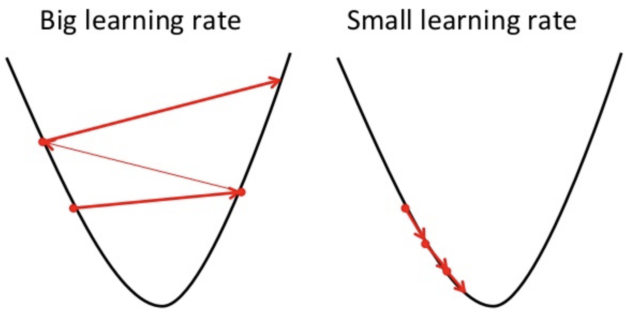


Figure 3.3: *If the learning rate (red arrows) is too high, it might be impossible to identify the minimum of the function F(x) (black line), while a small learning rate allows to identify it with higher precision.*

Training the NN with gradient-based learning methods might cause the *vanishing gradient problem*, which occurs when the gradient becomes really small, preventing the weight from changing its value. In the worst case, this may also interrupt the process of training. This problem can actually be solved with the activation funtion, which should be chosen depending on the type of problem that needs to be solved. This analysis will use a NN for a multi-class classification problem (there will be four different output nodes) with mutually exclusive output classes; in this case, the most common activation functions are:

- the softmax function $\sigma(z)$, which requires the outputs of each unit to be between 0 and 1 and also divides each output such that the total sum of the outputs is equal

to 1, giving the probability of the input value being in a specific class j; it is defined via the relation:

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \tag{3.5}$$

where the index *k=1, .., K* refers to the output classes. This activation function should be used just for the output layers: the purpose of this function is to give the probabilities of a given class to occur, and this would not be of much use in the hidden layers.

- The ReLU function

$$f(x) = max(0, x), \tag{3.6}$$

where the gradient is either 0 or 1, which means the gradient cannot vanish.

Therefore, in this work, ReLU will be the activation function in the hidden layers, while the softmax function will be used for the output layer.

There are also various different loss functions; for multi-class classification problems the most common is the multi-class cross-entropy loss, defined as

$$L(X_i, Y_i) = -\sum_{j=1}^{c} y_{ij} \times log(p_{ij}) \tag{3.7}$$

where Y_i is a vector $(y_{i1}, .., y_{ic})$, with y_{ij} equal to one if the i-th element is in class j and zero otherwise, and $p_{ij} = f(X_i)$ is the probability that the i-th element is in the class j; p_{ij} is calculated with softmax (which is then required as activation function in the output layer).

In general, too little training will mean that the model will *underfit* the training and test sets, causing a poor discrimination between signal and backgrounds. On the other hand, too much training might cause the learner to adjust to very specific random features of the training data that have no causal relation to the target function; this process is called *overfitting*: the model will stop generalizing and start learning the statistical noise in the training dataset, meaning that predictions on new data will be less reliable. In order to avoid this problem, a process called *dropout* can be used: a percentage of inputs to each layer will be randomly removed, in order to keep the training as general as possible.

3.4 Hyperparameter optimization

The parameters used to control the learning process and define the architecture of the model are called *hyperparameters*; by contrast, the values of other parameters (like node weights) are learnt. The hyperparameter optimization (or tuning) is defined as the problem of choosing a set of optimal hyperparameters for the algorithm. This can be done

following many different approaches; the ones used to train the DNN for the analysis presented in Chapter 4 are the random and the Bayesian optimization.

In random optimization, each hyperparameter is randomly sampled from a statistical distribution that must be given as input. Trainings with different hyperparameters can be done at the same time, since they are independent from each other, and the process goes on until a model with a good performance is found; this can be determined, for example, from the value of the loss or from the area under the ROC curve, which is created by plotting the true positive rate against the false positive rate and evaluates the discrimination between the different output classes. This algorithm allows a fast optimization procedure, but it must be taken into account that the finally selected hyperparameter set might not be the true best out of the ranges that has been set in the search.

The Bayesian optimization [18, 19] can be used when the analytical expression of the objective function $f(x)$ (which might be either the loss or the area under the ROC curve as function of one or more hyperparameters) is totally unknown and also expensive to evaluate, meaning that sampling at many points via random search would take too much time. This algorithm attempts to find the global optimum in a minimum number of steps by incorporating prior beliefs about f and updates it with samples drawn from the function itself, in order to get a better approximation. The model used for the approximation of the objective function is called *surrogate model*. Bayesian optimization also uses an *acquisition function* that directs sampling to areas where an improvement over the current best observation is likely.

A popular surrogate model for Bayesian optimization are Gaussian processes (GPs); they are defined as stochastic processes such that every finite linear combination of them is normally distributed and are usually cheap to evaluate. The acquisition function is then used to propose sampling points where the surrogate model predicts an optimal value for f, or where the prediction uncertainty is high. Both correspond to high values for the acquisition function, which will be maximized in order to determine the next sampling point. A common acquisition function is the Expected Improvement EI, defined as

$$EI(x) = \mathbb{E}\max(f(x) - f(x^+), 0) \tag{3.8}$$

where $f(x^+)$ is the value of the best sample so far and x^+ is the location of that sample. The training is repeated unitl the optimal value (which might be either a maximum or a minimum) is reached.

3.5 Applications in high energy physics

Due to an increased amount and complexity of the data collected, MVA started to be used in high energy physics since the late nineties [20, 21], even though with limited success: good results started to be achieved only in the 2000s. In terms of types of applications, MVA techniques, particularly ANNs, have been used for both online triggers (details

are reported in reference [22]) and offine data analysis, for which applications include a variety of tasks such as:

- track and vertex reconstruction,
- particle identification and discrimination,
- calorimeter energy estimation and jet tagging.

Many important physics results were extracted using an ANN; in these applications ANNs gave better results than the standard methods, mainly due to the highly non-linear character of the algorithm. It should also be considered that the training requires a datatset for which the target output is known: this is usually obtained by MC simulations, meaning that the result is very much dependent on the quality of the MC simulation.

Chapter 4

Analysis of tZq in the dilepton final state

Existing direct measurements of the coupling between the top quark and the Z boson consider events with at least three isolated leptons, in which two of them originate from the Z boson decay and one lepton appears in the decay of a top quark. Other final states in which the top quark or the Z boson decay hadronically are currently being studied in order to bring complementary information, exploring a different phase space of the data; in this analysis tZq events with exactly two leptons are investigated. The two leptons are assumed to originate from the Z boson, and the top quark decays hadronically. So far, the process has not been observed in this final state, since the identification of signal events is challenging due to the overwhelming $t\bar{t}$ and DY backgrounds. This channel has more statistics compared with the one having three leptons in the final state, meaning it can bring additional sensitivity to future combinations of all measured channels, which would help to improve the understanding of top-Z coupling and production process.

4.1 MC samples and pre-selection

The production of tZq with a dilepton final state is shown in Figure 4.1. In this channel dominant backgrounds originate from Drell-Yan production with additional jets (Fig. 4.2a) and $t\bar{t}$ in the dilepton final state (Fig. 4.2b), due to their large cross sections. Also, the dilepton decay channel of $t\bar{t}Z$ (Fig. 4.3a) must be considered: its cross section is comparable to that of tZq as it can be seen in Fig. 4.3b: cross sections of $t + Z$ and $\bar{t} + Z$ together are more or less equal to the one of $t\bar{t} + Z$. The structure of tZq and $t\bar{t}Z$ events is similar and this makes it hard to distinguish them from one another. Minor backgrounds have been included as well: Table 4.1 lists all the processes and MC samples that have been used, separated into five different groups (tZq, Drell-Yan, $t\bar{t}$, $t\bar{t}Z$ and minor backgrounds); corresponding cross sections are also reported. The signal sample is generated in the 4-flavour scheme at next-to-leading order (NLO) using MADGRAPH5_aMC@NLO (v5.2) [23]. Different MC generators were used to generate the background processes; DY samples have been generated with MADGRAPH at leading

order (LO), while both $t\bar{t}$ and $t\bar{t}Z$ have been simulated at NLO with POWHEG [24] and MADGRAPH respectively.

Table 4.1: *List of all the processes and MC samples considered in the analysis, separated into five different groups: tZq, Drell-Yan, $t\bar{t}$, $t\bar{t}Z$ and minor backgrounds; the middle column contains the naming convention of the samples as used within CMS, while the last column reports the values of the corresponding cross sections.*

Process	Sample name	Cross section [fb]
tZq		
tZq (dileptonic)	tZq_ll	94.18
Drell-Yan*		
Z+jets$\to \ell\ell$	DYJetsToLL_M_50_LO	6225420.0
Z+jets$\to \ell\ell$	DYJetsToLL_M_50_HT_70to100	208977.0
Z+jets$\to \ell\ell$	DYJetsToLL_M_50_HT_100to200	181302.0
Z+jets$\to \ell\ell$	DYJetsToLL_M_50_HT_200to400	50417.7
Z+jets$\to \ell\ell$	DYJetsToLL_M_50_HT_400to600	6983.94
Z+jets$\to \ell\ell$	DYJetsToLL_M_50_HT_600to800	1681.41
Z+jets$\to \ell\ell$	DYJetsToLL_M_50_HT_800to1200	775.4
Z+jets$\to \ell\ell$	DYJetsToLL_M_50_HT_1200to2500	186.222
Z+jets$\to \ell\ell$	DYJetsToLL_M_50_HT_2500toInf	4.38495
$t\bar{t}$		
$t\bar{t}$ (dileptonic)	TTTo2L2Nu	87310.0
$t\bar{t}$ (semileptonic)	TTToSemiLeptonic	364360.0
$t\bar{t}Z$		
$t\bar{t}Z$ (dileptonic)	TTZToLLNuNu_M_10	281.36
Minor backgrounds		
Single t (t-channel)	ST_t_channel_antitop	80959.0
Single $\bar{t}$ (t-channel)	ST_t_channel_top	136020.0
$W + jets \to \ell\nu$	WJetsToLNu	61526700.0
$W\gamma \to \ell\nu\gamma$	WGToLNuG	585800.0
$Z\gamma \to \ell\ell\gamma$	ZGToLLG_01J	124900.0
$t\bar{t}\gamma$	TTGamma_Dilept	1026.0
ZZZ	ZZZ	12.98
WZZ	WZZ	55.65
WWZ	WWZ	165.1

Table continues on the next page

*The inclusive DY sample is only included for HT<70 GeV, while the HT-binned samples are used in all the other cases; in this way the full phase space in HT is covered without double counting.

Process	Sample name	Cross section [fb]
WWW	WWW	208.6
$WW \to 2\ell 2\nu$	WWTo2L2Nu	12178.0
$WW \to 2\ell 2\nu$ (DS)	WWTo2L2Nu_DoubleScattering	1640.0
$ZZ \to 2\ell 2q$	ZZTo2L2Q	3220.0
$ZZ \to 2\ell 2\nu$	ZZTo2L2Nu	564.0
$WZ \to 2\ell 2q$	WZTo2L2Q	5606.0
tW	ST_tW_top	35850.0
$\bar{t}W$	ST_tW_antitop	35850.0
$t\bar{t}WH$	TTWH	1.582
$t\bar{t}ZH$	TTZH	1.535
$t\bar{t}HH$	TTHH	0.7565
$t\bar{t}WW$	TTWW	11.5
$t\bar{t}WZ$	TTWZ	3.884
$t\bar{t}ZZ$	TTZZ	1.982
$t\bar{t}t\bar{t}$	TTTT	9.103
$t\bar{t}W + jets \to l\nu$	TTWJetsToLNu	204.3

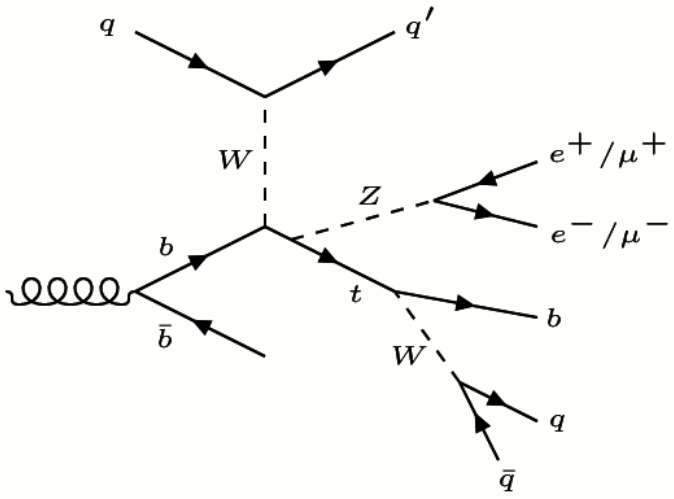
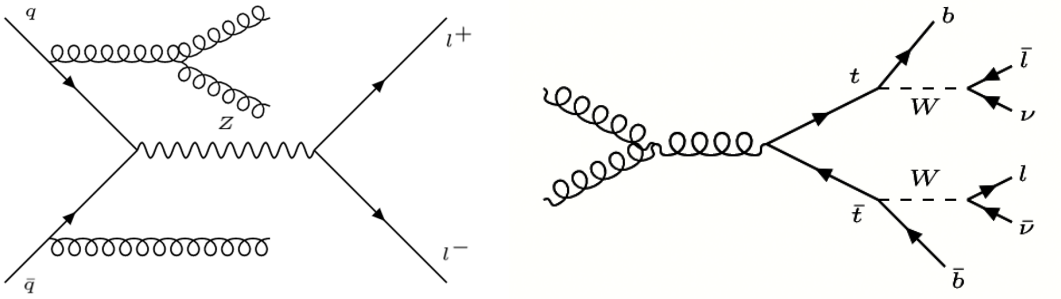


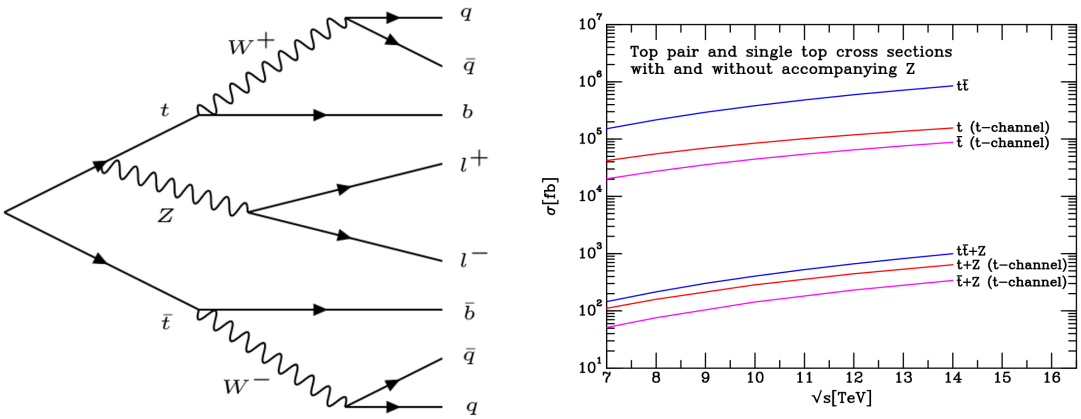
Figure 4.1: *A diagram for a tZq event with two leptons, originating from the Z boson, and hadronically decaying top quark.*

For the analysis, data recorded with single and double lepton triggers were used, most of them requiring a transverse momentum threshold between 20 and 30 GeV; similar requirements are imposed on MC events as well. A set of selection requirements is applied



(a) (b)

Figure 4.2: *Diagrams for (a) Drell-Yan plus jets and (b) $t\bar{t}$ processes, which form the biggest fraction of background.*



(a) (b)

Figure 4.3: *Diagram for $t\bar{t}Z$ in the dilepton final state (a) and its cross-section as function of the center of mass energy compared with tZq productions for top quark and anti-quark (b).*

for lepton identification (ID), based on the lepton's signature in the detector. The selections together, are known as "medium lepton ID". Current simulations and calibrations lead to a data-to-MC scale factor of $\simeq 99.2\%$ derived with a systematic uncertainty $< 0.4\%$. To identify jets originating from b-quarks, information about the energy of particles, tracks and vertices inside a jet are fed into a deep neural network (DeepCSV). The tight working point of the discriminator output is used, which corresponds to 0.1% misidentification rate.

The signal region is defined with the following criteria which already reduce the background contribution to some extent:

- exactly 2 leptons, with same flavor and opposite charge, and
  - $p_{\mathrm{T}} > 25$ GeV (first lepton) and $p_{\mathrm{T}} > 15$ GeV (second lepton),
  - $|m_{\ell\ell} - m_{\mathrm{Z}}| < 15$ GeV;
- at least 3 jets, with $p_{\mathrm{T}} > 25$ GeV and $|\eta| < 5$;
- at least one b-jet (included in the 3 previous jets).

Events fulfilling these requirements are used in the later stages of the analysis, including the reconstruction of the hadronically decaying top quark and the recoil jet.

4.2 Event reconstruction

Different methods have been considered for both the top quark and the recoil jet reconstruction, in order to identify the combination that gives the highest reconstruction efficiency, defined as the fraction of correctly reconstructed events. These methods will be used to select the input variables for a DNN, which will be subsequently trained to distinguish the signal from the three main background classes.

4.2.1 Hadronic top quark reconstruction

The hadronically decaying top quark is reconstructed from a b-jet and two additional jets coming from the W boson. The reconstruction method used here is based on minimizing a χ^2 defined as

$$\chi^2 = \left(\frac{|m_{\mathrm{t}}^{\mathrm{gen}} - m_{\mathrm{t}}^{\mathrm{reco}}|}{\Gamma_{\mathrm{t}}}\right)^2 + \left(\frac{|m_{\mathrm{W}}^{\mathrm{gen}} - m_{\mathrm{W}}^{\mathrm{reco}}|}{\Gamma_{\mathrm{W}}}\right)^2 \tag{4.1}$$

where $m_{\mathrm{t}}^{\mathrm{gen}}$ and $m_{\mathrm{W}}^{\mathrm{gen}}$ are the theoretical predictions for the masses of the top quark and the W boson, while $m_{\mathrm{t}}^{\mathrm{reco}}$ and $m_{\mathrm{W}}^{\mathrm{reco}}$ are their reconstructed values; Γ_{t} and Γ_{W} are the decay amplitudes of the two particles.

Three different ways to treat identified b-jets are studied:

- method I: all possible permutations of jets are considered (requiring b-tagging for exactly one of them), and the one permutation that gives the best χ^2 is taken for the top quark reconstruction;
- method II: the jet with the highest p_{T} is set as the b-jet coming from the top quark, and it is required that it passes the b-tag selection;
- method III: the jet with the highest b-tag probability is set as the b-jet coming from the top quark.

In the last two methods, the remaining two jets are also selected calculating the minimum χ^2, and the method that gives the highest efficiency is selected for the event reconstruction. A reconstructed top quark is considered as "correct" if its distance to the generated top quark in the ϕ (azimuthal angle)-η (pseudorapidity) plane is less than 0.2. The distance ΔR is defined as: $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$. The efficiency is then the ratio between correctly and all reconstructed top quarks, i.e.

$$\frac{N_{\mathrm{events}}(\Delta R < 0.2)}{N_{\mathrm{events}}(\mathrm{all})}. \tag{4.2}$$

The results obtained in this first part of the analysis are listed in Table 4.2, where the efficiency is reported for all the three years and collectively for Run 2. It can be seen that the first method turned out to be the most efficient, even if the results obtained from the different methods are quite similar to one another; this is due to events having just one b-jet (the second one, coming from the initial state, might not be reconstructed because of its low momenutm), for which the performances of the methods are identical.

Table 4.2: *Hadronic top quark reconstruction efficiency with the three tested methods. Results for the three years and for Run 2 are reported: it can be seen that the first method performs better, even if only slightly, and thus it is selected for the top quark reconstruction.*

Year	method I (χ^2)	method II (p_{T})	method III (b-tag)
2016	21.09 %	20.53 %	20.54 %
2017	24.75 %	23.97 %	24.16 %
2018	24.64 %	23.83 %	23.95 %
Run 2	23.89 %	23.14 %	23.26 %

4.2.2 Recoil jet reconstruction

Following a similar procedure, three methods are tested for the identification of the recoil jet:

- method I: the jet with the highest $|\eta|$ value is taken as recoil jet, requiring that it is not b-tagged;

- method II: the jet with the highest p_{T} value is taken as recoil jet, requiring that it is not b-tagged;
- method III: the jet with the highest likelihood ratio is taken as recoil jet.

The first and second methods are derived from the fact that the jet is recoiling against the whole system, therefore it is often emitted in the forward region and it is supposed to have a large p_{T}. In the third one a likelihood-ratio test is performed: three-dimensional histograms, one for the recoil jet and another one for all the other jets, filled with jet multiplicity, p_{T} and η of the jets from the signal MC samples are projected into the $(|\eta|, p_{\mathrm{T}})$ plane in order to obtain the likelihood-ratio distribution shown in Figure 4.4. This is done by considering the ratio:

$$\frac{(|\eta|, p_{\mathrm{T}})\ of\ the\ recoil\ jet}{(|\eta|, p_{\mathrm{T}})\ of\ other\ jets} \tag{4.3}$$

for each bin in the histograms. For each event, the p_{T} and pseudorapidity values are then used to find the jet with the highest likelihood-ratio, which will be identified as the recoil jet.

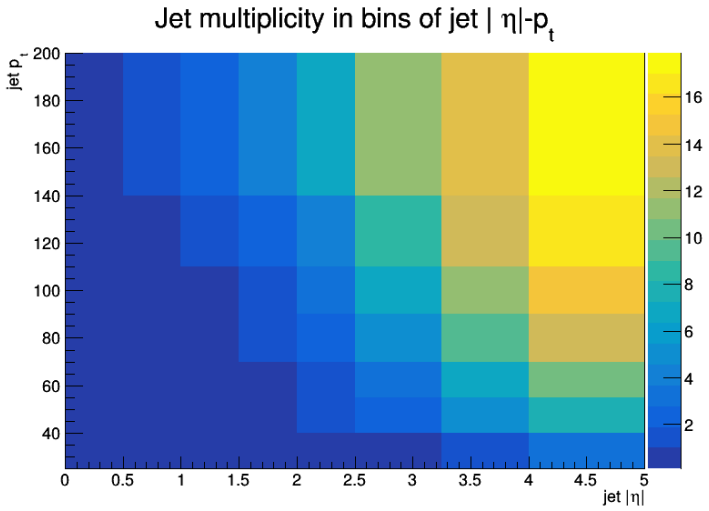


Figure 4.4: *Likelihood-ratio distribution obtained starting from three-dimensional histograms, one for the recoil jet and another one for all the other jets, filled with jet multiplicity, p_T and η of the jets from the signal MC samples. These are projected into the $(|\eta|, p_T)$ plan and the bins of the recoil jet distribution are divded by the values referring to the other jets.*

Results are reported in Table 4.3, where the fraction of correctly identified jets has been calculated taking into accout the ΔR between the generated and the reconstructed recoil jets, similarly to what has been done for the hadronic top quark reconstruction. The third method has been then selected to perform the identification of the recoil jet in the events.

Table 4.3: *Comparison between the three methods for the recoil jet identification. Numbers refer to the fractions of correctly reconstructed jets for the three years and for Run 2: it can be seen that the third method gives the highest identification efficiency, therefore it has been selected for the identification of the recoil jet.*

Year	I method ($\|\eta\|$)	II method (p_{T})	III method (Likelihood-ratio)
2016	50.66 %	50.86 %	62.33 %
2017	54.97 %	53.82 %	66.07 %
2018	54.87 %	54.97 %	66.54 %
Run 2	53.91 %	53.67 %	65.41 %

4.3 Signal-background discrimination with a DNN

It is hardly possible to isolate the signal from irreducible backgrounds with a cut-based approach using single variables. Therefore a DNN has been designed making use of Keras [25] and Tensorflow [26], two deep learning library written in Python and C++.

4.3.1 Input variables

The list of the input variables for the DNN is reported in Table 4.4. In addition to kinematic properties of the physics objects, variables coming from the event reconstruction are also employed for the training: vectors with variables from the four leading jets (p_{T}, $|\eta|$, ...) might be given as inputs instead of the ones from the reconstructed top quark and the identified recoil jet, but the discrimination was proven to be slightly higher if the event reconstrucition is taken into account. In this last case the loss and the value of the area under the ROC curve both reach a smaller value, as it can be seen in Figures 4.5a, 4.5b and Table 4.5.

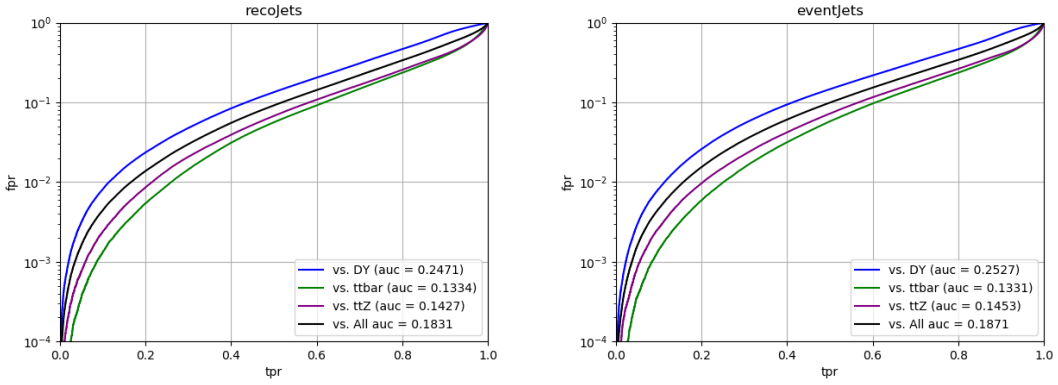
Table 4.4: *List of the input variables for the DNN.*

Variable
Properties of the physics objects
Number of jets
Number of b-jets
Maximum DeepCSV and QGLikelihood* values from the jets
$\Delta\phi$ and ΔR between the two leptons
Sum of lepton p_{T}
Sum of jet p_{T}
Maximum invariant mass and p_{T} of all any jet pair

Table continues on the next page

*Discriminator giving the probability that a jet was produced from a quark intead of a gluon. It can be particularly useful to distinguish the signal from DY events.

Variable
Maximum $\Delta\phi$ and ΔR between any jet pair
Missing transverse energy
Variables coming from the event reconstruction
Mass, p_{T} and η of the reconstructed Z boson
p_{T} and η of the ℓ^- from Z oson
p_{T} and η of the ℓ^+ from Z boson
Mass, p_{T}, η of the reconstructed top quark
Mass, p_{T}, η of the reconstructed W boson
p_{T}, η, DeepCSV and QGLikelihood values of the reconstructed b-jet from top quark
p_{T}, η, DeepCSV and QGLikelihood values of the two reconstructed jets from top quark
p_{T}, η, DeepCSV and QGLikelihood and likelihood-ratio of the reconstructed recoil jet
ΔR between reconstructed recoil jet and top quark
ΔR between reconstructed recoil jet and b-jet
ΔR between reconstructed W boson and b-jet
ΔR between reconstructed top qark and Z boson
ΔR between the two jets from the reconstructed W boson
Invariant mass of the reconstructed recoil jet and top quark
Invariant mass of the reconstructed recoil jet and b-jet



(a) (b)

Figure 4.5: *ROC curve obtained with (a) and without (b) event reconstruction. Plots have the true positive rate on the x axis and the false positive rate on the y axis, meaning that lower values imply higher discrimination. Four different types of ROC curves are shown: three refer to the discrimination between signal and each of the other DNN output nodes, while the black one considers all the main backgrounds as a single class.*

The neural network has been trained with signal samples (tZq) and with samples of the three main background classes (Drell-Yan, $t\bar{t}$ and $t\bar{t}Z$, all in the dilepton channel);

Table 4.5: *Values of the loss and the area under the ROC curve if variables from event reconstrucition are taken into account (second column) and if only kinematic properties of the physics objects are used (third column).*

	With event reconstruction	Without event reconstruction
loss	0.674	0.683
auc (tZq vs all)	0.1831	0.187

minor backgrounds have not been considered for this part of the analysis: in the end, the DNN will have four output nodes. About 50% of the whole MC dataset has been used for training, 30% for validation and 20% for testing.

4.3.2 Hyperparameter optimization

For practical reasons, only one or two hyperparameters are optimized at once, while keeping the default or already optimized values for the others. The expected fraction of tZq events is quite low ($\simeq$0.3%), as well as the fraction of $t\bar{t}Z$ ($\simeq$0.4%), while the ones for DY and $t\bar{t}$ are 67.8% and 31.5% accordingly: the four classes are imbalanced and the weights for tZq and $t\bar{t}Z$ need to be optimized in the process of hyperparameter tuning. The hyperparameters are then optimized in the following order:

- tZq and $t\bar{t}Z$ class weights,
- learning rate,
- number of layers and neurons,
- dropout.

The batch size is also considered in the hyperparameter tuning, but it turns out that its values don't affect much the loss or the area under the ROC curve.

The Bayesian optimization algorithm (see Section 3.4) is used for tZq and $t\bar{t}Z$ class weights, which have been optimized by minimizing the area under the ROC curve, while random values are set for the other hyperparameters. Figure 4.6 shows the results for this first part of the tuning: the black bullets refer to the points where the acquisition function extractes the samples for the training, while the optimum value is marked with a yellow star. It can be seen that the minimum of the surrogate model is reached if the weight for tZq is $\simeq$212 and the one for $t\bar{t}Z$ $\simeq$135. The same algorithm is used for the optimization of the learning rate, this time minimizing both the loss function and the area under the ROC curve; in both cases, optimal values are reached for a learning rate $\simeq 0.6 \times 10^{-3}$. These results are shown in Figures 4.7a and 4.7b. The Bayesian optimization performs well for class weights and learning rate, but it is not the best algorithm for the remaining hyperparameters optimal values are reached where the neural network

is overtrained. The number of neurons and layers, as well as the dropout, are then chosen using the random optimization algorithm, aiming to find a combination of the three hyperparameters with a good discrimination and not affected by overtraining.

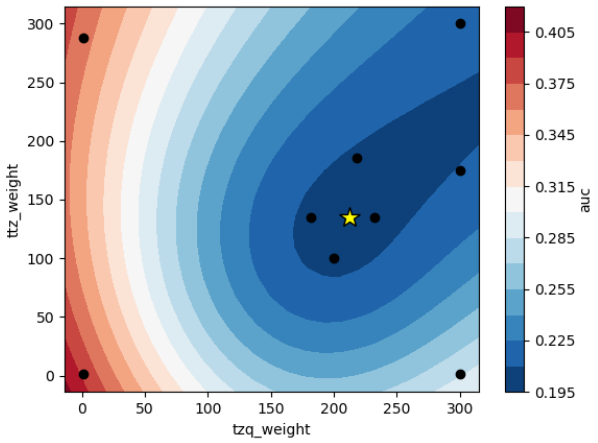
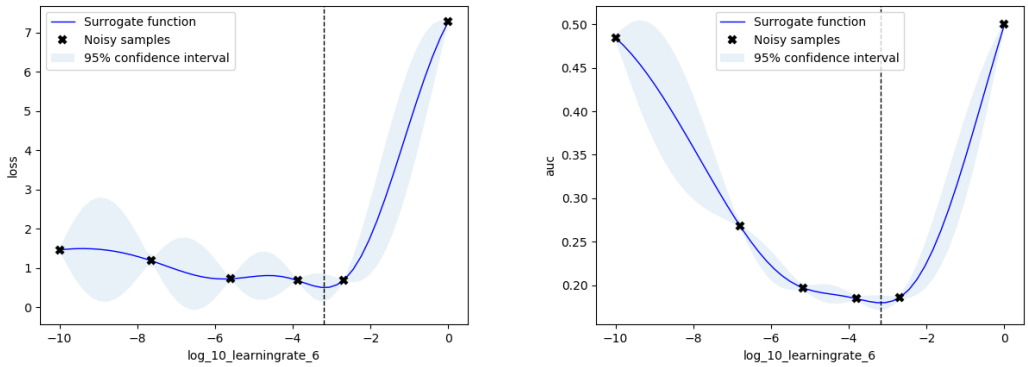


Figure 4.6: *Results for the Bayesian optimization algorithm applied to tZq and $t\bar{t}Z$ class weights; the area under the ROC curve is minimized. Black bullets refer to the points where the acquisition function extracted the samples for the training, while the optimum value is marked with a yellow star. The minimum of the surrogate model is reached if the weight for tZq is $\simeq$212 and the one for $t\bar{t}Z$ $\simeq$135.*

The final choices for the hyperparameters are listed in Table 4.6 and the output of the DNN for tZq node is shown in Figure 4.8a, together with results of Kolmogorov-Smirnov test between histograms obtained from training and testing; the plot in Figure 4.8b refers to the ROC curve of tZq against the three background classes; it can be seen that a good discrimination between tZq and the three main background classes is reached.

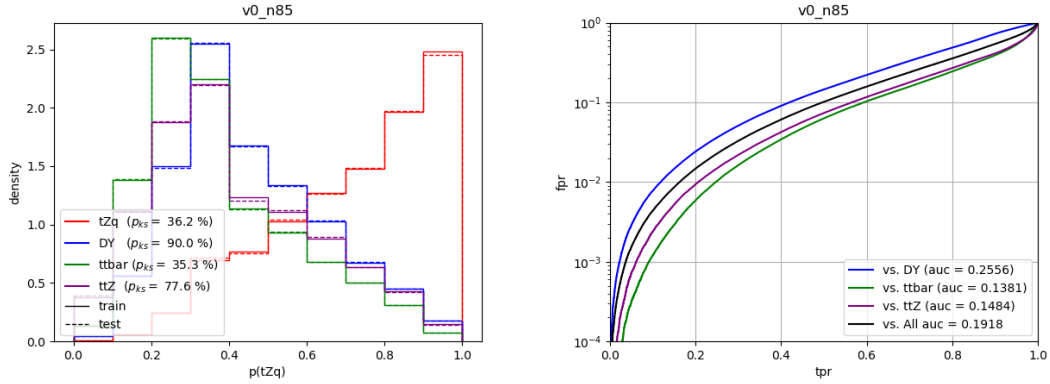
Table 4.6: *List of the hyperparameters considered for the tuning, together with their optimized values and the algorithm that has been used.*

Hyperprameter	Value	Optimization algorithm
tZq class weight	212	Bayesian optimization
$t\bar{t}Z$ class weight	135	
Learning rate	0.6×10^{-3}	
Number of neurons	85	Random optimization
Number of hidden layers	3	
Dropout	0.6	



(a) (b)

Figure 4.7: *Outputs of the Bayesian optimization algorithm for learning rate with the minimization of the loss (a) and the area under the ROC curve (b). The blue line shows the surrogate model and the light-blue area corresponds to the confidence level, while the 'noisy' samples are the points where the acquisition function extracts data for the training; these samples are called 'noisy' since it's possible that if the training is repeated with the exact same hyperparameters, the surrogate function might have a slightly different value.*



(a) (b)

Figure 4.8: *Output for tZq node (a), together with results of the Kolmogorov-Smirnov test between histograms obtained from training and testing, and ROC curve of tZq against the three background classes (b).*

4.4 Data-MC agreement

Control plots of the variables employed in the analysis are inspected in order to check the data-MC agreement. Examples of them are reported in this section: Figures 4.9a and 4.9b show the distributions of the maximum $\Delta\phi$ between any jet pair and the sum of the p_{T} of all the jets, while the distributions in Figures 4.9c and 4.9d refer to the reconstructed mass of the hadronic top quark and the η of the reconstructed Z boson. In the figures, all the minor backgrounds are grouped together and classified as "others". It can be seen that a good agreement between data and MC is reached, even though only statistical uncertainties are included so far.

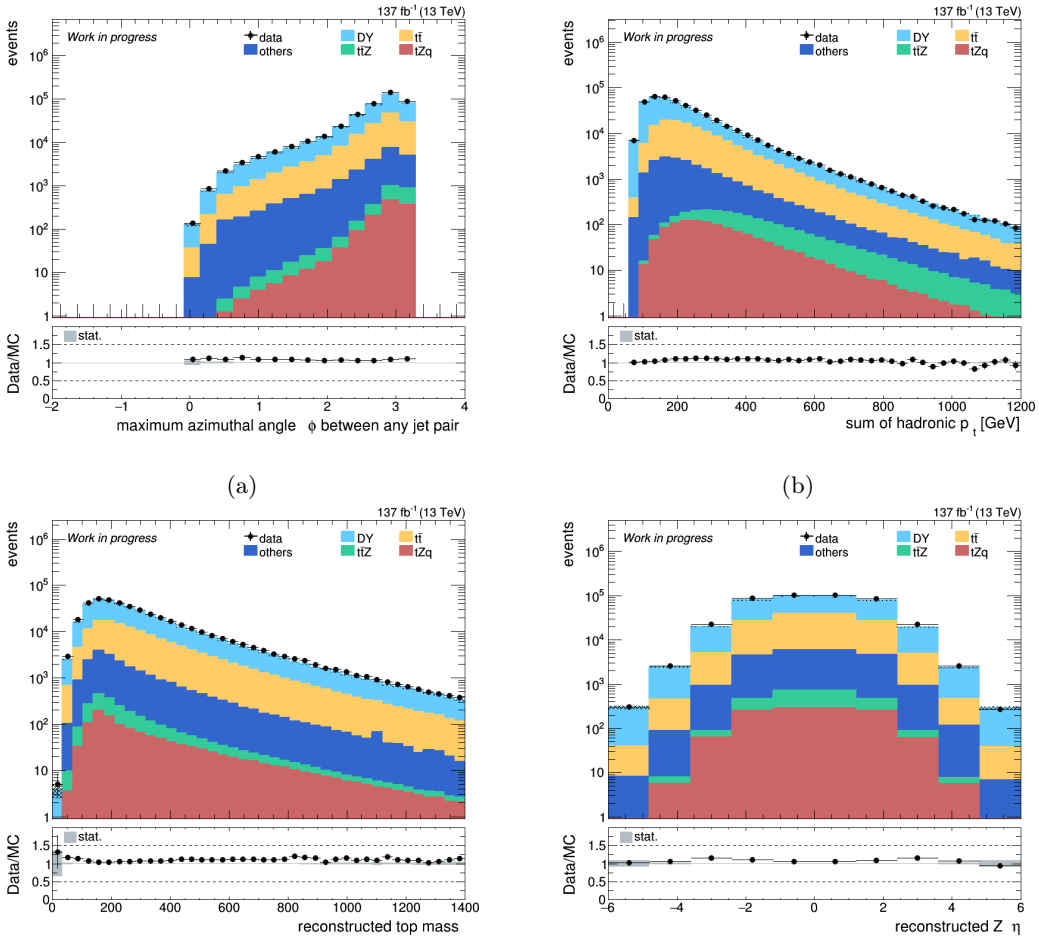


Figure 4.9: *(a) distribution of the maximum $\Delta\phi$ between any jet pair and (b) the sum of the p_T of all the jets (b); (c) distribution of the reconstructed mass of the hadronic top quark and (d) η of the reconstructed Z boson.*

4.5 Statistical fit of the DNN output

The last part of the analysis aims to perform a statistical fit to the DNN output for the signal extraction. The CMS Higgs Combination toolkit [27] is used to extract the significance, taking into account both statistical and systematic uncertainties; this is done with the ratio of profiled likelihoods, where the signal strength r is set to zero in the numerator and it is free to float in the denominator,

$$-2\ln\frac{\mathcal{L}(data|r=0,\hat{\theta}_0)}{\mathcal{L}(data|r=\hat{r},\hat{\theta})}, \tag{4.4}$$

where $\hat{\theta}_0$ denotes a set of nuisance parameters parameter that maximizes $\mathcal{L}$ for a specified r and the denominator is the maximized likelihood function; more information on the extraction of the significance in a likelihood-based fit can be found in reference [28].

Systematic errors are described for every source of systematic uncertainty. A distinction is made between *rate* and *shape* uncertainties. The first expresses the possible systematic deviation of a source in normalization. Shape uncertainties are more complex and are taken into account by vertical interpolation of the histograms: for each shape uncertainty and for each channel affected by it, two additional input histograms are provided. These are obtained by shifting the relevant parameter of uncertainty up and down by one standard deviation.

The fit is then performed simultaneously on three DNN output distributions corresponding to tZq, DY and $t\bar{t}$ nodes. In order to have orthogonal categories, only events for which the node value is maximum among all nodes have been included in each distribution. In each bin of the distributions, the number of events is assumed to be the sum of the yields for signal and for DY, $t\bar{t}$, $t\bar{t}Z$ and minor backgrounds. Contributions from DY and $t\bar{t}$ are treated as freely floating rate parameters, in order to be determined from data without any prior constraint.

As mentioned above, both statistical and systematic uncertainties are included in the fit. Systematics are listed in Table 4.7, grouped according to whether they affect the normalization or the shape of the distributions; it should be noted that the list is not fully complete, since the fit is still being finalized: matrix element uncertainties for DY, $t\bar{t}$ and minor backgrouds are still missing, and JECs (Jet Energy Corrections) should be splitted into the different sources. Uncertainties on the normalization of backgrounds other than $t\bar{t}$ and DY are set as log-normal effects, with an uncertainty equal to 15% for both $t\bar{t}Z$ and minor backgrounds.

Table 4.7: *List of systematic uncertainties grouped according to whether they affect the normalization or the shape of the distributions. The list is not complete, since the fit still needs to be finalized: matrix element uncertainties for DY, $t\bar{t}$ and minor backgrouds are still missing, and JECs should be split into the different sources.*

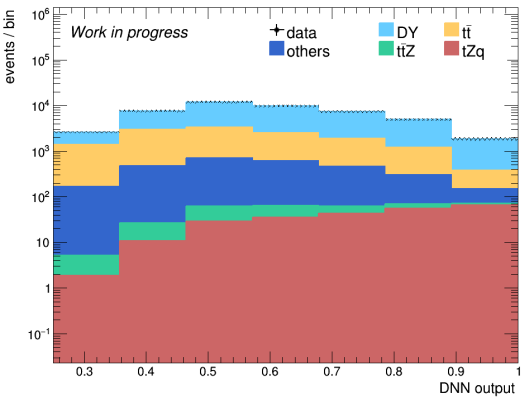
Systematics
Normalization uncertainties
luminosity (2.5% for 2016 and 2018, 2.3% for 2017)
trigger (2% for all the three years)
normalization for $t\bar{t}Z$ and *others*
Shape uncertainties
JEC
MET
matrix element (only for tZq and $t\bar{t}Z$)
b-tag scale factors
lepton scale factors
parton shower weights (2017 and 2018)
pile-up weights
L1 prefire weights (2016 and 2017)

Pre-fit plots for the three categories, together with statistical and systematic uncertainties, are shown in Figure 4.10.

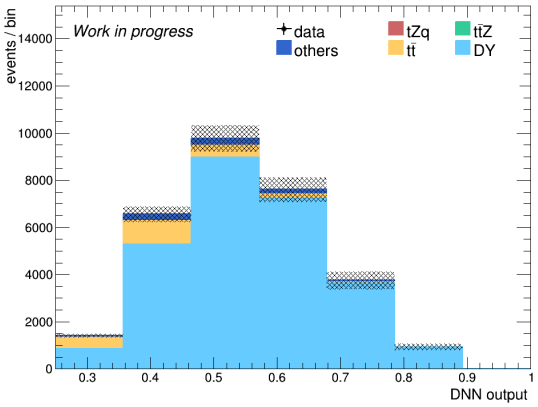
Each systematic is included in the fit as a nuisance parameter θ; its impact on the signal strength r is defined as the shift Δr that is induced as θ is fixed and brought to its $\pm 1\sigma$ post-fit values, with all other parameters profiled as normal. This is a measure of the correlation between the nuisance parameters and the signal strength, which is useful to determine which systematics have the largest effect. The plots in Figures 4.11a and 4.11b summarise the nuisance parameter values and impacts on the signal strength. The direction of the $+1\sigma$ and -1σ impacts (i.e. when θ is moved to its $+1\sigma$ or -1σ values) on r indicates whether the parameter is correlated or anti-correlated with it. The left panel in the summary plot shows the value of $(\theta - \theta_0)/\Delta\theta$, where θ and θ_0 are the post and pre-fit values of the nuisance parameter and $\Delta\theta$ is the pre-fit uncertainty; the difference between θ and θ_0 is equal to zero in this case, since a toy data is constructed using MC and the result does not depend on the observed dataset. The error bars show the pre-fit uncertainty divided by the post-fit uncertainty; parameters with error bars smaller than ± 1 are constrained in the fit, meaning that the initial estimation of the uncertainty needs to be revisited or the fit should be investigated further.

It can be seen that major systematics are JECs and they are also strongly constrained: this still needs to be examined in detail, for example splitting to the different JEC sources,

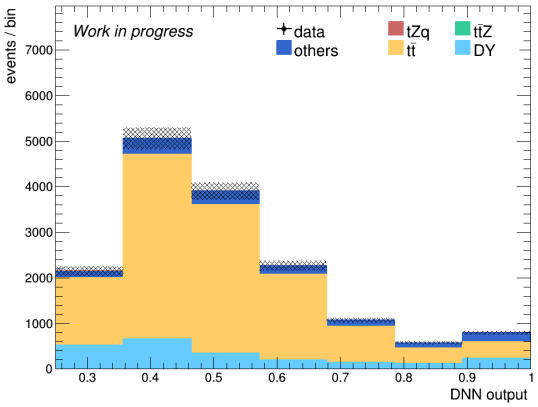
which could give additional information and help identify the problem.



(a)



(b)

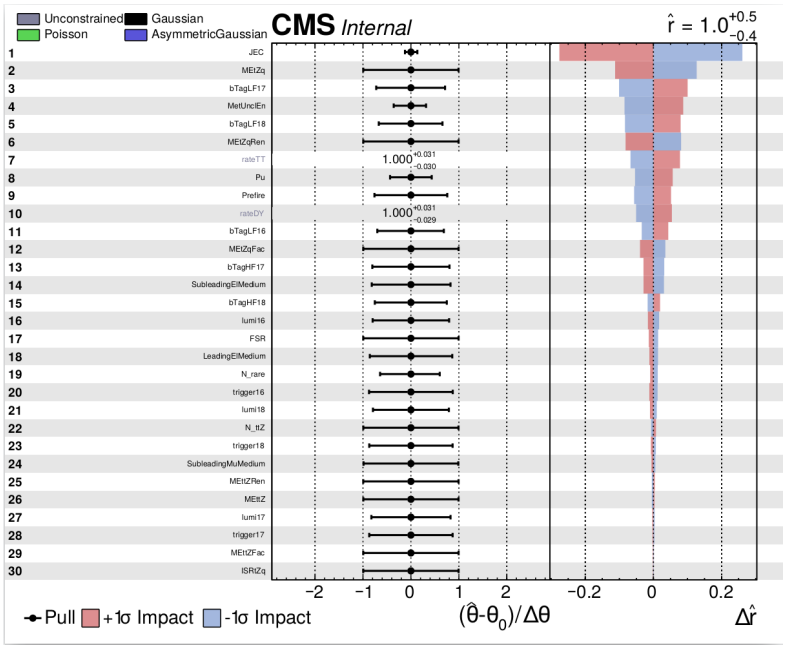


(c)

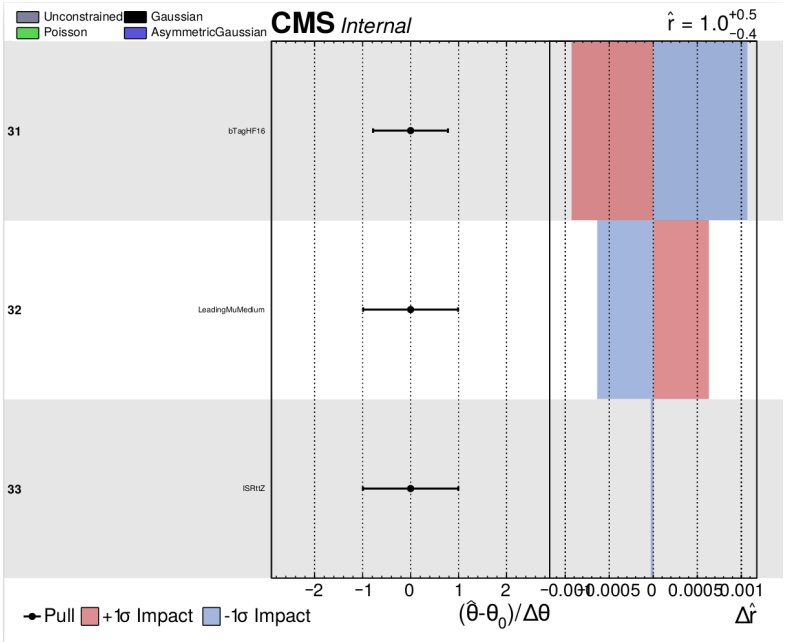
Figure 4.10: *Pre-fit distributions for the three categories, corresponding to (a) tZq, (b) DY and (c) $t\bar{t}$ output nodes. Statistical and systematic uncertainties are shown at the top of each histogram.*

Under these conditions, the expected significance of this analysis is 2.4σ, while it was equal to 4σ before considering systematics. The result may change after finalizing the fit with a more thorough study of uncertainties.

Additional categories could also be added in the fit in order to increase the expected significance: looking into regions with looser lepton IDs or with a different number of jets might allow to recover signal events which are lost with the current selections, or give additional information on the backgrounds.



(a)



(b)

Figure 4.11: *Nuisance parameter values and impacts on the signal strength. The direction of the impacts on r indicates whether the parameter is correlated or anti-correlated with it. The left panel shows the value of* $(\theta - \theta_0)/\Delta\theta$*, where* θ *and* θ_0 *are the post and pre-fit values of the nuisance parameter and* $\Delta\theta$ *is the pre-fit uncertainty. The error bars show the pre-fit uncertainty divided by the post-fit uncertainty.*

Chapter 5

Conclusions

The Z-boson associated single top quark production (tZq) has been investigated in this work, particularly focusing on final states with two leptons originating from the Z boson and the top quark decaying hadronically. Major backgrouds in this channel come from the dileptonic decays of Drell-Yan (DY), $t\bar{t}$ and $t\bar{t}Z$ events.

The first part of the analysis aims to study the event reconstruction, focusing on the top quark and the recoil jet, in order to achieve the highest reconstruction efficiency. For the top quark, taking the permutation of jets that gives, respectively within the resolutions, the closest dijet and trijet masses to the W boson and top quark mass turned out to be the method with the highest efficiency, while a likelihood ratio test using η and p_{T} of the recoil jet is done for the identification of the recoil jet.

A Deep Neural Network (DNN) has been designed in order to further discriminate signal from backgrounds; the training has been made with the outcome of the event reconstruction and other variables corresponding to kinematic properties of the physic objects: it has been shown that the use of jet kinematics that are associated with the top quark and the recoil jet improves the performance over cases where the event reconstruction information is not used. The Bayesian and the random optimization algorithms have been used for the process of hypermarameter tuning.

In the last part of the analysis, a statistical fit to the DNN output is performed for signal extraction. The fit is performed simultaneously on three DNN output distributions, corresponding to tZq, DY and $t\bar{t}$ nodes. Contributions from DY and $t\bar{t}$ have been treated as freely floating rate parameters, in order to be determined from data without any prior constraint. Major systematics are Jet Energy Corrections (JECs) and they are also strongly constrained: this needs to be studied in more detail.

Under these conditions, the expected significance of the tZq signal over backgrounds is 2.4σ, but the result may change after finalizing the fit with a more thorough study of uncertainties. Additional categories could also be added in the fit in order to increase the

expected sensitivity: looking into regions with looser lepton IDs or with a different number of jets might allow to recover signal events which are lost with the current selections, or give additional information on the backgrounds.

Acknowledgements

My deepest thanks goes to every member of the DESY CMS Top Group, for the friendly and stimulating working environment.

A special thanks goes to Abideh Jafari, Andreas Meyer and Pietro Govoni, for giving me the opportunity to work at DESY. Without them, this analysis would not have been possible.

Many thanks also go to David Walter and Nicolas Tonon, for their help and suggestions during my work.

Finally, I would like to thank all my family and friends, for always giving me love and support.

Bibliography

[1] CERN - Education, Communications and Outreach Group, *LHC Guide*, 2017. [Online]. Availabe: http://cds.cern.ch/record/2255762

[2] CERN, *The Large Hadron Collider*, 2020. Online. Available: http://home.cern/topics/large-hadron-collider

[3] The CMS Collaboration, *The CMS Experiment at the CERN LHC*, Journal of Instrumentation, vol. 3.

[4] The CMS Collaboration, *The CMS tracker system project: Technical Design Report*, Technical Design Report CMS, CERN, Geneva, 1997.

[5] The CMS Collaboration, *The CMS electromagnetic calorimeter project: Technical Design Report*, Technical Design Report CMS, Geneva, 1997.

[6] The CMS Collaboration, *The CMS hadron calorimeter project: Technical Design Report*, Technical Design Report CMS, Geneva, 1997.

[7] The CMS Collaboration, *The CMS muon project: Technical Design Report*, Technical Design Report CMS, Geneva, 1997.

[8] The CMS Collaboration, *The CMS trigger system*, Journal of Instrumentation, vol. 12, 2017.

[9] V. Sordini, *JEC levels*, 2016. [Online]. Available: https://twiki.cern.ch/twiki/bin/viewauth/CMS/IntroToJEC,

[10] Mark Thomson, *Modern Particle Physics*, Cambridge University Press, 2013, pp. 1-12.

[11] F. Abe, et al. (CDF collaboration), Phys. Rev. Lett. 74, 2626 (1995).

[12] 2. S. Abachi, et al. (DØ collaboration), Phys. Rev. Lett. 74, 2632 (1995).

[13] Joscha Knolle, *Rare top quark production at the LHC: $t\bar{t}Z$, $t\bar{t}W$, $t\bar{t}\gamma$, tZq, $t\gamma q$, and $t\bar{t}t\bar{t}$*, XXVII International Workshop on Deep-Inelastic Scattering and Related Subjects, 2019.

[14] CMS Collaboration, *Observation of single top quark production in association with a Z boson in proton-proton collisions at* $\sqrt{s}$*= 13 TeV*, CERN, 2018.

[15] ATLAS Collaboration, *Observation of the associated production of a top quark and a Z boson in pp collisions at* $\sqrt{s}$*= TeV with the ATLAS detector* , CERN, 2019.

[16] Alex Smola and S.V.N. Vishwanathan, *Introduction to Machine Learning*, Cambridge University Press, 2008.

[17] Diederik P. Kingma, Jimmy Ba *Adam: A Method for Stochastic Optimization*, International Conference for Learning Representations, San Diego, 2015.

[18] Peter I. Frazier *A Tutorial on Bayesian Optimization*, Cornell University, New York, 2018.

[19] Martin Krasser *Bayesian optimization*, 2018. [Online]. Available: http://krasserm.github.io/2018/03/21/bayesian-optimization/

[20] Dimitri Bourilkov *Machine and Deep Learning Applications in Particle Physics*, International Journal of Modern Physics, 2019.

[21] Liliana Teodorescu *Artifcial neural networks in high-energy physics*, Brunel University, United Kingdom, 2008.

[22] Bruce Denby *Computer Physics Communications*, Université de Versailles, France, 1999.

[23] Johan Alwall, Michel Herquet, Fabio Maltoni, Olivier Mattelaer, Tim Stelzer *MadGraph 5 : Going Beyond*, JHEP, 2011.

[24] S. Alioli, P. Nason, C. Oleari, E. Re, *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*, JHEP, 2010.

[25] *Home - Keras Documentation*, 2020. [Online]. Available: https://keras.io/

[26] *Tensorflow*, 2020. [Online]. Available: https://www.tensorflow.org/

[27] *Home - Combine - GitHub Pages*, 2020. [Online]. Available: https://cms-analysis.github.io/HiggsAnalysis-CombinedLimit/

[28] Glen Cowan, Kyle Cranmer, Eilam Gross, Ofer Vitells *Asymptotic formulae for likelihood-based tests of new physics*, Eur. Phys. J. C, 2011.