

Development of a multivariate analysis for the study of the production of a W boson in association with 2 b -jets with the ATLAS experiment at the LHC^(*)

ATENA HARAREH⁽¹⁾ and EVELIN MEONI⁽¹⁾⁽²⁾

⁽¹⁾ *Department of Physics, University of Calabria - Cosenza, Italy*

⁽²⁾ *INFN - Cosenza, Italy*

received 2 December 2024

Summary. — This study investigates the production of a W boson in association with two b -jets in proton-proton collisions using the ATLAS detector at the LHC. A Neural Network approach is optimized to suppress the huge top-quark background, marking a step toward the first differential cross-section measurement of this process. The analysis is based on Monte Carlo simulations emulating the dataset collected by the ATLAS experiment at $\sqrt{s} = 13$ TeV. The study assumes an integrated luminosity of 140 fb^{-1} , corresponding to the full Run-2 dataset.

1. – Introduction

The production of a W boson in association two jets originated by b -quarks (b -jets) in proton-proton collisions, referred to as $W+2b$ in the following, is a process predicted by the Standard Model (SM). It is mediated by the strong interaction, therefore its study at the LHC contributes to the improvement of the understanding of perturbative QCD and the understanding of the proton fundamental structure. Furthermore, the measurement of this process provides a benchmark to probe and improve the modelling of Monte Carlo (MC) generators used to estimate this process in Higgs-boson measurements and searches for new physics where it constitutes an important background. In particular, $W+2b$ process is one of the main backgrounds for the processes $pp \rightarrow VH$, $H \rightarrow b\bar{b}$, with V being a W or Z boson, which is crucial for the measurement of the Higgs-boson coupling with the b quark. The uncertainties on the modeling of $W/Z+2b$ result among the main limiting factors of such measurements. The ATLAS and CMS Collaborations at the LHC performed various measurements of $W/Z+b$ -jets. The measurements are done selecting the leptonic decays on the W or Z boson ($W \rightarrow \ell\nu$ and $Z \rightarrow \ell\ell$, with ℓ =electron or muon) [1] and the cross-sections are measured in a fiducial phase-space close to the detector-level one to minimise the uncertainties. The CMS Collaboration conducted two studies focusing on the measurement of the production cross-section of

^(*) IFAE 2024 - “Poster” session

$W+b$ -jets using a proton-proton dataset of 5.0 fb^{-1} at $\sqrt{s} = 7 \text{ TeV}$, and a second proton-proton dataset of 19.8 fb^{-1} at $\sqrt{s} = 8 \text{ TeV}$. The ATLAS Collaboration published a study [2] using 5.0 fb^{-1} of proton-proton collisions data at $\sqrt{s} = 7 \text{ TeV}$ where the differential cross-sections of a W -boson production in association with b -jets has been performed, selecting first a fiducial phase-space with exactly 1 b -jet and second a fiducial phase-space with 2 jets of which at least one being a b -jet. While both Collaborations have already performed various differential cross-section measurements of $Z+b$ -jets and in particular $Z+2b$ -jets [3], neither ATLAS nor CMS has published measurements yet of the differential cross-section of the $W+2b$ -jets process, due to very large top-quark background. Moreover, at present there is no available $W+2b$ measurement (differential or inclusive) at $\sqrt{s} = 13 \text{ TeV}$. The study documented here is meant as the first step for the first ATLAS differential cross-section measurement of $W+2b$ -jets with the largest and best known dataset available so far, the full Run-2 pp dataset, corresponding to an integrated luminosity of 140 fb^{-1} at $\sqrt{s} = 13 \text{ TeV}$.

2. – MC simulation and Preselection

This study is based on simulated MC samples of the signal and the background processes officially produced by the ATLAS Collaboration. The simulated samples reproduce the experimental conditions of the full Run-2 dataset. They are processed using the full detector simulation based on Geant4. To account for multiple interactions in the same and neighbouring bunch crossings (pile-up), multiple overlaid inelastic proton-proton collisions are simulated with Pythia 8.186 using the A3 tune and the NNPDF 2.3 LO PDF set. The distribution of the average number of interactions per bunch crossing in the simulations is weighted to reflect the one in data. Two fully simulated signal samples of W boson production in association with jets ($W \rightarrow \ell\nu$) have been used in this analysis.

The nominal sample is produced using the MadGraph5 aMC@NLO (v2.6.5). It generates the matrix element (ME) of the process up to 3 jets at NLO accuracy in QCD and for the higher multiplicities it uses the parton shower provided by Pythia (MGaMC+Py8 FxFx) [4]. The alternative sample is generated with Sherpa 2.2.11, using the ATLAS Sherpa v2.2.11 configuration [5]. In this case the ME of the process is up to 2 jets at NLO accuracy and from 3 up to 5 jets at LO accuracy, for the higher multiplicities the Sherpa parton shower is used. For all the other backgrounds simulations the same path has been followed as in [6].

Muons, jets, b -jets, and missing transverse energy are the primary objects that have been used in this analysis, excluding electrons. The Inner Detector and Muon System independently reconstruct muons, which are combined with sub-detectors' information and calorimeter energy loss to create muon tracks for physics analyses. Muons selected in this analysis are required to fulfill "Medium" identification requirements [7] based on quality criteria applied to the inner-detector and muon-spectrometer tracks. Muons must be isolated from tracking and calorimeter energy deposits, with a p_T -dependent variable cone. Muon candidates for signal selection must have $p_T > 27 \text{ GeV}$ and $|\eta| < 2.5$, while muon candidates for veto of the presence of a second muon must have $p_T > 7 \text{ GeV}$ and $|\eta| < 2.5$. Jets are reconstructed using the anti- k_t algorithm [8] implemented in the FastJet package [9], with radius parameter $R = 0.4$, from particle-flow objects. Jets are calibrated using a simulation-based calibration scheme, followed by in situ corrections to account for differences between simulation and data. Central jets with rapidity $|\eta| < 2.5$ are required to have $p_T > 20 \text{ GeV}$, while forward jets, with $2.5 < |y| < 4.5$, are required to have $p_T > 30 \text{ GeV}$. Jets originating from pileup vertices are excluded by using the

“Tight” working point of JVT discriminant [10], which requires JVT to be above 0.5 for jets within the 20-60 GeV p_T range and $|\eta| < 2.4$. Jets originating from b -quarks are selected using a cut on the DL1r b -tagging discriminant, a deep-learning MV algorithm trained using information on the track and the secondary vertex. Jets are selected if they pass the 70% working point, which corresponds to an efficiency of 70% in selecting jets containing a b -hadron, 12.5% for jets containing a c -hadron and a rejection rate of about 0.3% for jets containing only light hadrons. Such b -tagged jets are required to have $p_T > 20$ GeV and rapidity $|y| < 2.5$. E_T^{miss} is determined as the negative vector sum of the transverse momenta from all identified hard physics objects (such as e , μ , jets), along with an extra track-based soft term that includes the contribution of unclustered particles [11]. After the selection of the W -boson candidates, two different strategies have been developed to select events containing 2 b -jets, which result in the definition of two different signal regions: the Geq2BiJ and the 2B2J regions. The Geq2BiJ strategy involves selecting events with at least two b -jets allowing other jets to be present in the event. This inclusive approach is similar to the one used in the ATLAS analyses for $Z+2b$ -jets. The results show that after the Geq2BiJ selection, the ratio of signal to $t\bar{t}$ background is approximately 5.3%. The ratio has a very small decreases if one considers the sum of all background (S/B), indeed it is 4.8%, indicating as expected that the $t\bar{t}$ is the dominant background, comprising about 90% of the total background. This strategy provides a good starting point but requires further selection to improve the signal-to-background ratio (S/B), which can be achieved using a Neural Network (NN) classifier.

The 2B2J strategy aims to reduce the significant background from semi-leptonic $t\bar{t}$ decays, which typically produce more jets than the $W+2b$ -jets signal. This strategy involves selecting events with exactly two b -jets and applying a jet veto. The 2B2J strategy results in a higher ratio of signal to $t\bar{t}$ background, at level of about 40%, with a S/B of about 26% but at the cost of reducing the signal to 30% (the number of signal events indeed passes from 1M to 30K) of the selected events with the Geq2BiJ strategy. The study uses Machine Learning (ML) techniques to achieve precise differential cross-section measurements of $W+2b$ -jets at the LHC, identifying a strategy for event selection

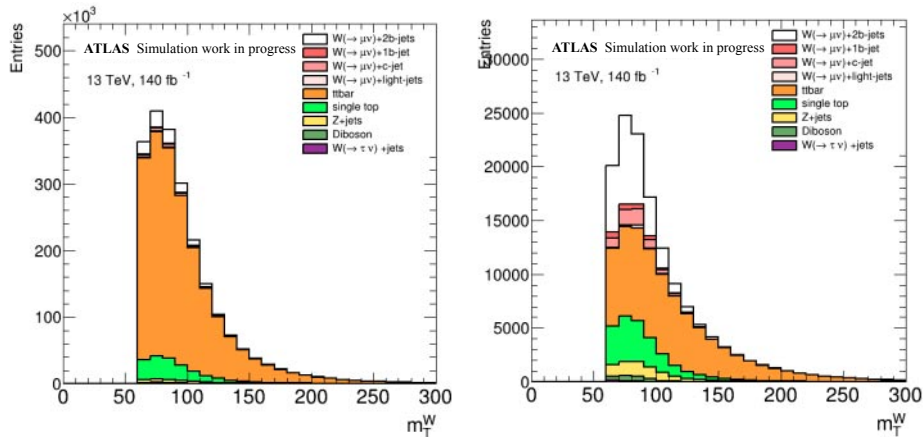


Fig. 1. – Distributions of m_T^W in events of Geq2BiJ (left) and 2B2J (right) regions.

at the detector level. Figure 1 shows the distribution of m_T^W (the transverse mass of the W boson, reconstructed with the charged lepton and measured missing momentum) as an example on the left after the Geq2BiJ selection and on the right after the 2B2J selection. The peak of m_T^W as expected is well visible, and the $t\bar{t}$ background dominates in both regions.

3. – Neural Network analysis

The study developed a NN to distinguish events involving W boson decay into muons and neutrinos, along with two jets, from background events. The NN is optimized to reject the most copious background, with the final state being the semileptonic decay. This study uses sequential models with dense layers optimized for binary classification using Keras 2.15.0 and Tensorflow 2.11. The hidden layers use the Rectified Linear Unit (RELU) activation function, while the output layer uses the Sigmoid function. Input features are preprocessed, and two networks are trained on even and odd events. The performance of the networks is evaluated using a test and validation set, with optimized hidden layers, nodes, and training iterations. After applying criteria that define the 2B2J selection, a study has been conducted to identify the observables that can effectively distinguish signal events from top background events. The physics objects present in events passing the 2B2J selection are: the muon, the missing transverse energy (that represents the neutrino), and the 2 b -jets. Using these objects a set of kinematic variables has been studied. Figure 2 illustrates the distributions of muon transverse momentum (p_T^μ), muon pseudorapidity ($|\eta_\mu|$), missing transverse energy (E_T^{miss}), the angle between the muon and E_T^{miss} in the transverse plane ($\Delta\phi(\mu, E_T^{miss})$), and the transverse mass of the W boson (m_T^W). The best separation of various observables is identified using signal and background event distributions, with an NN model implementing these variables and training with 50k signal events and 600k $t\bar{t}$ events. In order to perform an optimisation study a scan search approach has been implemented to establish the best hyperparam-

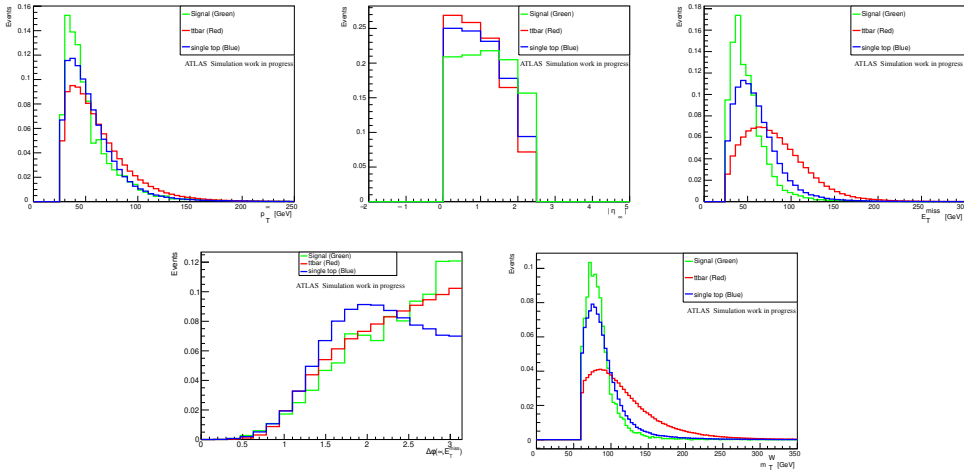


Fig. 2. – Distributions of: p_T^μ , $|\eta_\mu|$, E_T^{miss} , $\Delta\phi(\mu, E_T^{miss})$, and m_T^W in the 2B2J region for: nominal signal MC sample (green), $t\bar{t}$ (red), and single top (blue).

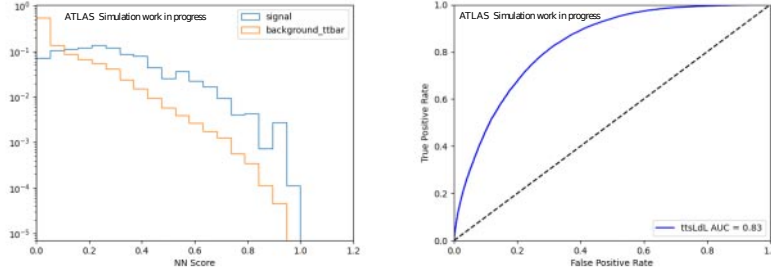


Fig. 3. – NN score distributions for nominal signal MC and $t\bar{t}$ background in the test sample (left). NN performance in terms of signal efficiency and $t\bar{t}$ background mis-identification efficiency obtained on the test sample (right). NN is trained using 5 features within the 2B2J region.

eters of the NN. The chosen figure of merit for the optimisation is the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve. This curve is generated by computing the rate of true positive signal events as a function of the false positive background events, and it is obtained by changing the NN value used to classify signal against background events. The chosen set of hyperparameters is as follows: 2 hidden layers, 15 nodes per layer, 100 epochs, and a learning rate of 10^{-3} . The model was then trained using the identified optimal hyperparameters, and the resulting performance has been evaluated to ensure that it provided the best results. Figure 3 shows the ROC curve of the NN model developed for this analysis using the hyperparameters defined in the optimisation procedure. It is evaluated using the testing sample. The AUC on the testing sample is 0.83 which suggests good ability of model to distinguish the two classes of events. The 2B2J selection is already rejecting a large fraction of signal events therefore a tight cut (at very large background rejection, but also quite reduced signal efficiency) would make impossible the measurement of differential cross-sections. A NN cut, selecting the 70% of the signal events (corresponding to 20K events) provides a $S/t\bar{t}$ of 1.3 and an overall S/B of 0.6 that seems adequate for this analysis. Considering that in the Geq2BiJ region in addition to the decay products of the W boson and the 2 b -jets there are additional jets in the final state, A NN has been developed including in addition to the 5 observables used for 2B2J region also other ones related to jets, namely: the number of central jets, N_{jet}^{centr} , the number of forward jets, N_{jet}^{fwd} , the scalar sum of the transverse momenta of central jets H_T^{centr} ($H_T^{centr} = \sum_{i=1}^{N_{jets}} |p_{T,i}|$), the sum of the masses of central jets M_{centr} ($M_{centr} = \sum_{i=1}^{N_{jets}} m_{jet}$), the scalar sum of the transverse momenta of all jets H_T^{all} , and the sum of the masses of all jets M_{all} . The number of signal events used for the training is 0.8M while 15M $t\bar{t}$ events are used. The chosen set of hyperparameters after the optimisation study is as follows: 2 hidden layers, 20 nodes per layer, 250 epochs, and a learning rate of 10^{-3} . Figure 4 shows the ROC curve of the NN model evaluated using the testing sample, the AUC on the testing sample is 0.85. The 11-features NN approach largely improves the $S/t\bar{t}$ ratio with respect to the 2B2J cut (3.6 against 1.3 at a working point selecting for both approaches 20K signal events), and an improvement is observed as expected also in the S/B ratio (0.95 against 0.6). One of the advantages of employing the Geq2BiJ region is the possibility to enhance the number of signal events, while still keeping a favorable S/B ratio. In this region one works at a signal efficiency between 25% and 40% keeping the number of signal events

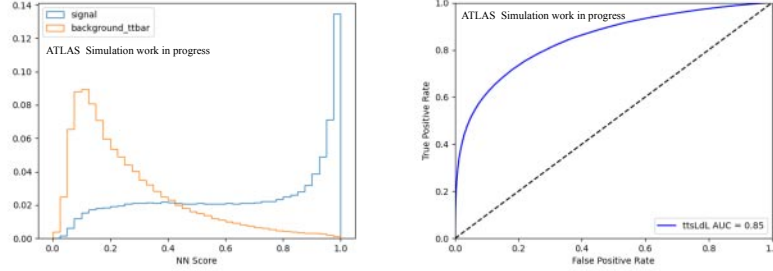


Fig. 4. – NN score distributions for nominal signal MC and $t\bar{t}$ background in the test sample (left). NN performance in terms of signal efficiency and $t\bar{t}$ background mis-identification efficiency obtained on the test sample (right). NN is trained using 11 features within the Geq2BiJ region.

between 25k and 40k with an extremely favorable $S/t\bar{t}$ ranging between 2.2 and 0.9, and still favorable S/B ranging between 0.8 and 0.5. The development of the 11-features NN in the Geq2BiJ has been repeated using the alternative signal sample, Sherpa 2.2.11. The NN developed with the second signal sample looks very similar to the nominal one in terms of performances: same AUC, and similar $t\bar{t}$ rejection. Similar results are also obtained using the alternative signal sample in the 2B2J region compared with the ones obtained with nominal MC.

4. – Conclusions

To conclude, this work is the first documented study of the event selection of the $W+2b$ -jets final state using the full Run-2 dataset of proton-proton collisions at 13 TeV of the ATLAS experiment at the LHC. The study is based on a ML technique for the background suppression. Two strategies proposed for the suppression of the $t\bar{t}$ background: the exclusive 2B2J selection and the inclusive Geq2BiJ one. Between the two approaches, the inclusive one is preferable. The study conducted using alternative signal MC samples shows that the systematic uncertainties related to the NN used for signal modeling in the final cross-section measurement will be small. This study represents the basis of the first measurement of the differential cross-sections of the production of a W boson in association with 2 b -jets at ATLAS.

REFERENCES

- [1] CMS COLLABORATION, *Phys. Lett. B*, **735** (2014) 204.
- [2] ATLAS COLLABORATION, *JHEP*, **06** (2013) 084.
- [3] ATLAS COLLABORATION, *JHEP*, **07** (2020) 044.
- [4] SJÖSTRAND T. *et al.*, *Comput. Phys. Commun.*, **191** (2015) 159.
- [5] ATLAS COLLABORATION, *JHEP*, **08** (2021) 089.
- [6] ATLAS COLLABORATION, arXiv:2403.15093 [hep-ex] (2024).
- [7] ATLAS COLLABORATION, *Eur. Phys. J. C*, **76** (2016) 292.
- [8] CACCIARI M., SALAM G. P. and SOYEZ G., *JHEP*, **04** (2008) 063.
- [9] CACCIARI M., SALAM G. P. and SOYEZ G., *Eur. Phys. J. C*, **72** (2012) 1896.
- [10] ATLAS COLLABORATION, *Eur. Phys. J. C*, **76** (2016) 581, arXiv:1510.03823 [hep-ex].
- [11] ATLAS COLLABORATION, ATLAS-CONF-2018-023.