Application of Machine Learning Based Top Quark and W Jet Tagging to Hadronic Four-Top Final States Induced by SM as well as BSM Processes

Petr Baroň[a], Jiří Kvita[a], Radek Přívara[a], Jan Tomeček[b], Rostislav Vodák[b]

[a]Joint Laboratory of Optics of Palacký University Olomouc and Institute of Physics of Czech Academy of Sciences, Czech Republic
petr.baron01@upol.cz, jiri.kvita@upol.cz, radek.privara@upol.cz

[b]Department of Mathematical Analysis and Applications of Mathematics, of Palacký University Olomouc, Czech Republic
rostislav.vodak@upol.cz, jan.tomecek@upol.cz

(Poster for the 16th International Workshop on Top Quark Physics, Traverse City, USA, Sept. 2023)

Palacký University Olomouc

GAČR CZECH SCIENCE FOUNDATION

Introduction

We apply gradient boosting machine learning techniques to the problem of hadronic jet substructure recognition using classical subjettiness variables available within a common parameterized detector simulation package DELPHES. Per-jet tagging classification is being explored. Jets produced in simulated proton-proton collisions are identified as consistent with the hypothesis of coming from the decay of a top quark or a W boson and are used to reconstruct the mass of a hypothetical scalar resonance decaying to a pair of top quarks in events where in total four top quarks are produced. Results are compared to the case of a simple cut-based tagging technique for the stacked histograms of a mixture of a Standard Model as well as the new physics process Machine learning (ML) techniques are getting growing applications in many research areas, one of them being the events classification in high energy physics (HEP).

1 Objects

Using the MADGRAPH5 version `2.6.4` simulation toolkit [1], proton-proton collision events at $\sqrt{s} = 14$ TeV were generated for the SM process $pp \to t\bar{t}$ in the all-hadronic $t\bar{t}$ decay channel at next-to-leading order (NLO) in QCD in production, using the MLM matching [2, 3], *i.e.* with additional processes with extra light-flavoured jets produced in the matrix element, matched and resolved for the phase-space overlap of jets generated by the parton shower using MADGRAPH5 defaults settings. The parton shower and hadronization were simulated using PYTHIA8 [4].

As a train BSM model, the resonant s-channel $t\bar{t}$ production via an additional narrow-width (sub-GeV) vector boson Z' as $pp \to Z' \to t\bar{t}$ using the model [5, 6, 7] were generated, to provide a sample of top quarks with large transverse momenta, enhancing the boosted regime.

As a representative model of a BSM process for testing, the production of a scalar resonance decaying to a pair of top quarks $y_0 \to t\bar{t}$ was adopted [6] at the leading-order (LO) in the $t\bar{t}$ production with the gluon-gluon fusion loop , with inclusive $t\bar{t}$ decays, selecting the all-hadronic channel later in the analysis.

1.1 Parameterized detector simulation

Using the DELPHES (version `3.4.1`) detector simulation [8] with the ATLAS card, jets with distance parameters of $R = 1.0$ (dubbed as large-R jets) were reconstructed using the anti-k_t algorithm using the FastJet package [9] at both particle and detector levels.

The trimming jet algorithm [10] as part of the DELPHES package was used to obtain jets with removed soft components, using the parameter of $R_{\text{trim}} = 0.2$ and modified p_{T} fraction parameter $f_{\text{trim}}^{p_{\text{T}}} = 0.03$ (originally 0.05). The trimming algorithm was chosen over the standard non-groomed jets, soft-dropped [11] and pruned jets [12], with parameters varied, in terms of the narrowness of the mass peaks.

1.2 Objects of interest

The interest is the identification of large-R hadronic jets coming from the hadronic decays of top quarks and W bosons. In the naïve picture of the hadronic decays of $W \to q\bar{q}'$ and $t \to Wb \to bq\bar{q}'$, these three and two prong decays, respectively. Different jet substructure is thus expected for such t and W jets.

1.3 Cut-based tagging

As input variables to both cut-based as well as ML-based tagging we utilize simple yet powerful "classical" variable called n-subjettiness [13], τ_N, which is related to the consistency of a jet with the hypothesis of containing N subjets. These variables are combined into ratios τ_{32} and τ_{21}, defined as $\tau_{ij} \equiv \frac{\tau_i}{\tau_j}$.

In order to identify jets coming from the hadronic decays of the W boson or a top quark by a simple cut-based algorithm, large-R jets were tagged as

- W-jets if $0.10 < \tau_{21} < 0.60 \wedge 0.50 < \tau_{32} < 0.85 \wedge m_J \in [70, 110]\,\text{GeV}$;
- top-jets if $0.30 < \tau_{21} < 0.70 \wedge 0.30 < \tau_{32} < 0.80 \wedge m_J \in [140, 215]\,\text{GeV}$.

Shapes of the variables used as input to the ML classifier are shown in Figure 1 for the individual samples. One can observe the enhancement in the Z' samples at the place of the expected top quark mass peak, the larger the higher the mass of the Z' particle, while the lower mass Z' sample provides enhanced region at the W boson mass. The various $t\bar{t}$ samples exhibit a large continuum of masses, with non-resonant bulk contribution below 60 GeV of different sizes due to different jet p_{T} kinematics cut for the samples.

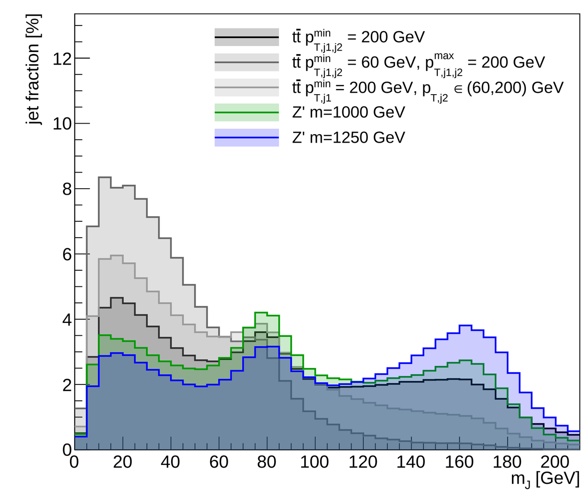

Figure 1: Shapes of the large-R jet mass variables in the five samples for training and testing.

2 ML-based top and W tagging

2.1 General data structure

Three samples corresponding to the SM $t\bar{t}$ production were generated, with different cuts at the generator level on the transverse momentum of the jets, in order to cover regions with various fractions of t, W as well as non-resonant (light) jets. These have been used as both training and testing data sets.

The two Z' samples with the Z' masses of 1000 and $1250\,\text{GeV}$ provide a $t\bar{t}$ sample with enhanced boosted top quarks, thus leading to events with enhanced fractions of t and W jets.

Variables defined and used for each jet in the classification are as follows

- $\Delta R(J, W)$, the minimal distance of the jet to the nearest W [1];
- $\Delta R(J, t)$, the minimal distance of the jet to the nearest top parton;
- Jet tr. momentum p_{T}^J and jet four-vector invariant mass m_J.
- η and ϕ of the jet.
- Jet substructure variables τ_{32} and τ_{21}.

The true type jets labels are based on the following criteria

1. truth t-jets: $\Delta R(J, t) < 0.25 \wedge 150\,\text{GeV} \leq m_J \leq 210\,\text{GeV}$;
2. truth W-jets; $\Delta R(J, W) < 0.25 \wedge 60\,\text{GeV} \leq m_J \leq 110\,\text{GeV}$;
3. truth light jets: otherwise.

For the predictions we used the machine learning (ML) model based on the Gradient Boosting technique, which is a popular and widely used algorithm for supervised learning, see e.g. [14]. This classifier is one of the two most used types of *ensemble methods*, which are methods combining multiple simple predictors (esp. decision trees) in order to create a more powerful model. The method does not work with weights but it tries to fit the predictor to the *residual errors* made by the previous predictor. The new prediction is made by simply adding up the predictions of all the predictors. We decided to test per-jet predicting of t-, W- or l-jets.

2.2 Results

Figure 2 displays the jet mass distribution, where the red filled areas represent top-tagged jets using both the cut-based (dark red) and ML approach (light red). Although the cut-based method exhibits higher tagging efficiency, it is evident that it suffers from a significant mistagging of light jets as tops. In contrast, the ML method appears to tag top jets with higher purity, and the boundary of the red filled area aligns with the background of light jets. To further support this argument, the real tagging efficiencies as a function of jet p_T are presented in Figure 3, where the ML approach demonstrates lower efficiency in all bins (solid red) compared to the cut-based approach (dashed red). Additionally, the ML method exhibits lower fake efficiency in the QCD background (blue lines). Figure 4 showcases the invariant mass of leading and subleading top-tagged jets for the Standard Model $t\bar{t}t\bar{t}$ process, alongside the stacked invariant mass of leading and subleading jets from the $y_0 t\bar{t} \to t\bar{t}t\bar{t}$ process (scaled by a factor of 0.075). The light filled areas and dashed lines represent the top-tagged jets that are matched to the parton top within $\Delta R < 0.15$. The efficiencies, as indicated in the legend of Figure 4, are provided for each set of histograms as

$$\epsilon = \frac{\text{top tagged jets \&\& matched to parton level}}{\text{top tagged jets}}. \quad (1)$$

[1]The angular distance between two objects is defines as $\Delta R \equiv \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}$ where the pseudorapidity $\eta \equiv -\ln\tan\frac{\theta}{2}$ is related to the standard azimuthal angle θ of the spherical coordinates, where the beam axis coincides with the z axis, and ϕ is the polar angle in the xy plane.

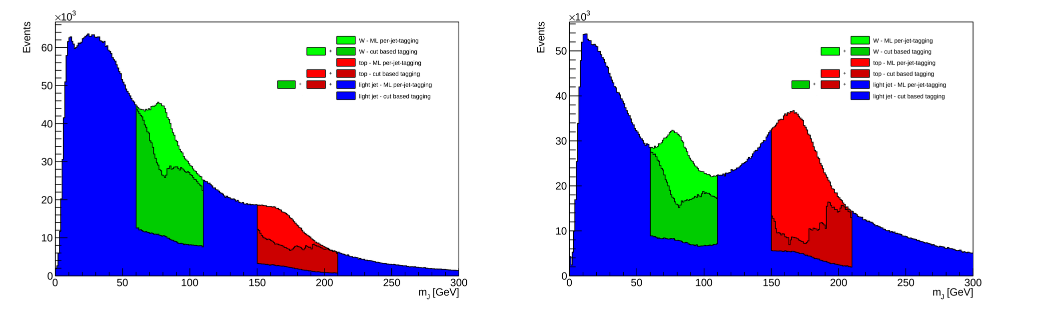

Figure 2: Jets mass of Standard Model $t\bar{t}t\bar{t}$ (left) and $y_0 t\bar{t} \to t\bar{t}t\bar{t}$ (right) with cut-based and maching learning tagging of W boson (green) and top (red).

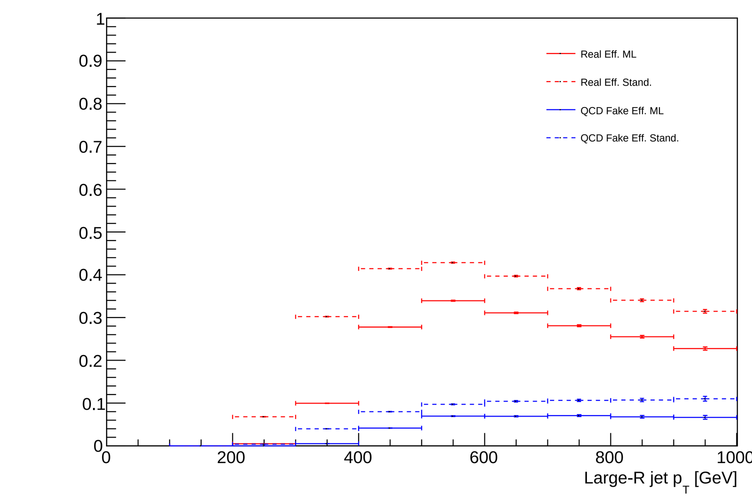

Figure 3: Real efficiencies of the top-tagging using cat-based (dashed red line) and ML approach (solid red line) and QCD background fake efficiencies using cut-based (dashed blue lines) and ML (solid blue line) .

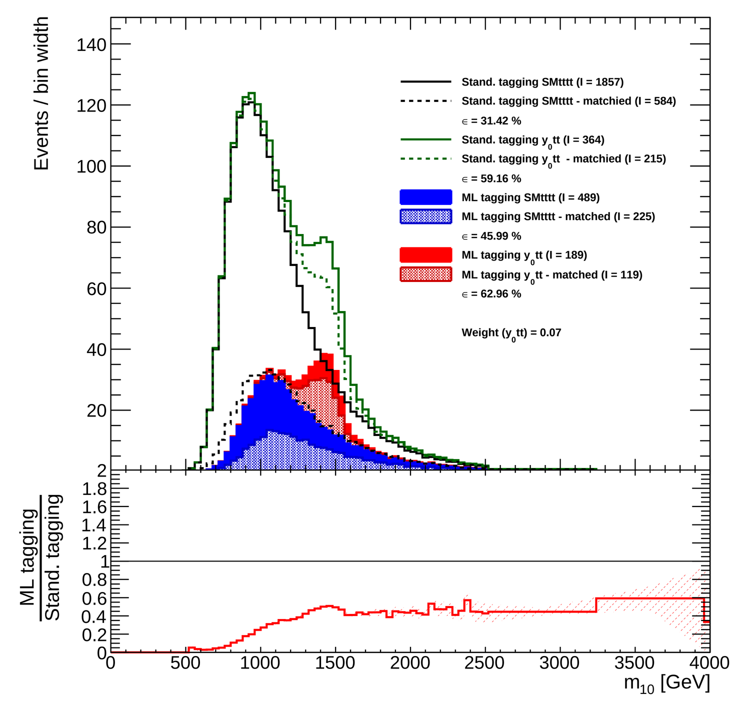

Figure 4: Invariant mass of leading and subleading jets of Standard Model $t\bar{t}t\bar{t}$ and $y_0 t\bar{t} \to t\bar{t}t\bar{t}$ ($m_{y_0} = 1.5\,\text{TeV}$) tagged as top quark. The light red / blue filled areas and dashed lines represent matched jets to top quark within $\Delta R < 0.15$.

Conclusion

The ML approach demonstrates a higher tagging efficiency compared to top-tagged jets matched at the parton level, with an improvement of 14.6% for the Standard Model $t\bar{t}t\bar{t}$ process and 3.8% for the $y_0 t\bar{t} \to t\bar{t}t\bar{t}$ process (with $m_{y_0} = 1.5\,\text{TeV}$) when compared to the cut-based selection.

References

[1] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *JHEP*, 07:079, 2014.

[2] Stefan Hoeche, Frank Krauss, Nils Lavesson, Leif Lonnblad, Michelangelo Mangano, Andreas Schalicke, and Steffen Schumann. Matching parton showers and matrix elements. In *HERA and the LHC: A Workshop on the Implications of HERA for LHC Physics: CERN - DESY Workshop 2004/2005 (Midterm Meeting, CERN, 11-13 October 2004; Final Meeting, DESY, 17-21 January 2005)*, pages 288–289, 2005.

[3] Michelangelo L. Mangano, Mauro Moretti, Fulvio Piccinini, Roberto Pittau, and Antonio D. Polosa. ALPGEN, a generator for hard multiparton processes in hadronic collisions. *JHEP*, 07:001, 2003.

[4] Torbjörn Sjöstrand, Stefan Ask, Jesper R. Christiansen, Richard Corke, Nishita Desai, Philip Ilten, Stephen Mrenna, Stefan Prestel, Christine O. Rasmussen, and Peter Z. Skands. An introduction to PYTHIA 8.2. *Comput. Phys. Commun.*, 191:159–177, 2015.

[5] Duhr C. FeynRules Implementation of Abelian Higgs Model. 2011. `https://feynrules.irmp.ucl.ac.be/wiki/HiddenAbelianHiggsModel`.

[6] Neil D. Christensen and Claude Duhr. FeynRules - Feynman rules made easy. *Comput. Phys. Commun.*, 180:1614–1641, 2009.

[7] James D. Wells. How to Find a Hidden World at the Large Hadron Collider. 2008.

[8] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi. DELPHES 3, A modular framework for fast simulation of a generic collider experiment. *JHEP*, 02:057, 2014.

[9] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. FastJet User Manual. *Eur. Phys. J.*, C72:1896, 2012.

[10] David Krohn, Jesse Thaler, and Lian-Tao Wang. Jet Trimming. *JHEP*, 02:084, 2010.

[11] Andrew J. Larkoski, Simone Marzani, Gregory Soyez, and Jesse Thaler. Soft Drop. *JHEP*, 05:146, 2014.

[12] Stephen D. Ellis, Christopher K. Vermilion, and Jonathan R. Walsh. Recombination Algorithms and Jet Substructure: Pruning as a Tool for Heavy Particle Searches. *Phys. Rev. D*, 81:094023, 2010.

[13] Jesse Thaler and Ken Van Tilburg. Identifying boosted objects with n-subjettiness. *Journal of High Energy Physics*, 2011(3), Mar 2011.

[14] Andreas G. Müller and Sarah Guido. *Introduction to Machine Learning with Python*. O'Reilly, Beijing Boston Farnham Sebastopol Tokyo, 2016.

Acknowledgement

The author would like to thank the grants of MSMT, Czech Republic, GACR 23-07110S for the support.