Frascati Physics Series Vol. 72 (2022)
7TH YOUNG RESEARCHERS' WORKSHOP 'PHYSICS CHALLENGES IN THE LHC ERA'
July 11 and 14, 2022

COLOURING THE JETS AT LHC FOR *Xbb* TAGGER IMPROVEMENT

Giulia Manco
Università degli studi di Pavia and INFN, Sezione di Pavia

Abstract

After 10 years from its discovery, the Higgs boson is still one of the most investigated particles. The Higgs decay in two b-quarks is very interesting and the most probable decay, but, due to the large QCD background, it is not straightforward to study. For this reason, the LHC community is investing in the direction of Xbb taggers (X=Z or Higgs boson), which aims at finding an optimal Higgs-tagger using jet substructure information. In this document, colour-sensitive variables will be studied as Xbb tagger, exploiting the different colour configuration of a colour-singlet and a colour-octect. Observable performances are tested on the $VHbb$ channel in the boosted limit.

1 Introduction

The Higgs boson was discovered in 2012 at LHC by the ATLAS and CMS Collaborations [1) 2)]. Since then, the high energy physics community has been involved in the measurements of its proprieties. The Higgs boson gives the opportunity to test the Standard Model (SM) predictions and discover new physics. In particular, the coupling of the Higgs particle is the only interaction that can feel the difference between fermion generations.

At a Higgs boson mass of 125 GeV, the most probable decay is in two b quarks, with a branching ratio of about 58%. The direct measurement of the $b\bar{b}$ channel provides a test of the Yukawa coupling to a down-type quark and constrains the overall Higgs decay width. While this decay is the most frequent, it is a real experimental challenge to observe it. This is due to the overwhelming large QCD background that can mimic the signal signature. For these reasons it took six years until ATLAS and CMS obtained the necessary 5σ significance for the evidence of this decay channel [3) 4)]. The production mode used in these analyses is Higgs-boson (H) production in association with a vector boson V (W or Z), with V decaying

leptonically and the Higgs hadronically into a pair of b-quarks, which provides a clean experimental signature. The hard b-quarks produced by the Higgs boson decay are usually detected as two separate b-jets. When the momentum of the jets is higher than their invariant mass, the regime is called *boosted*. In such a situation, the two b-jets are close in angle and hence reconstructed as a single jet, also known as a large-radius jet.

In order to better discriminate the $H(b\bar{b})$ process over the production of b-jets from a gluon collinear splitting ($g \to b\bar{b}$), many strategies have been developed. Several jet substructure techniques have been designed, which aim at improving the discrimination performance by finding hard prongs inside the large-radius jet. Specifically, the different radiation pattern of signal and background can be exploited. In the signal case, the b-jets originate from a colour singlet and the radiation is more constrained inside the two b-quark system. In the background case, the radiation is more diffuse, due to the colour connection with the initial state, as shown in Figure 1.

In this paper, observables sensitive to the different colour configuration will be exploited, referring to this recent article [7]. The idea is build a tagger that can be applied to the decay products of a generic colour singlet X. In this regards, the Xbb tagger group in ATLAS aims at providing recommendations for $H \to b\bar{b}$ tagging and tools for its use within analysis. It is an activity which involves the investigation of both jet substructure and b-tagging performance in boosted $H \to b\bar{b}$ topologies [8]. The identified tagger with colour-sensitive variables is matter of interest of this group.

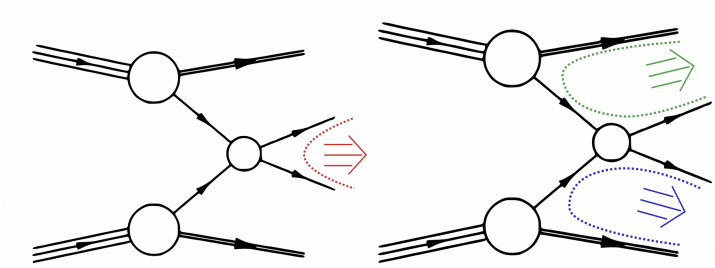

Figure 1: *Possible colour connections for the signal on the left ($pp \to H \to b\bar{b}$) and for the background on the right ($pp \to g \to b\bar{b}$)* [5].

2 Observables

In the following, a selection of high-level colour-sensitive variables are presented. They were introduced in the literature in the past few years.

2.1 Jet Pull

Let us consider a hard jet J_a. The *pull vector* $\vec{t}$ is the jet shape observable defined as: [5]

$$\vec{t} = \frac{1}{p_{Ta}} \sum_{i \in J_a} p_{Ti} |\vec{r}_i|^2 \hat{r}_i, \tag{1}$$

where p_{T_a} is the transverse momentum of the jet, and the sum runs over all the the jet constituents. y and ϕ represent rapidity and azimuthal angle and $\vec{r}_i$ is the distance vector between the jet axis and its

i-th constituent in the y-ϕ plane

$$\vec{r}_i = (y_i - y_a, \phi_i - \phi_a). \tag{2}$$

The pull vector is sensitive to the different colour connections of the event and points toward the direction of emitted radiation.

We can introduce the projections of the pull vector along the direction between the two jets $t_\parallel$ and in the perpendicular direction $t_\perp$ [10, 11]. We also consider the pull angle θ_p defined as [9]:

$$\theta_p = \arccos \frac{t_\parallel}{|\vec{t}|}. \tag{3}$$

2.2 Jet colour ring

The jet colour ring [11] is defined from the ratio of the squared matrix elements of signal and background, where the signal is considered as the decay of a colour singlet and the color octect is the background. In the soft-collinear limit approximation, such a ratio becomes:

$$\mathcal{O} = \frac{\Delta_{ak}^2 + \Delta_{bk}^2}{\Delta_{ab}^2}, \tag{4}$$

where Δ_{ij} are the distances between jets (or subjets) in the azimuth-rapidity plane, a is the leading jet, b the subleading jet and k a soft emission. The observable name originates from its geometric interpretation: radiation from colour singlets will tend to fall between the two jets, leading to values of $\mathcal{O} < 1$, while in the case of colour octets, one will tend to have $\mathcal{O} > 1$.

2.3 D_2

The variable D_2 [12] is defined as the ratio of two normalized N-point energy correlation functions (ECFs) [6], e_k^β:

$$D_2^{(\beta)} = \frac{e_3^{(\beta)}}{(e_2^{(\beta)})^3}. \tag{5}$$

β is a parameter which we have set to $\beta = 2$. The variable is usually calculated on a large radius jet, and is useful to discriminate 2-prong jets from 1-prong jets.

2.4 Lund jet plane

The Lund jet plane is defined in reference [13]. It is formed by parsing backwards the Cambridge–Aachen (C/A) clustering history of the jet. The procedure starts by undoing the final clustering step and by recording the kinematics of the splitting. The primary Lund jet plane is obtained by iterating the above procedure, always following the hardest branching in each splitting and recording the azimuth-rapidity separation of the branchings involved in the splitting and the relative transverse momentum of the emission.

3 Observable performances on *VHbb* channel

3.1 Event simulation and selection

In order to test the observable discrimination performance, 300k events for $pp \to H(b\bar{b})Z(\nu_\ell\overline{\nu}_\ell)$ signal and 4M events for the $pp \to b\bar{b}\nu_\ell\overline{\nu}_\ell$ background processes are generated. Number of events are chosen in order to have 50k events for both signal and background, accounting for the efficiency after applying

Table 1: *Percentage of events which pass the analysis selections.*

	Truth	Reco
Signal	20%	17%
Background	1.6%	1.3%

Table 2: *Area under the ROC curves for different combination of observables.*

	Truth	Reco
CS observables	0.826	0.788
D_2+CR	0.817	0.787
$\mathrm{LP_{CNN}}$	0.876	0.828
CS + $\mathrm{LP_{CNN}}$	0.893	0.846

the selection cuts, shown in Table 1. Hard events are generated with `MG5_aMC@NLO v2.8.3.2` [14] in a boosted regime and parton-level events are then showered in `Pythia` v8.305 [15]. Detector effects are considered with a fast detector simulation of `Delphes` v3.5.0 [16]. From `Delphes`, the Monte Carlo truth is extracted, containing the particle-level information. Reference [7] gives a complete description of analysis selection and simulation used here.

3.2 Discrimination performance

In Fig. 2 the normalised distributions for eight colour sensitive variables (CS) are shown, both for signal and background, and at truth and reco level. Looking at the plots, the discrimination power of D_2 and $\mathcal{O}$ can be appreciated and the detector effects, in particular on pull variables, can be observed. In Fig. 2 the average Lund images for the signal and background processes in the truth and reco case are presented. From the plot, it is possible to appreciate the detector effect on the images, which adds in the reco case a radiation for the middle values of Δ and k_t. After having determined the distributions of the CS observables and the Lund jet images, these are used as inputs to ML algorithms in order to build combined classifiers. Specifically, a Boosted Decision Tree (BDT) is trained on the CS observables, whereas Lund images are classified using a Convolutional Neural Network (CNN). The output distribution of CNN Lund jet plane classifier is shown in Figure 2. More details about these methods and architectures are provided in [7]. Different combinations of variables are also considered in order to improve the total discrimination power. In this case the procedure is in two steps and uses the CNN Lund jet plane classifier as an additional input to the BDT.

In Fig. 3 the receiver operating characteristic (ROC) curves for several combinations of observables are shown. The background rejection ($1/\epsilon_b$) vs the signal efficiency (ϵ_s) is presented: the higher the curve, the better the discriminant power. Namely, we have considered: all the colour-sensitive observables (CS) or just the D_2 and the colour ring (D_2+CR), combined through a BDT; the CNN Lund jet plane classifier ($\mathrm{LP_{CNN}}$); the combination of all the CS observables with the (CS+$\mathrm{LP_{CNN}}$), by means of the two-step procedure explained above. For each curve in Fig. 3, the value of the area under the ROC curve (AUC) is reported in Table 2.

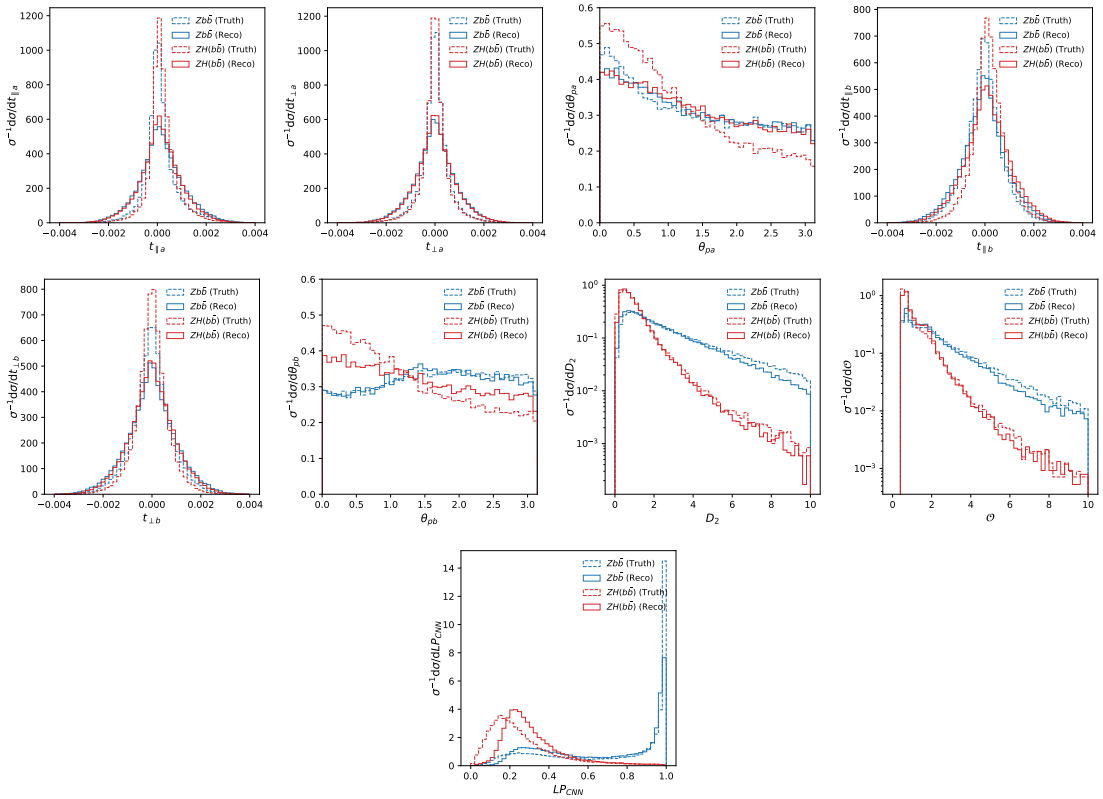

Figure 2: *Observables for signal and background, truth and reco cases, as defined in Section 2* [7].

3.3 Results

As expected, the performances are worse in the reco case, due to detector resolution. However, discrimination is still good for most combinations, close to 0.85 for CS + $\mathrm{LP_{CNN}}$. It is evident that most of the discriminating power of CS is due to D_2+CR alone, both in AUC values and in distributions. It is clear that pull variables are not as powerful in discrimination as the other variables. Moving to combination with Lund jet plane, Lund jet plane alone performs better than the whole set of CS observables. When $\mathrm{LP_{CNN}}$ is combined with CS observables, there is a noticeable improvement of the overall classification power, with a value of AUC equal to 0.893 in the truth case and 0.846 in the reco case.

4 Conclusions

In this paper, the problem of finding a Xbb tagger, namely how to distinguish the b-jets originating from a colour singlet, such as Higgs boson, from those originating from the QCD background is investigated. Colour-sensitive observables present in literature are exploited in combination in order to perform a powerful discriminator. These observables are tested on the signal process $pp \to H(b\bar{b})Z(\nu_\ell\overline{\nu}_\ell)$, but the strategy can be valid in a more general context. The discrimination performance is estimated using ML techniques, namely BDT and CNN architecture. The BDT is trained with the colour-sensitive variables, including the Lund jet plane CNN discriminator. The results are encouraging, with a power in discrimination of 0.893 AUC for the combination of CS + $\mathrm{LP_{CNN}}$.

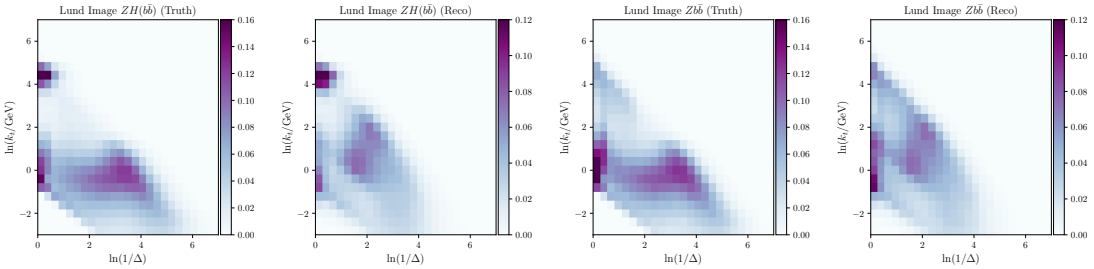

Figure 3: *Averaged primary Lund jet plane images for $ZH(b\bar{b})$ and $Zb\bar{b}$ in the truth and reco case* [7].

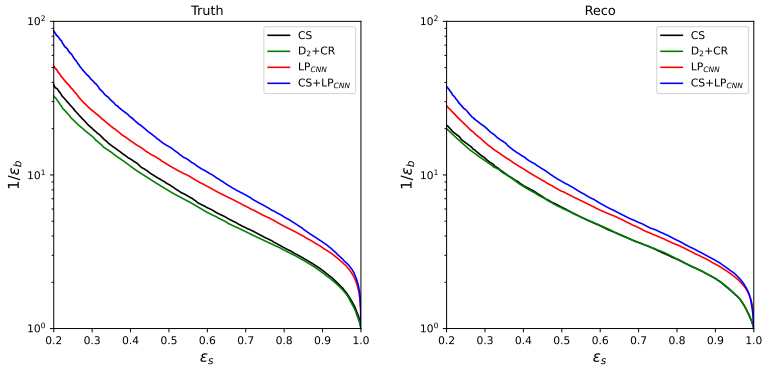

Figure 4: *The ROC curves showing background rejection as a function of signal efficiency for the truth (left) and reco case (right) for CS variables, LP_{CNN} and the combined cases* [7].

In the end, this tagger, which is a combination of several theory-driven single-variable observables with a representation of the radiation pattern within a jet, is not only effective in theory, but also shows promising prospects for application to experimental analyses.

5 Acknowledgements

The author acknowledges the University of Pavia, in particular her supervisor Daniela Rebuzzi, for the opportunity to attend the Frascati Summer School. We also thank the authors of the publication which inspired this paper [7]: Luca Cavallini, Andrea Coccaro, Charanjit Khosa, Simone Marzani, Fabrizio Parodi, Daniela Rebuzzi, Alberto Rescia and Giovanni Stagnitto. In the end, the author acknowledges the organizers of the LNF 'Bruno Touschek' Summer School for the opportunity to present this work.

References

1. ATLAS Collaboration, G. Aad et al., *Phys. Lett. B* **716**, 1 (2012).
2. CMS Collaboration, S. Chatrchyan et al., *Phys. Lett. B* **716**, 30 (2012).
3. ATLAS Collaboration, M. Aaboud et al., *Phys. Lett. B*, **786**, 59 (2018).
4. CMS Collaboration, A.M. Sirunyan et al., *CMS-PAS-HIG-18-016*, (2018).

5. J. Gallicchio and M.D. Schwartz, *Phys. Rev. Lett.* **105**, 022001 (2010).

6. A.J. Larkoski, G.P. Salam and J. Thaler, *JHEP* **2013**, 108 (2013).

7. L. Cavallini, A. Coccaro, C.K. Khosa, G. Manco, S. Marzani, F. Parodi, D. Rebuzzi, A. Rescia and G. Stagnitto, *Eur. Phys. J. C*, **82**, 493 (2022).

8. ATLAS Collaboration, *ATL-PHYS-PUB-2020-019*, (2020).

9. A.J. Larkoski, S. Marzani, C. Wu, *Phys. Rev. D* **99**, 091502 (2019).

10. Y. Bao and A.J. Larkoski, *JHEP* **12**, 035 (2019).

11. A.J. Larkoski, S. Marzani and C. Wu, *SciPost Phys.* **9**, 026 (2020).

12. I. Moult, L. Necib and J. Thaler, *JHEP* **12**, 153 (2016).

13. F.A. Dreyer, G.P. Salam and G. Soyez, *JHEP* **12**, 064 (2018).

14. J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer et al., *JHEP* **07**, 079 (2014).

15. T. Sjstrand, S. Ask, J.R. Christiansen, R. Corke, N. Desai, P. Ilten et al., *Comput. Phys. Commun.* **191**, 159 (2015).

16. DELPHES 3 Collaboration, *JHEP* **02**, 057 (2014).