Simultaneous measurement of tt+X processes in the semileptonic channel at the CMS experiment

Rufa Kunnilan M Rafeek

ETP, Karlsruhe Institute of Technology

The Compact Muon Solenoid (CMS) is one of the two general purpose de- tectors at the largest and the energetic particle accelerator in the world, the Large Hadron Collider (LHC).

A large number of top quark anti-quark pairs are produced, thus acting as main background to many Standard Model (SM) processes, thus making their analysis of prior important at the LHC. The top quark anti-quark pairs ($t\bar{t}$) are produced in association with other particles (X) where X can be the Higgs boson, Z/W boson or QCD-initiated heavy flavour jets ($b\bar{b}/c\bar{c}$). The $t\bar{t}$ pairs are pro- duced via strong interaction and in this analysis, the final state used is the semileptonic one, which contains a single isolated lepton (muon or electron).

The measurement of tt+X processes is important as it is a direct probe of the coupling of standard model particles to the top quark, gives better understanding of the background of many SM as well as Beyond SM processes and can also be a bridge to unravel new physics.

The measurement of tt+X processes with multiple heavy flavour jets in the final state, originating from different particles, like Higgs boson or Z-boson, for example, ttH(bb) or ttZ(bb) production, is challenging. These processes are the signal events in this analysis. In such events, it is important to not only identify the jet flavour but also to find the origin of each jet. Also, due to the limited spatial resolution of clustered jets, they can overlap, leading to a very small separation, particularly for events with b or c jets produced from collinear gluon splitting. Such events, tt+bb or tt+cc, are an irreducible background to ttH(bb) ot ttZ(bb) production. These events also serve as a major background of many SM as well as Beyond SM processes. Thus separating these processes with high efficiency gives better understanding of the SM, validating the SM and helping to identify any discrepancies that might suggest new physics phenomena.

Signal and background separation is very crucial for any analysis. The signal and background regions are defined by classifying events by their jet and b-jet multiplicities. The signal, tt+X, seemed to be concentrated more in the region with at least 6 jets with 4 b-jets since the signals, ttH with H ⇒ bb, ttZ with Zbb, ttbb and ttcc demand a large number of b-jets and the similar kinematic features of the b-jet and c-jet leads to the mis-tagging of a c-jet as a b-jet.

The phase-space of our analysis, as mentioned, is with at least 5-jets selected with $p_T \geq 30$ GeV within the tracking fiducial region of pseudorapidity, $|\eta| < 2.4$. These selection cuts are applied to the dataset (era = 2018) and trained using the simulated data provided by the CMS collaboration.

Further, due to the high jet multiplicity of final states of these tt+X events, the reconstruction and identification of these processes are very difficult, making them challenging to differentiate from each other. Hence, advanced multivariate methods have to be used for their classification. Here we utilise Graph Neural Networks (GNNs), which operate on a representation of the data as a mathematical graph, for this purpose. For the training, the jet categories used are the two additional jets, the b-jets associated with top quark decaying hadronically and leptonically and the other two light flavour jets associated with the hadronically decaying top quark.

Initially, the classification task is focused on the identification of the additional b-jet (and not the b-jet from the tt-system) as a binary classification problem. This is referred to as the Node Level Prediction (NLP) of the GNN, the output of which is used as an additional input for the multi-class (tt+X: tt+bb, ttH(bb), ttZ(bb) and tt+cc) event classification or the Graph Level Prediction (GLP). Here, each event is considered as a signal and is trained and classified against all the other event classes in the dataset. One of the assessment technique is analysing the Receiver Operating Characteristic (ROC) curve plots. Another methodology to compare the quality of an assignment strategy, the rate of identifying the two out of two (2/2) additional jets correctly with NLP approach over a baseline approach based on the minimal value of ΔR. When these methods are compared, there is a very significant improvement across all processes using the NLP aproach.

Secondly, events are classified into the various processes present within the phase space, the GLP. The framework is similar to that of the NLP and the difference here is that the input feature contains an additional input of the node being an additional jet or not. The multi class event classification can be analysed based on confusion matrices and ROC AUC values. The results are promising.

The next steps in the analysis of tt+X events involve a detailed study of systematic uncertainties, which will be crucial in understanding the robustness of the results. This systematic uncertainty study will encompass the evaluation of various sources of uncertainty, such as those arising from detector effects, modeling of the signal and background processes, and theoretical assumptions. Following this, a simultaneous fitting procedure will be employed to accurately model the tt+X events across different datasets and channels. Finally, the analysis will culminate in the simultaneous cross-section measurements of tt+X processes specifically within the semileptonic channel, providing key insights into the production mechanisms and contributing to the broader understanding of top quark physics.