IL NUOVO CIMENTO **47 C** (2024) 112
DOI 10.1393/ncc/i2024-24112-2

Colloquia: IFAE 2023

Colorflow Studies in the ATLAS Boosted Higgs Boson Tagger(*)

M. D'Andrea(1)(2) on behalf of the ATLAS Collaboration

(1) *Istituto Nazionale di Fisica Nucleare (INFN), Sezione di Roma 1 - Rome, Italy*
(2) *University of Rome, "La Sapienza" - Rome, Italy*

received 13 February 2024

Summary. — Enhancing the discrimination power to identify Higgs boson decay events into a pair of bottom quarks in the boosted regime is of significant importance within the ATLAS experiment at the Large Hadron Collider (LHC). Effectively identifying this specific process enhances the collected sample of Higgs boson production events, hence reducing the statistical uncertainties in the determination of the properties of this particle. The production of the Higgs boson in the so-called boosted regime, *i.e.*, with large transverse momentum with respect to the beam direction, results in the final-state quark pair being highly collimated, reconstructed as a single large-radius jet (LargeR-jet). Discriminating between these signal events and the background is a rather challenging task. To enhance the discriminative power between signal and background events, it is possible to employ variables that are sensitive to the SU(3) color representation of the decaying particle, which produces the LargeR-jet. This study has demonstrated how the new ATLAS boosted Higgs boson tagger (GN2X) appears to autonomously utilize color information from the jet, showcasing its robustness, thus offering valuable insights for future analyses within the collaboration focusing on such events.

1. – Introduction

Within the ATLAS collaboration at the LHC, a recent development includes a Machine Learning algorithm known as GN2X, as referenced in [1]. This algorithm is designed for the identification of the Higgs boson decay into a heavy-flavored quark-antiquark pair in the boosted decay regime.

These boosted events give rise to two closely aligned jets, often referred to as subjets, which are subsequently reconstructed into a single jet with a radius defined as $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2} = 1$(1), known as a "Large-$R$" jet. The jet opening angle, originating from the highly boosted Higgs boson decay, can be assessed in this limit as $\Delta R = \frac{2m_H}{P_T}$; therefore, it is imperative to have a radius of $\Delta R = 1$ for a Higgs boson with high transverse momentum to encompass the complete dynamics of the process.

(*) IFAE 2023 - "Poster" session
(1) Where the pseudorapidity, η, is defined as $-\log\left(\tan\frac{\theta}{2}\right)$ while ϕ is the azimuthal angle in the ATLAS detector coordinate system.

The identification model is built upon a graph neural network designed to discern the flavour of a jet originating from the fragmentation of a quark. In particular, the algorithm GN1, originally developed for tagging the flavour of such single jets [2], was tailored to discriminate Large-R jets formed by $H \to b\bar{b}(c\bar{c})$ from background processes (primarily QCD processes such as $g \to b\bar{b}(c\bar{c})$). Specifically, the algorithm takes in the kinematic parameters of the tracks left by the charged particles traversing the ATLAS Inner Detector [3].

Through these features, the model is trained to recognize and effectively utilize the dynamics of quark hadronization and the decay in the final state of the process. Notably, it focuses on the discrimination of flavours, bypassing the use of global jet parameters like mass and transverse momentum for signal-to-background differentiation.

Recent studies have been conducted regarding the usage of kinematic variables of jets that are sensitive to the $SU(3)$ color representation of the particle giving rise to the jet, as described in [4]. These variables, in principle, can be employed to enhance signal-to-background discrimination within the ATLAS GN2X algorithm.

2. – Color Flow

Color Flow pertains to the examination of the flow of color charge of particles (charged and neutral) under the symmetry group of $SU(3)_c$ during a high-energy physics process. At the LHC, hard interactions between partons contained within protons generate a flow of color charge. The analysis of this color flow is important both for a better understanding of the theory of strong interactions, thereby extending our knowledge to non-perturbative regimes of QCD, and for identifying characteristic processes that exhibit specific behavior from the perspective of this phenomenon.

Recent theoretical studies, as referenced in [5], have revealed how the color representation of a particle decaying into two quarks in the final state partially influences the topological distribution of jets generated during the parton shower process. Specifically, if an uncolored particle, residing in a singlet representation of $SU(3)c$, decays, the resulting two jets exhibit a higher degree of collimation compared to the scenario where a particle carrying color charge under $SU(3)c$ decays.

This characteristic can be exploited to enhance the discrimination power between the hadronic decays of the Higgs boson (a singlet under $SU(3)$) and the gluon (a particle carrying color charge). Particularly, this distinction is valuable in distinguishing between processes like $H \to b\bar{b}$ and $g \to b\bar{b}$.

To capture information regarding the color flow of the decay, it is essential to consider variables sensitive to the dynamics of color flow. These variables are extracted during the jet clustering process, where the tracks of particles traversing the detector are grouped to define the jet. In this specific case, the clustering process is entrusted to the *Cambridge-Aachen* (C-A) algorithm, which groups tracks into clusters, also referred to as pseudojets, by minimizing the distance in the y-ϕ plane. Here, y represents rapidity, defined as:

$$y = \frac{1}{2} \ln \frac{E + p_z}{E - p_z} \xrightarrow[\beta=1]{} \eta = -\ln \tan \frac{\theta}{2}. \tag{1}$$

At each step of track/pseudojet clustering, color variables are extracted. These primarily include the distance (Δ) in the y-ϕ plane between the two grouped entities and the transverse momentum of the softer cluster relative to the harder one (k_t).

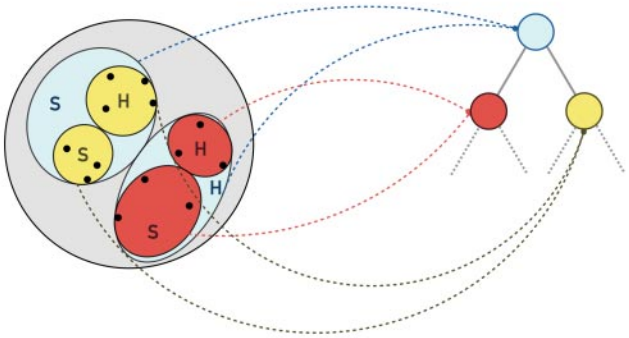

Fig. 1. – Illustration of the graph construction during the track (black dots) clustering of the jet. The colors represent a clustering step, while the labels "S" and "H" refer to which cluster is "soft" and "hard" in terms of energy in the given process.

During this phase, it is possible to construct a graph where each node represents a step of clustering and thus serves as a container for the aforementioned color variables as shown in fig. 1.

These graphs can be used as input for a Graph Neural Network, which outputs an inference regarding whether a graph (representing a jet) is generated by the decay of a colorless particle or a particle with color charge under $SU(3)$. This task is assigned to the neural network known as LundNet, as referenced in [6].

3. – GN2X with Color Information

As initial approach to incorporate color information into the current ATLAS tagger, the output of LundNet was utilized to tag Large-R jets, which were subsequently used to train and evaluate the model.

Specifically, LundNet was trained using two million jets, equally split between signal and background. This model was evaluated on a sample of seven million events (comprising both signal and background), to which the respective LundNet inferences were appended. These events were then employed to train a GN2X model with jets containing color information, followed by an assessment of its performance.

The results of this network test are presented in fig. 2.

The graphs in fig. 2 illustrate the variation in background event rejection concerning the efficiency of correctly tagging signal events. These curves can be analyzed at fixed signal efficiency values. For instance, at a value of 0.75 for the signal-jet tagging efficiency, the GN2X models and GN2X with color information reject approximately 100 QCD background events, whereas LundNet rejects around four events. These numbers are obtained for the three models trained and evaluated through the approach described earlier. In particular, the violet curve represents the performance of the GN2X model([2]), the green curve depicts GN2X trained with the color information derived from the LundNet output, and the blue curve represents LundNet operating independently.

([2]) In fig. 2, the label GN1Xbb is used because it was the name initially used during the early stages of development, later replaced by GN2X.

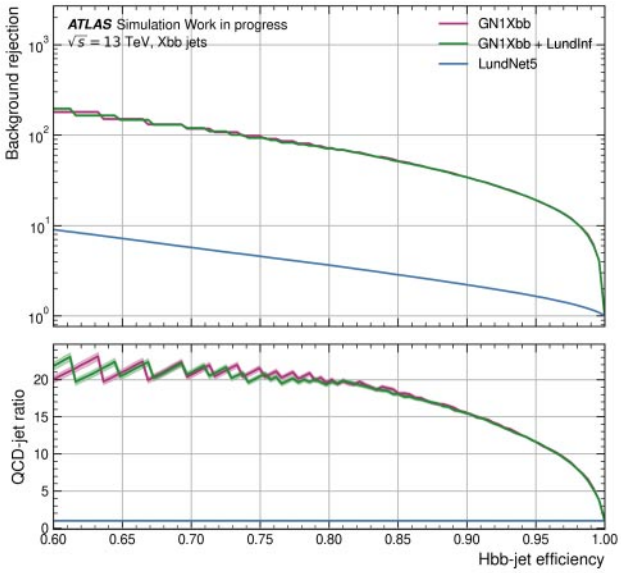

Fig. 2. – The graph displays the background rejection as a function of signal efficiency for the GN2X models (GN1Xbb), GN2X with color information (GN1Xbb + LundInf), and LundNet (LundNet5). The graph also includes a ratio plot comparing the GN2X and LundNet models, with the latter serving as the baseline.

As evident from fig. 2, the rejections for GN2X and GN2X with color information are similar. This nearly identical behavior between GN2X and GN2X with color information suggests that the network is already leveraging the discriminative power of color variables, even without these variables being explicitly provided as input. This highly non-trivial behavior underscores the robustness of the network developed by the ATLAS Flavour Tagging group, which holds profound implications and relevance for the current Run 3 analysis at the LHC.

REFERENCES

[1] ATLAS Collaboration, *Transformer Neural Networks for Identifying Boosted Higgs Bosons decaying into $b\bar{b}$ and $c\bar{c}$ in ATLAS*, ATL-PHYS-PUB-2023-021.
[2] ATLAS Collaboration, *Graph Neural Network Jet Flavour Tagging with the ATLAS Detector*, ATL-PHYS-PUB-2022-027.
[3] ATLAS Collaboration, *JINST*, **3** (2008) S08003.
[4] Cavallini Luca *et al.*, *Eur. Phys. J. C*, **82** (2022) 493.
[5] Buckley Andy *et al.*, *Sci. Post. Phys.*, **9** (2020) 2.
[6] Dreyer Frédéric A. and Qu Huilin, *J. High Energy Phys.*, **2021** (2021) 52.