Measurement of $t\bar{t}$ production in association with a Z boson in pp collisions at $\sqrt{S} = 13.6\text{TeV}$

Konstantinos Pourgourides[1,2], Sergio Sanchez Cruz[2] and Giovanni Petrucciani[2]

[1] *Department of Physics, University of Cyprus, 1678, Nicosia, Cyprus*
[2] *European Organization for Nuclear Research (CERN), 1211, Geneva, Switzerland*

ABSTRACT

We present a measurement of the inclusive cross section of $t\bar{t}$ production in association with a Z boson in proton-proton collisions at a center-of-mass energy of 13.6TeV. The data samples used were collected by the CMS experiment in 2022 during run-3 of the Large Hadron Collider (LHC) at CERN, and they correspond to an integrated luminosity of 34.7fb^{-1}. The measurement is performed using final states that contain 3 charged leptons, where 2 of them are an oppositely charged, same flavour pair with invariant mass compatible to that of the Z boson. The production cross section is measured to be $\sigma(t\bar{t}Z) = 0.10 \pm 0.01(\text{stat}) \pm 0.01(\text{syst})\text{pb}$.

I. INTRODUCTION

The coupling of the Z boson with the top quark is of large significance in elementary particle physics as it can serve as a probe for several interesting phenomena, such as the violation of CP-symmetry or BSM processes.

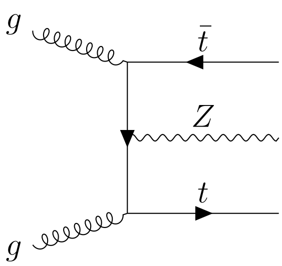

Figure 1: Leading order Feynman diagram of $t\bar{t}Z$ production

The primarily insightful process for this is the associated production of a top quark pair with a Z boson; the leading order Feynman diagram of this process is shown in figure (1). In this project, we focus on events with 3 charged leptons in the final state: the Z boson is hypothesized to decay to two oppositely charged, same flavour leptons, while the top quarks follow the decay mode $t \to W^- + b$ (or $\bar{t} \to W^+ + \bar{b}$), where b quarks give rise to b-jets. The W bosons can either decay to a lepton (anti-lepton) and an anti-neutrino (neutrino) or a $q\bar{q}$ pair (for example, u and d quarks). In order to achieve the desired final state, one W boson follows the hadronic channel decay and the other the leptonic one. Difficulty in the experimental detection of this process arises from the fact that other processes imitate it by possessing similar characteristics. In the next sections, methods to overcome this issue and discriminate between the two are discussed. This will contribute to later stages of the analysis, where signal or background enhanced regions will need to be defined.

II. METHODS

SAMPLES

The dataset used for this project was collected during two different time periods (or eras), which are denoted as 2022 and 2022EE, with luminosities 8fb^{-1} and 26.7fb^{-1} respectively. The difference between the two is that for 2022EE part of the ECAL of the CMS detector did not function properly due to damage. Monte Carlo (MC) generated events were used to model the signal and background processes and to predict their yields based on the luminosity and their theoretical cross sections. They were generated as follows: $t\bar{t}\ell\ell$ (signal; here the invariant mass of the di-lepton system is compatible to that of the Z boson, and thus this processes is referred to as $t\bar{t}Z$), $t\bar{t}W$, rares, conversions, tZq, $W+$jets, WW, WZ, tribosons, $t\bar{t}$ + tW, Drell-Yan, ZZ (background).

SELECTIONS

All events, signal or background, must satisfy certain conditions in order to be considered relevant for analysis. Initially, energy and noise filters as well as high-level trigger (HLT) selections were applied to prevent events of poor quality from entering the analysis. Additionally, events were required to contain at least 3 leptons that satisfy a tight ID criterion, $min\left[m\left(Z\rightarrow\ell\ell\right)\right]=$ 12GeV (this is the minimum invariant mass of the di-lepton system), and the transverse momenta of the 3 leading leptons had to satisfy the condition p_T^1 ≥25GeV, $P_T^{2,3}$ ≥15GeV. We select jets that pass the transverse momentum and pseudorapidity thresholds of p_T >25GeV and $|\eta|$ <2.4. MC events were corrected to be up to par with experimental data for jet energy scale and resolution, b-jet tagging and lepton efficiencies. In order to understand the behaviour of the signal and background processes, some plots with quantities of interest were produced and are shown in figure (2), with the two eras merged. l_W in this case denotes the lepton that derives from the leptonic decay of the W boson.

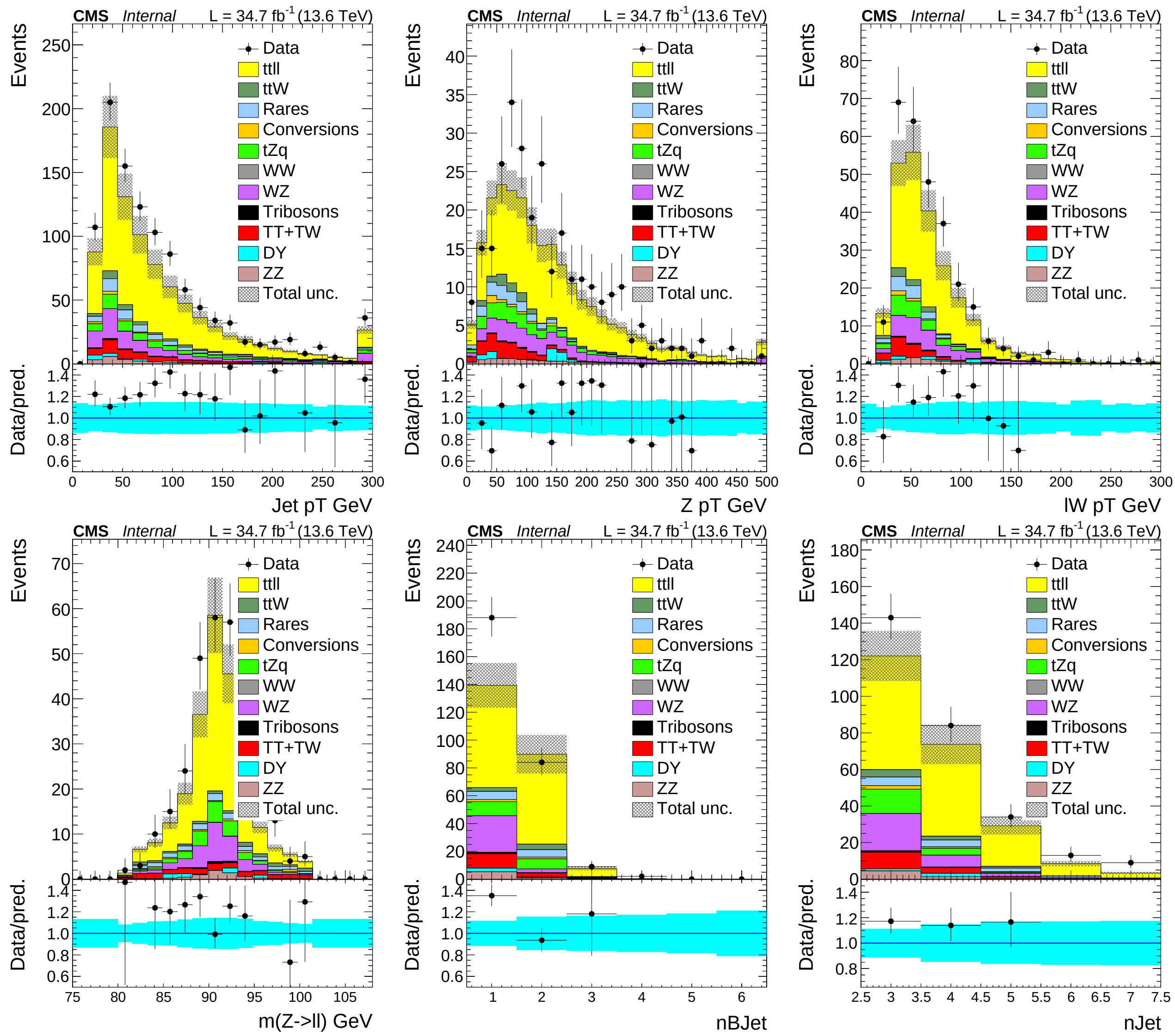

Figure 2: From left to right, up to down: jet p_T, Z boson p_T, l_W p_T, $m\left(Z\rightarrow\ell\ell\right)$, nBJet, nJet

Apart from the previous selections, in order to enhance the signal, the following selections were also made based on the nature of the process as described in the introductory section

- $|m(Z \to \ell\ell) - m_Z| \leq 10\text{GeV}$
- nJet ≥ 3 and nBJet (Medium ID) ≥ 1

Where $m_Z \approx 91.16\text{GeV}$ is the mass of the Z boson, and $m(Z \to \ell\ell)$ is its reconstructed mass

$$m(Z \to \ell\ell) = \sqrt{\mathcal{P}^\mu \mathcal{P}_\mu} \tag{1}$$

Here,

$$\mathcal{P}^\mu = p^\mu_{\ell_1} + p^\mu_{\ell_2} = \begin{pmatrix} E_{\ell_1} + E_{\ell_2} \\ \mathbf{p}_{\ell_1} + \mathbf{p}_{\ell_2} \end{pmatrix}^T$$

and $\ell_{1,2}$ is the oppositely charged, same flavour lepton pair whose invariant mass is compatible to that of the Z boson. As seen in figure (2), these selections have effectively differentiated between signal and background.

MEASURING THE CROSS SECTION

When MC events are compared with observed experimental data, discrepancies between their yields are noticed mostly due to differences in various efficiencies when conducting the analysis. These can be mitigated through the introduction of scale factors

$$SF \equiv \frac{\varepsilon^{\text{data}}_{\text{ff}}}{\varepsilon^{MC}_{\text{ff}}}$$

After the implementation of scale factors, differences might still persist. In order to quantify the observed agreement (or disagreement), one can fit the MC predictions to the data using a certain histogram, which will determine the signal strength, denoted as μ. The signal strength indicates how much the MC samples should be scaled by in order to be up to par with the observed data. More formally, it is defined as

$$\mu_{t\bar{t}Z} = \frac{N^{obs}_{t\bar{t}Z} - N^{bkg}_{t\bar{t}Z}}{N^{sig}_{t\bar{t}Z}} = \frac{\sigma(t\bar{t}Z)^{obs}}{\sigma(t\bar{t}Z)^{SM}} \tag{2}$$

thus, after the determination of the signal strength, the cross section is measured to be

$$\sigma(t\bar{t}Z)^{obs} = \mu_{t\bar{t}Z} \times \sigma(t\bar{t}Z)^{SM} \tag{3}$$

where $\sigma(t\bar{t}Z)^{SM}$ is the theoretical prediction for the cross section in the context of the Standard Model, which encapsulates the current knowledge about the signal process. The following subsection describes the statistical procedure by which one can determine the signal strength μ.

STATISTICAL METHODS

The method used to determine the signal strength μ is called Maximum Likelihood Estimation, which goes as follows [1]: let's assume a counting experiment, where a set of data $\mathbf{x} = (x_1, \dots, x_m)$ is observed after multiple repeated executions. Let's also assume that the outcome of the experiment depends on a set of unknown nuisance parameters $\mathbf{\Theta} = [\theta_1, \dots, \theta_k]^T \in \mathbb{R}^k$; thus, the observed data are distributed according to a probability density function $P(\mathbf{x}|\mathbf{\Theta})$. Then, the probability of observing a data point (event count) x_i in the interval $[x_i, x_i + dx_i]$ (which represents a single bin) is given by $P(x_i|\mathbf{\Theta})dx_i$. If the event counts in different bins are assumed to be independent, the joint probability of having observed the dataset $\mathbf{x}$ is just the product of the probabilities of having observed data points $x_i, \dots, x_m$ separately. By taking all of the above into account, the likelihood function can be defined as

$$\mathcal{L}(\mathbf{\Theta}) = \prod_{i=1}^{m} P(x_i|\mathbf{\Theta}) \tag{4}$$

The likelihood is treated as a function of $\mathbf{\Theta}$, considering that the observed dataset $\mathbf{x}$ if fixed (the experiment is over). Since the nuisance parameters are unknown and assumed to derive from some probability density functions ρ themselves, the probability of a parameter having being estimated as $\hat{\theta}_j$, assuming that its true value is θ_j, also has to be included in the likelihood

$$\mathcal{L}(\mathbf{\Theta}) = \prod_{i=1}^{m} P(x_i|\mathbf{\Theta}) \prod_{j=1}^{k} \rho_j \left(\hat{\theta}_j|\theta_j\right) \tag{5}$$

The general idea here is that one has to calculate the values for the nuisance parameters $\mathbf{\Theta}$ that maximize the likelihood of having observed the dataset $\mathbf{x}$ during the experiment. In principle, the satisfactory condition for the occurrence of maximum likelihood is

$$\frac{\partial \mathcal{L}}{\partial \theta_k} = 0, \quad \text{for all} \quad \theta_k \in \mathbf{\Theta} \tag{6}$$

In general, no close-form solution exists for this maximization problem, and thus, it is often done through numerical optimization. In the context of this project, this optimization was carried out by the ROOT-based Combine software [2].

Data collected during a counting experiment correspond to a Poissonian probability distribution function $\mathcal{P}$, and the specific hypothesis in this project is that the number of events in each bin consist both of signal and background, where their yields are a function of the nuisance parameters. Thus, equation (5) is modified to

$$\mathcal{L}(\mu, \mathbf{\Theta}) = \prod_{i=1}^{m} \mathcal{P}(x_i|\mu s_i(\mathbf{\Theta}) + b_i(\mathbf{\Theta})) \prod_{j=1}^{k} \rho_j \left(\hat{\theta}_j|\theta_j\right) \tag{7}$$

In order to evaluate the validity of a certain hypothesis concerning the observed data, a statistical test ought to be constructed. The relevant test in our case is called the profile-likelihood test statistic $q(\mu)$, defined as

$$q(\mu) = -2 \ln \frac{\mathcal{L}(\mu, \hat{\hat{\mathbf{\Theta}}}(\mu))}{\mathcal{L}(\hat{\mu}, \hat{\mathbf{\Theta}})} \tag{8}$$

Where $\hat{\hat{\mathbf{\Theta}}}(\mu)$ are the nuisance parameter values that maximize the likelihood for a given μ (often constrained in a certain range $\mu \in [0, \mu^{\max}]$), and $\hat{\mu}$ and $\mathbf{\Theta}$ are the best fit values without any constraints. This construction prevents the signal strength from taking negative values while also providing a one-sided confidence interval. If the parameter values that simultaneously maximize the likelihood are indeed the same as the ones calculated under the conditions applied for the value of μ, then $q(\mu)$ is minimized.

PREPARING AND PERFORMING THE FIT

The likelihood function is constructed on the hypothesis of both signal and background, and for this reason both should be given in adequate amounts to the software to perform the fit. For this reason, 3 different sample classifications are defined and implemented in the analysis; **0 b-jets:** suppression of signal and enhancement of background, especially WZ, which could have the same final state as the signal without any b-jets in the case of leptonic decay of the W boson. Drell-Yan and ZZ also become noticeable. **1 b-jet:** mixture of both signal and background; more background processes enter the picture, especially those that involve top quarks, since their decays give rise to b-jets. **>1 b-jets:** signal is greatly enhanced, background processes that could still produce more than one b-jet, such as $t\bar{t}W$, are still noticeable. The fit is eventually performed using the MC samples and observed data from 3 plots of nJet, one for each sample classification with merged eras. Before continuing to this step, the correct implementation of relevant uncertainties in the analysis is treated in detail over the next section.

UNCERTAINTIES

Systematic uncertainties stemming from different sources affect the analysis; some of them are associated with scale factors, others with the energy scale and resolution for jets, and others with the cross section of various processes. All of them are eventually reflected in the yields of the MC samples. Since none of them can be determined with complete accuracy, we have to account for their uncertainties, which are propagated throughout the analysis and significantly impact the final result; thus, they should be treated with caution. The way in which they are treated in this project is by parametrizing them with so-called nuisance parameters $\boldsymbol{\Theta} = [\theta_1, \ldots, \theta_k]^T \in \mathbb{R}^k$. Each nuisance θ_i parametrizes (or describes) one source of systematic uncertainty $\mathcal{U}_i^{syst}$. The total systematic uncertainty $\mathcal{U}^{syst}$ is a function of the partial systematic uncertainties.

Number	Nuisance Parameter	Associated Uncertainty Source
1	JESTotal	Jet energy scale
2	JER0	Jet energy resolution
3	CMS_eff_ttHbtag__correlated_heavy	b-tagging efficiency
4	CMS_eff_ttHbtag__uncorrelated_heavy	b-tagging efficiency
5	CMS_eff_ttHbtag__light	b-tagging efficiency
6	CMS_eff_ttHl__el	electron ID efficiency
7	CMS_eff_ttHl__mu	muon ID efficiency
8	WZ_JetMultiplicity	NNLO QCD interactions in WZ process
9	WZ	WZ cross section
10	ZZ	ZZ cross section
11	Conversions	Conversions cross section
12	Tribosons	Tribosons cross section
13	$t\bar{t}$ + tW	$t\bar{t}$ + tW cross section
14	Drell-Yan	Drell-Yan cross section

Table 1: All nuisance parameters used in the project and the uncertainty source they describe; uncertainties for cross sections are partly inspired by the run-2 measurement [3]

Apart from this type of uncertainty, the result also possesses statistical uncertainty $\mathcal{U}^{stat}$. The total uncertainty, assuming that systematic and statistical uncertainties are independent, is given by

$$\mathcal{U}^{tot} = \mathcal{U}^{syst} \oplus \mathcal{U}^{stat} \tag{9}$$

When performing the fit, this total uncertainty is calculated, and is attributed to both systematic and statistical effects. If the nuisance parameters are frozen to their estimated values, then the entire uncertainty is attributed to statistical effects

$$\mathcal{U}^{tot} = \left(\cancelto{0}{\mathcal{U}^{syst}}\right)_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}} \oplus \mathcal{U}^{stat} = \mathcal{U}^{stat} \tag{10}$$

When trying to calculate the contribution of a single systematic source (parametrized by nuisance θ_i) to the total uncertainty, all other nuisances are frozen to their estimated values. Hence, the only systematic contribution to the total uncertainty stems from θ_i

$$\mathcal{U}_i^{tot} = \mathcal{U}_i^{syst} \oplus \mathcal{U}^{stat} \quad \text{when} \quad \boldsymbol{\Theta'} = \hat{\boldsymbol{\Theta}}' \tag{11}$$

where $\boldsymbol{\Theta'}$ is a proper subset of $\boldsymbol{\Theta}$ ($\boldsymbol{\Theta'} \subset \boldsymbol{\Theta}$), because it includes all of its elements apart from θ_i. Therefore, the systematic uncertainty associated only with nuisance θ_i is given by

$$\mathcal{U}_i^{syst} = \sqrt{(\mathcal{U}_i^{tot})^2 - (\mathcal{U}^{stat})^2} \tag{12}$$

This uncertainty breakdown is calculated for each nuisance parameter included in $\boldsymbol{\Theta}$, as listed in table (1). Their respective contributions to the result are listed analytically in the next section.

III. RESULTS

PRE-FIT & POST-FIT DISTRIBUTIONS

As mentioned previously, the fit was performed using MC samples and observed data from 3 plots of nJet with merged eras, one for each of the different classifications. The pre-fit and post-fit distributions are presented in figure (3). It is clearly evident that after the fit there was a dramatic improvement in the ratio between data and MC samples, as well as much smaller uncertainty bands. The large uncertainty bands in pre-fit distributions are caused by the intended overestimation of each source's contribution, to avoid any uncertainty underestimations. These were eventually constrained by the fit, despite the large values that were assigned to them initially. Additionally, signal and background enhancement regions behave exactly as predicted.

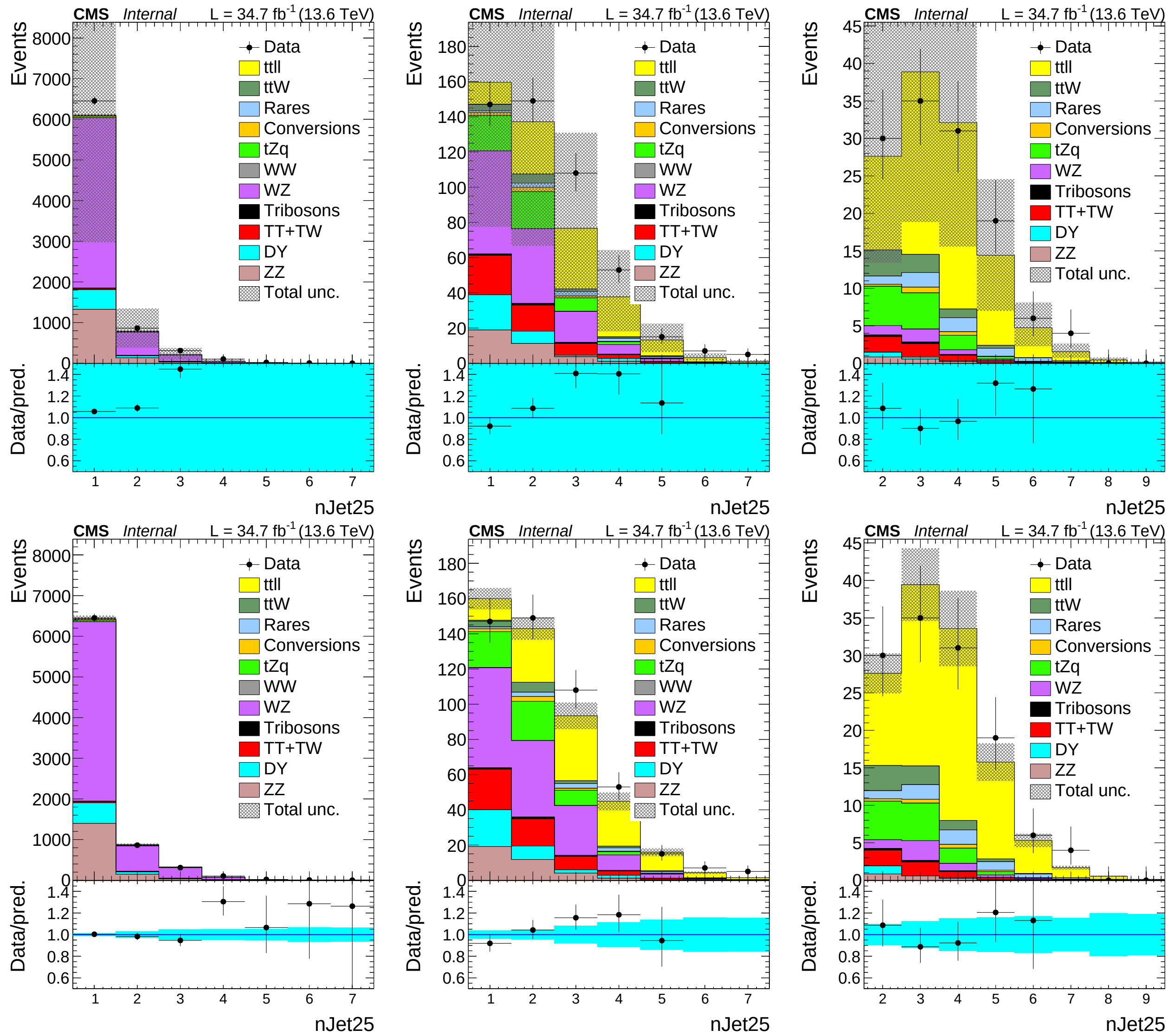

Figure 3: Pre-fit distributions used for the fit (above) and post-fit distributions after the fit was performed (below)

UNCERTAINTY BREAKDOWN

Number	Nuisance Parameter	Value	Impact on the $t\bar{t}Z$ x-sec (%)
1	JESTotal	1.45 ± 0.88	5
2	JER0	0.08 ± 0.33	<1
3	CMS_eff_ttHbtag__correlated_heavy	-0.85 ± 0.96	4
4	CMS_eff_ttHbtag__uncorrelated_heavy	-0.81 ± 0.97	4
5	CMS_eff_ttHbtag__light	-0.53 ± -0.92	2
6	CMS_eff_ttHl__el	0.07 ± 1.02	3
7	CMS_eff_ttHl__mu	-0.01 ± 1.00	1
8	WZ_JetMultiplicity	1.22 ± 0.32	3
9	WZ	0.02 ± 1.02	3
10	ZZ	0.02 ± 1.02	3
11	Conversions	0.03 ± 1.10	3
12	Tribosons	0.07 ± 0.82	3
13	$t\bar{t}$ + tW	0.04 ± 1.27	3
14	Drell-Yan	0.04 ± 1.21	3

Table 2: Nuisance parameter values and their impact on the $t\bar{t}Z$ cross section measurement

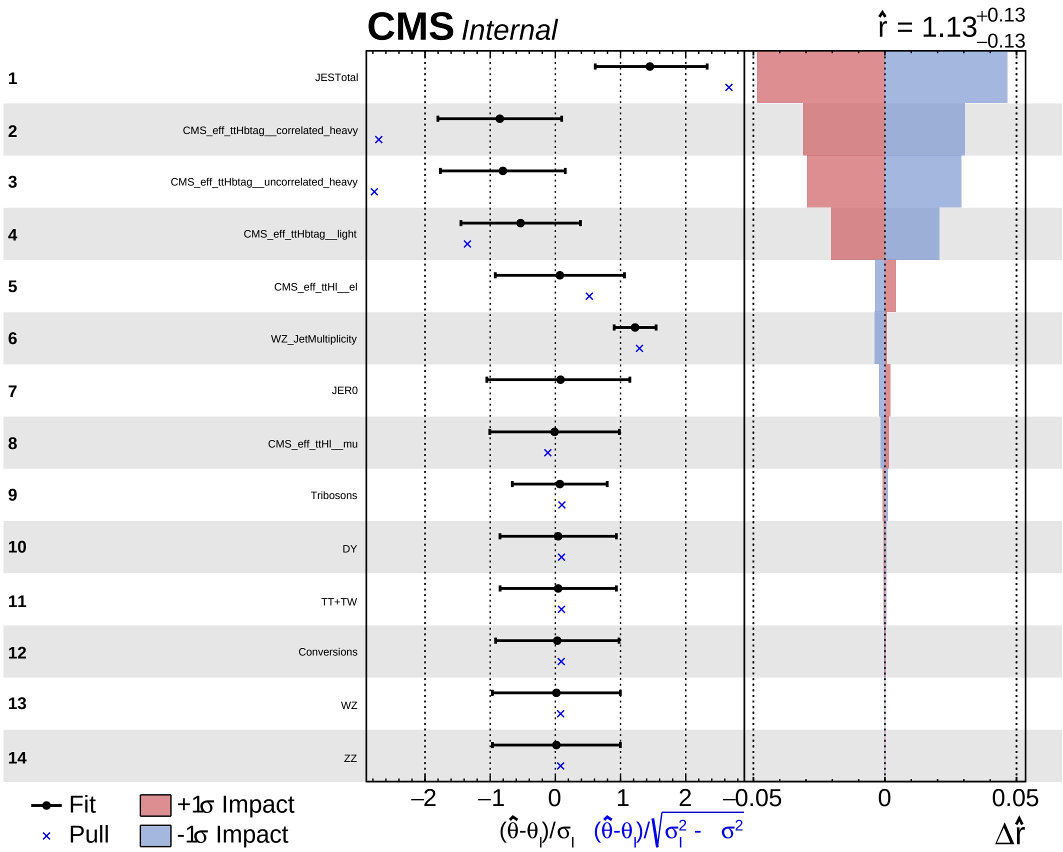

Figure 4: Impact plot: left side depicts the deviation of nuisance parameters from their true values; right side shows the impact of each nuisance parameter on the signal strength when assigned their $\pm 1\sigma$ values.

After the complete uncertainty breakdown was performed, the values of the nuisance parameters and their contribution to the total systematic uncertainty of the result were obtained, and are presented in table (2) and visualized in figure (4).

FINAL RESULTS

After taking everything into careful consideration, the signal strength is measured to be

$$\mu_{t\bar{t}Z} = 1.13 \pm 0.09(\text{stat}) \pm 0.10(\text{syst})$$

Subsequently, according to formula (3), the production cross section of $t\bar{t}$ in association with a Z boson at $\sqrt{S} = 13.6$TeV and $\mathcal{L} = 34.7\text{fb}^{-1}$ is measured to be

$$\sigma(pp \to t\bar{t}Z) = 0.10 \pm 0.01(\text{stat}) \pm 0.01(\text{syst})\text{pb}$$

The calculation was done by taking into account the SM prediction for the cross section in the case of $m(Z \to \ell\ell)$ >50GeV ($(\sigma(t\bar{t}Z)^{SM} = 0.08646$pb), since events with lower values of the invariant mass of the dilepton system could possibly correspond to fakes.

IV. CONCLUSIONS

To summarize this project, a measurement of the production cross section of $t\bar{t}$ in association with a Z boson at $\sqrt{S}$=13.6TeV was performed and presented using run-3 data of $\mathcal{L}$=34.7fb^{-1} collected during 2022 at the CMS experiment at CERN. The measurement regarded only final states with 3 charged leptons, one deriving from the leptonic decay of the W boson and an oppositely charged, same flavour pair with an invariant mass compatible to that of the Z boson. Signal and background discrimination was achieved through the utilization of appropriate selections, which contributed to the measurement, finalized by the ROOT-based Combine software. The final result is $\sigma(pp \to t\bar{t}Z) = 0.10 \pm 0.01(\text{stat}) \pm 0.01(\text{syst})$pb. Future re-measurement of this quantity with larger datasets (higher luminosity) and further addition of uncertainty sources could provide a better result.

V. ACKNOWLEDGEMENTS

I would like to express my gratitude towards my supervisors Dr. Sergio Sanchez Cruz and Dr. Giovanni Petrucciani for providing me with their support and guidance throughout the entirety of this project. I would also like to extend my thanks to the European Organization for Nuclear Research (CERN) for giving me the opportunity to spend a great summer at the largest research facility in the world, expanding my knowledge, curiosity and appreciation for science.

VI. REFERENCES

[1] Glen D. Cowan, "Statistical Data Analysis", Clarendon Press, Oxford Science Publications (1997)

[2] CMS Collaboration, "The CMS statistical analysis and combination tool: COMBINE", Computing and Software for Big Science (submitted April 2024), doi:10.48550/arXiv.2404.06614

[3] CMS Collaboration, "Measurement of top quark pair production in association with a Z boson in proton-proton collisions at $\sqrt{S}$ = 13 TeV", JHEP **03** (2020) 056, doi:10.1007/JHEP03(2020)056