IL NUOVO CIMENTO **48 C** (2025) 160
DOI 10.1393/ncc/i2025-25160-8

COLLOQUIA: IFAE 2024

Boosted jet identification at the CMS experiment(*)

D. TROIANO(1)(2) on behalf of the CMS COLLABORATION

(1) *INFN, Section of Bari - Bari, Italy*
(2) *Physics department, University of Bari - Bari, Italy*

received 2 December 2024

Summary. — A fundamental aspect of the CMS experiment's research concerns the identification of jets produced in high energy proton-proton collisions. W and Z bosons and the Higgs boson can be produced with a high Lorentz boost and, under such conditions, their decay products can be reconstructed as large radius jets, *i.e.*, anti-k_t jets of radius 0.8 (AK8). The identification of the particle, which initiates the large radius jet, therefore plays a crucial role in distinguishing boosted particles from the dominant QCD background. Several AK8 identification algorithms, based on sophisticated machine learning techniques, have been developed by the CMS collaboration. Here an overview of them in terms of their performance and use within the collaboration will be provided.

1. – Jet flavor-tagging at CMS

Jets are the experimental signatures of the production of quarks and gluons in high energy processes, typically detected as collimated sprays of particles that are clustered together. At the Large Hadron Collider (LHC) [1] collision energies, electroweak resonances (W/Z bosons) and Higgs bosons are often produced with a high transverse momentum (p_T), as consequence their decay products become collimated and result in one large and massive jet (fat jet). Investigating the internal structure of fat jets allows for separating the process of interest from the large QCD multijet background. Currently jet reconstruction at the Compact Muon Solenoid (CMS) experiment [2] is based on clustering algorithms which are applied to either reconstructed particle tracks, calorimeter clusters or particle candidates reconstructed with the Particle Flow (PF) approach [3]. The most commonly adopted algorithm is the anti-k$_t$ algorithm [4]. In particular, AK8 (AK4) jets are reconstructed with the anti-k$_t$ algorithm with a radius of 0.8 (0.4). Since the identification of jets from radiation and hadronization of b- or c-quarks is of particular interest, tagging algorithms have been designed to identify jets originated from heavy-flavored quarks. Recent developments in this field are based on Deep Neural Networks to exploit the full potential of the CMS detector and event reconstruction, combining high level

(*) IFAE 2024 - "Poster" session

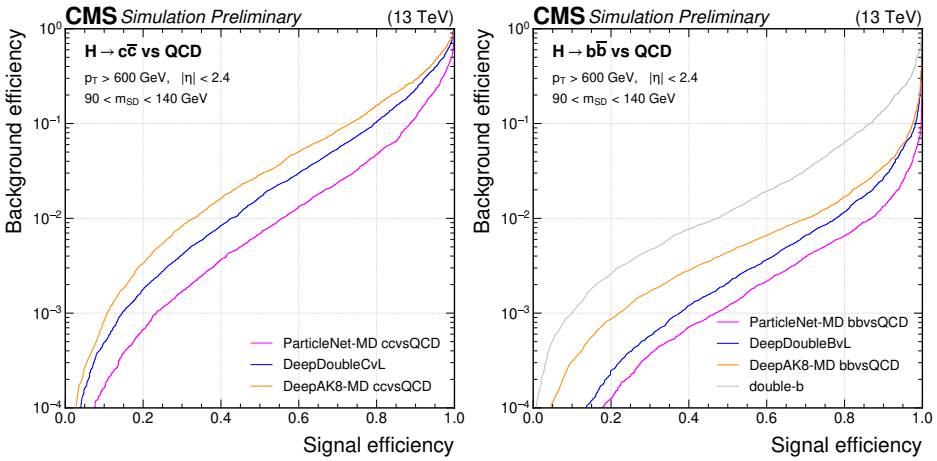

Fig. 1. – Comparison of the performance of the $X \to c\bar{c}$ and $X \to b\bar{b}$ (where X is a Lorentz-boosted spin-0 particle) identification algorithms in terms of receiver operating characteristic curves for $H \to c\bar{c}$ and $H \to b\bar{b}$ signal jets *versus* the inclusive QCD jets as background, respectively on the left and on the right, using simulated events in the 2018 data-taking conditions [7].

inputs (*e.g.*, jet substructure observables) with PF candidates and secondary vertexes. The efficiency and performance of ML techniques on jet physics rely heavily on how a jet is represented. One of the latest and best performing CMS tagging neural network architectures, called ParticleNet (PN) [5], represents fat jets as particles clouds, where each jet is considered as an unordered, permutation-invariant set of particles. This strategy permits the inclusion of any feature for each particle, thus rendering it considerably flexible.

1˙1. *ParticleNet.* – PN is a customized neural network architecture that operates directly on particle clouds for jet tagging. However, standard convolution operations cannot be utilized for point clouds due to the non-uniform distribution of points. Consequently, a new approach, with a convolution able to focus on the "local patch" of each point where the convolution kernel operates and maintaining the permutation symmetry of the point clouds, is required. To represent a jet as a graph, by exploiting the potentials derived from the representation as a particle cloud, the PN architecture uses three EdgeConv [6] blocks. After generating edges to describe the relationships between each jet constituent and its neighbors, each EdgeConv block updates the features of each jet constituent by aggregating the features of the k-nearest neighbors. After the EdgeConv blocks a softmax function is used to generate the output for the jet identification.

1˙2. *ParticleNet-MD.* – Using the PN architecture, ParticleNet-MD is a mass-decorrelated (no dependence on the mass of the AK8 jet) particle identification algorithm designed for identifying two-prong hadronic decays of highly Lorentz-boosted resonances X with four probability-like output scores: p($X \to b\bar{b}$), p($X \to c\bar{c}$), p($X \to q\bar{q}$), and p(QCD). Trained on CMS Run 2 Monte Carlo (MC) simulated events, PN-MD shows the best performance among the mass-decorrelated tagging algorithms. Figure 1 shows that, for the same signal efficiency, ParticleNet-MD has a lower background efficiency than the other mass-decorrelated algorithms. Since the detector conditions have changed,

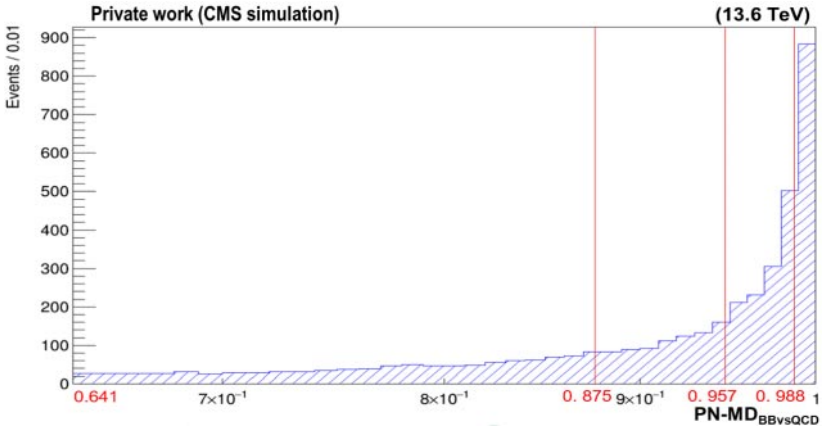

Fig. 2. – Z-boson candidate PN-$\text{MD}_{\text{BBvsQCD}}$ distribution from 0.641 to 1 for the MC $Z \to q\bar{q}$ after the event selection, normalized to 34.4 fb^{-1}. For each event, the Z-boson candidate is compatible with a Z decaying to $b\bar{b}$ at generation level. Red lines delimit the four highest score regions.

ParticleNet-MD has been retrained for the CMS Run 3. In this proceeding the validation of the PN-$\text{MD}_{\text{BBvsQCD}}$ score, defined as $\text{p}(X \to b\bar{b})$ / $[\text{p}(X \to b\bar{b})$ + p(QCD)], will be shown for 2022 data. The validation consists of the comparison of the PN-$\text{MD}_{\text{BBvsQCD}}$ distribution of the simulated MC $Z \to b\bar{b}$ + jets with the one obtained from data, subtracting the QCD background.

2. – ParticleNet-MD score validation

2˙1. *The boosted jet event selection.* – At the CMS High-Level Trigger level [8], selections require at least one of the following conditions: the scalar sum of jet transverse momenta (H_T) > 1050 GeV, a jet with p_T > 500 GeV, a AK8 jet with p_T > 400 GeV and mass (m) > 30 GeV, H_T > 800 GeV and a AK8 jet with m > 50 GeV.

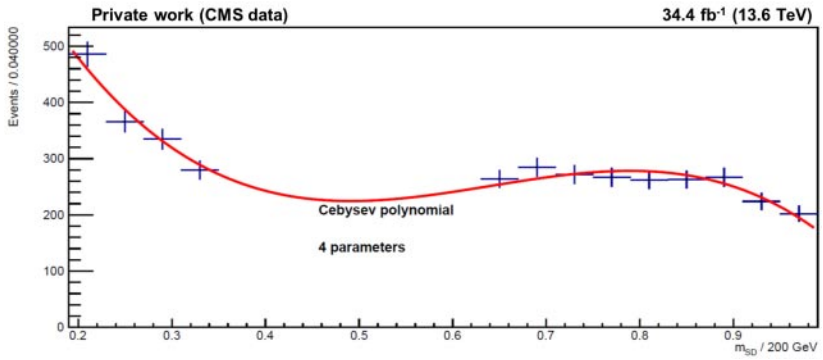

Fig. 3. – 2022 data for 0.988 < PN-$\text{MD}_{\text{BBvsQCD}} \leq 1$, fitted with a 3^{rd} order Cebysev polynomial (red line).

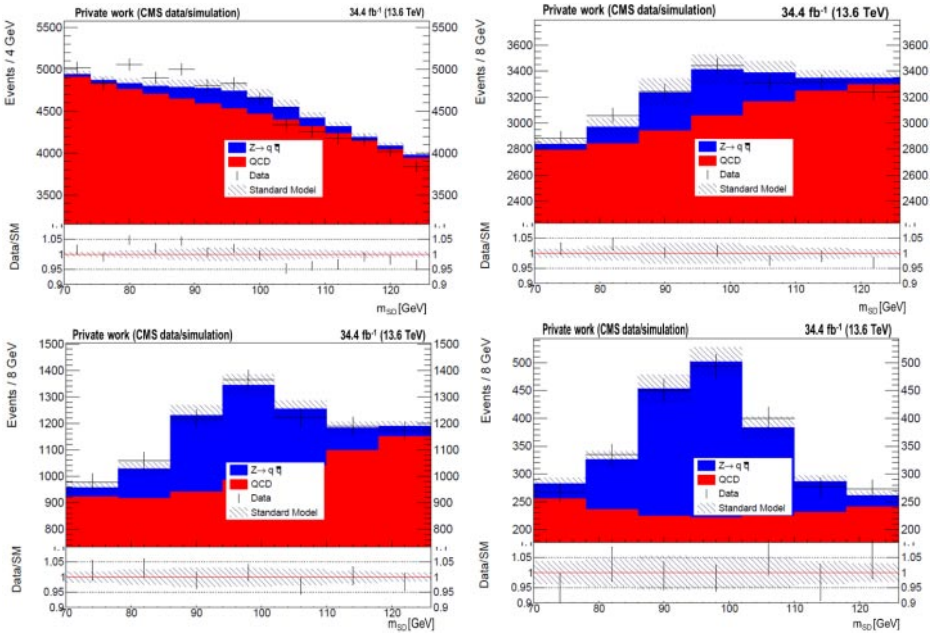

Fig. 4. – Stacked histograms showing $Z \to q\bar{q}$ and QCD m_{SD} distribution compared with the data after the likelihood fit for the fourth (top left), third (top right), second (bottom left) and first (bottom right) highest score regions. The ratio between data and the standard model prediction ($Z \to q\bar{q}$ + QCD) is reported in correspondence of each m_{SD} distribution.

Events passing the trigger are further required to pass the offline selection. It requires that the leading AK8 jet in p_T, identified as the Z-boson candidate, has $p_T > 450$ GeV and $|\eta| < 2.4$, and the subleading AK8 jet, identified as the recoil jet, has $p_T > 200$ GeV and $|\eta| < 2.4$. To suppress the $t\bar{t}$ background, which has leptons and extra quarks in the final state, events with at least one electron or muon with $p_T > 20$ GeV, $|\eta| < 2.4$, and satisfying the loosest identification and isolation working point [9,10] and events with a b-tagged AK4 jet satisfying $p_T > 30$ GeV and ΔR > 0.8 with respect to the Z-boson candidate are vetoed. PN-MD$_{\mathrm{BBvsQCD}}$ of the Z-boson candidate is used to separate events into five score regions. These have been chosen so that in each one there are approximately the 20% of the MC $Z \to b\bar{b}$ events. Among the five regions containing the 20% of the events, the one with the lowest score was not included in the validation because in the region the $Z \to b\bar{b}$ contribution in data is negligible compared with the QCD one. The PN-MD$_{\mathrm{BBvsQCD}}$ thresholds are 0.641, 0.875, 0.957, 0.988, 1. The four highest score regions are shown in fig. 2.

2˙2. *QCD estimate.* – Since the MC QCD sample does not yield a sufficient description of the data, the QCD is obtained with a data-driven technique using the soft-drop mass (m_{SD}) of the Z-boson candidate [11]. For each score bin the QCD yield is estimated in the signal mass window ($70 < m_{SD} < 126$ GeV) by fitting the data shape (which are mostly composed of QCD) with Cebysev polynomials in the mass sidebands ($38 < m_{SD} < 70 \cup 126 < m_{SD} < 198$ GeV). The polynomial order is established by means of the Fisher test with confidence level of 5%. An example of data sideband fitted is shown in fig. 3.

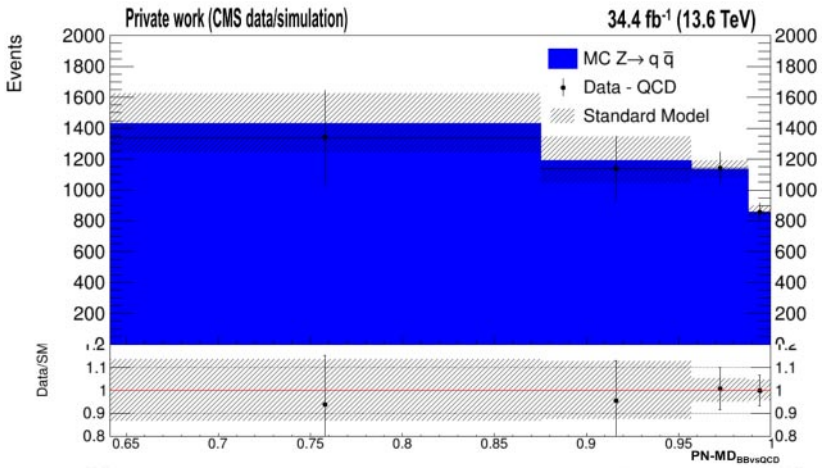

Fig. 5. – Z-boson candidate PN-$MD_{BBvsQCD}$ distribution of the $Z \to q\bar{q}$ yield, and the data, once the QCD contribution is subtracted (Data - QCD). In the bottom pad, the ratio between Data-QCD and the signal is shown. A good data-MC agreement is observed.

TABLE I. – *MC $Z \to q\bar{q}$ normalization factors at the four highest score regions.*

	r_Z
$0.641 <$ PN-$MD_{BBvsQCD} \leq 0.875$	1.3 ± 0.3
$0.875 <$ PN-$MD_{BBvsQCD} \leq 0.957$	1.3 ± 0.4
$0.957 <$ PN-$MD_{BBvsQCD} \leq 0.988$	1.31 ± 0.15
$0.988 <$ PN-$MD_{BBvsQCD} \leq 1$	1.04 ± 0.09

The QCD has two systematic uncertainties which take into account the uncertainties in the fit parameters and the reproducibility of the fit itself.

2˙3. *Analysis strategy.* – To evaluate the correct scalings of the MC templates, a likelihood fit [12] for the Z-boson candidate m_{SD} distribution has been done assuming as parameter of interest the MC $Z \to q\bar{q}$ normalization factor (r_Z), independently for each of the four highest score regions. The uncertainties included in the likelihood fit are the statistical uncertainty and the jet energy corrections for the MC $Z \to q\bar{q}$, the QCD uncertainties and the luminosity uncertainty. All the uncertainties, with the exception of the luminosity one, are assumed uncorrelated in the different score regions. Figure 4 shows for the four highest score regions the comparison between the stacked $Z \to q\bar{q}$ and QCD m_{SD} distribution and the data after the likelihood fit. After the likelihood fit there is a good data-simulation agreement. The Z peak becomes more visible as the score increases, which means that the tagger is able to discriminate the signal from the background.

3. – Results

The parameters, r_Zs, with uncertainties are reported in table I for each of the four highest score regions. In each score region, r_Z is compatible with unity within uncertainties. Figure 5 shows the comparison between the MC $Z \to q\bar{q}$ PN-$MD_{BBvsQCD}$ distributions and the data-driven one obtained from the data subtracting the QCD, after the likelihood fit. In each score bin the data-MC arrangement is compatible within one standard deviations.

REFERENCES

[1] Evans Lyndon and Bryant Philip, *JINST*, **3** (2008) S08001.
[2] The CMS Collaboration, *JINST*, **3** (2008) S08004.
[3] The CMS Collaboration, *JINST*, **12** (2017) P10003.
[4] Cacciari Matteo *et al.*, *JHEP*, **63** (2008) 0804.
[5] Qu Huilin and Gouskos Loukas, *Phys. Rev. D*, **101** (2020) 056019.
[6] Wang Y. *et al.*, *ACM Trans. Graph.*, **38** (2019) 146.
[7] The CMS Collaboration, *CMS Physics analysis summary*, CMS-PAS-BTV-22-001 (CERN, Geneva) 2023, https://cds.cern.ch/record/2804004.
[8] The CMS Collaboration, *Eur. Phys. J. C*, **46** (2006) 605.
[9] The CMS Collaboration, *JINST*, **16** (2021) P05014.
[10] The CMS Collaboration, *JINST*, **13** (2018) P06015.
[11] Larkosk A. J. *et al.*, *JHEP*, **05** (2014) 146.
[12] The CMS Collaboration, CMS-CAT-23-001 (CERN, Geneva) 2024, https://cds.cern.ch/record/2895097.