PROCEEDINGS OF SCIENCE

B-hadron identification in b-jets using novel deep learning technique in pp and PbPb collisions in CMS

Prince Kumar* **on behalf of the CMS Collaboration**

Indian Institute of Technology Madras,
Chennai, India

E-mail: kumar.prince@cern.ch

Advancements in geometric deep learning offer powerful tools to study the internal structure of jets initiated by heavy quarks, particularly in the context of dead-cone effect and jet quenching. The kinematics of b-hadron decays present a challenge for substructure measurements with inclusive b-jets, which are essential for quantum chromodynamics (QCD) studies. We propose an approach using graph-based deep learning that utilises charged decay products of the jets represented as point clouds to simultaneously identify tracks associated with b-hadron decay and perform b-jet tagging. The method is demonstrated in simulated p-p ($\sqrt{s}$ = 5.02 TeV) and Pb-Pb ($\sqrt{s_{NN}}$ = 5.02 TeV) collisions passed through the CMS detector framework, in both Run 2 and Run 3 conditions. We benchmark our method against traditional boosted decision tree classifiers, showcasing significant performance improvements in b-hadron identification of tracks.

The European Physical Society Conference on High Energy Physics (EPS-HEP2025)
7-11 July 2025
Marseille, France

*Speaker

 ISSN 1824-8039 . Published by SISSA Medialab.

PoS(EPS-HEP2025)664

1. Introduction

Jet substructure is an important tool to probe the properties of quantum chromodynamics (QCD) and heavy flavour dynamics. For b hadrons, the finite quark mass regulates the soft and collinear divergences, making the study of gluon suppression at small angles, known as the dead cone effect, particularly interesting. However, the presence of heavy-flavour decay daughters obscures radiation patterns in heavy-flavor jets. Identifying b hadron decay daughters is essential for precise b flavour studies and observing classification-sensitive track-level observables. We present a graph neural network based approach [1] for track identification in b jets in p-p and Pb-Pb collisions in CMS.

2. Methodology

2.1 Dataset Details

The study is performed using simulated pp ($\sqrt{s} = 5.02\ TeV$) and PbPb ($\sqrt{s_{NN}} = 5.02\ TeV$) collisions passed through the CMS detector framework generated with PYTHIA8 Monte Carlo (MC) generator [2].

Jets are reconstructed with the anti-kt [3, 4] algorithm with a distance parameter R=0.4. Charged particles from pileup vertices are identified and eliminated from the jet reconstructions [5]. The jets are selected to be within $|\eta| < 2$ and have $p_T^{jet} > 30\ GeV$ and $p_t^{track} > 1\ GeV$. The subset of b jets is selected using the ParticleNet tagger [6, 7] such that gluon and light-quark contamination is negligible. Jets are modelled as graphs with individual tracks as nodes, enabling network to capture complex dependencies inherent in particle interactions.

For the proton–proton (pp) dataset, the training sample comprises approximately 1.3×10^7 jets, while the test sample contains 4.0×10^4 jets. For the lead–lead (PbPb) dataset, the corresponding sizes are 5.0×10^6 jets for training and 1.4×10^5 jets for testing. The classification model, **FusionNet**, is trained on 16 input features spanning multiple levels of jet characterization, from global jet-level observables (broadcast to all constituent tracks) to secondary-vertex-level information.

2.2 Training Details

We train our model using pytorch and torch-geometric [8] with binary cross entropy (BCE) loss. Tracks matched to generator-level charged particles from b hadron decays are considered signal, and background comprises all tracks not from b hadron decays. While training we maintain even split of signal and background to avoid biases and robust training.

2.3 Network Architecture

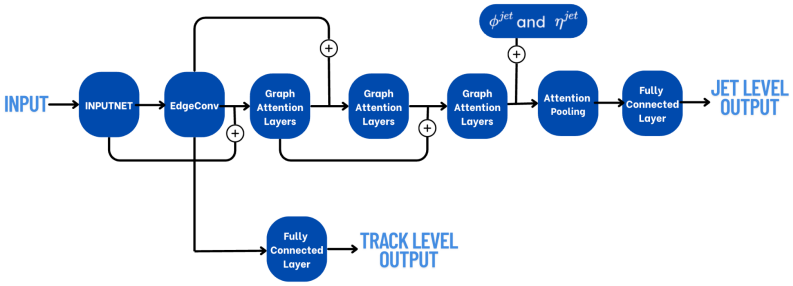

Figure 1: FusionNet - Model Architecture. Jet level tagging is an auxiliary task, targeting jet-level features helps in track-level feature performance. FusionNet-S: trained on p–p (S_{pp}) or Pb–Pb (S_{PbPb}) Run 2 simulations; FusionNet-T: pretrained on p–p, fine-tuned on Pb–Pb Run 2.

3. Results

3.1 Performance in proton-proton (pp) Monte Carlo (MC) simulated events

The performance of FusionNet was first evaluated on MC simulated pp events. Figure 2 shows that the inclusion of jet-level classification as a simultaneous task improves b-track identification performance compared to a model trained without this information. This suggests the network leverages global jet-level variables to enhance local track-level decisions. Additionally, the model was tested for robustness in important kinematic variable regions. Figure 3 displays the ROC curves binned by jet p_T, η, and track multiplicity. The performance remains consistent across different bins, with minor variations within error bands.

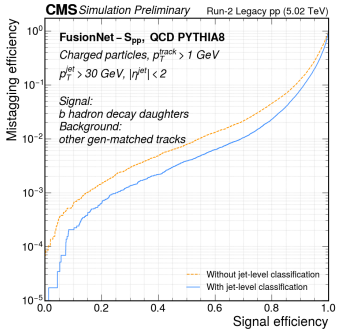

Figure 2: Comparative ROC curves of $FusionNet - S_{pp}$ trained with and without jet-level classification for b-track identification on pp MC simulated events.

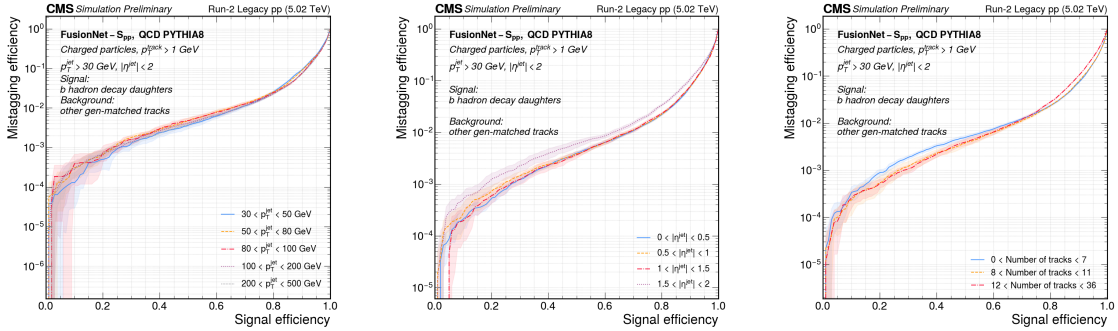

Figure 3: ROC curves for $FusionNet - S_{pp}$ in different bins of jet p_T (left), η (middle), and number of tracks (right). Error bands represent mean and standard deviation from bootstrapped samples.

3.2 Performance on ParticleNet-tagged b-jets

The model was evaluated in high-purity environments of pp collisions on jets already tagged by ParticleNet. As illustrated in Figure 4, $FusionNet - S_{pp}$ showcases improvements in identifying tracks originating from b-hadron decays compared to the Gradient BDT model [9], particularly at signal efficiencies above 0.75. The performance remains consistent across various kinematic bins, as shown in Figure 5.

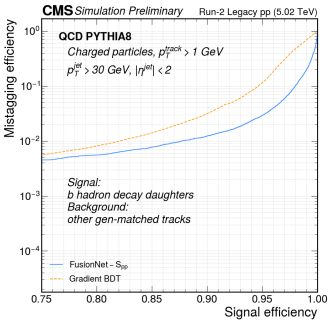

Figure 4: Comparative ROC curves for $FusionNet - S_{pp}$ and Gradient BDT models on ParticleNet-tagged b-jets in pp MC events.

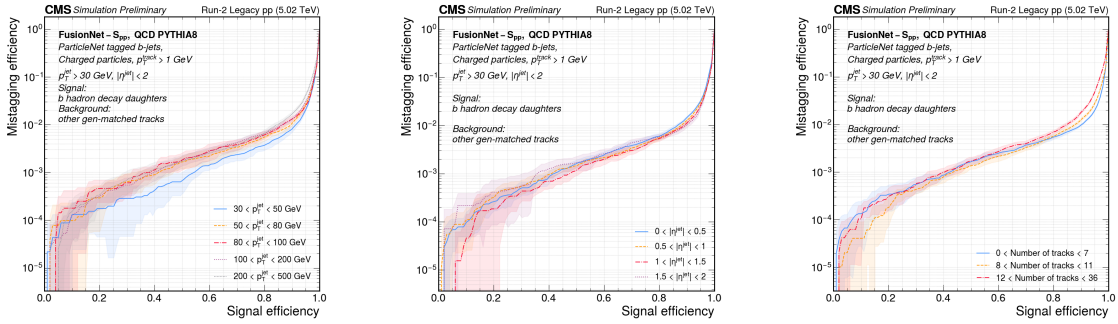

Figure 5: Performance evaluation on ParticleNet b-tagged jets in pp MC events, binned by jet p_T (left), η (middle), and track multiplicity (right).

3.3 Performance in lead-lead ($PbPb$) Monte Carlo simulated events

In the higher-multiplicity heavy-ion collisions ($PbPb$), we utilized FusionNet-T, which was pre-trained on pp samples before being fine-tuned on $PbPb$ samples. Figure 6 shows that FusionNet-T outperforms the $FusionNet - S_{PbPb}$ variant, which was trained and tested explicitly on $PbPb$ MC simulated events. Finally, Figure 7 demonstrates that FusionNet-T performance remains robust against complex variations in jet kinematic variables in $PbPb$.

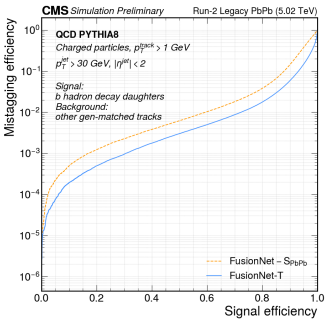

Figure 6: Comparative ROC curves for FusionNet-T and $FusionNet - S_{PbPb}$ in $PbPb$ MC samples.

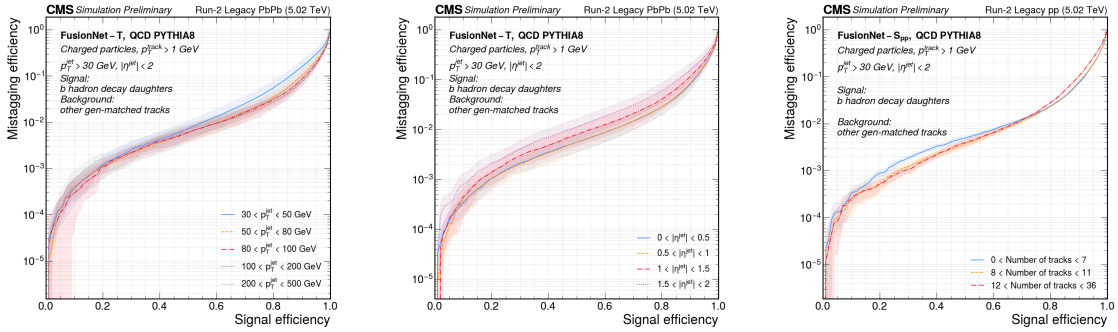

Figure 7: ROC curve of FusionNet-T evaluated on $PbPb$ MC simulated events in bins of jet p_T (left), η (middle), and number of tracks (right).

4. Conclusion

FusionNet demonstrates excellent performance in b track identification in simulated pp and PbPb collisions using the CMS Detector Framework. It surpasses traditional BDT approaches, particularly in regions of high signal efficiency. This study presents its architecture and performance with respect to key kinematic variables, showcasing its robustness.

References

[1] CMS Collaboration. (2025). A novel track finding algorithm to identify b hadron in b jets using FusionNet: a geometric deep learning model. CMS-DP-2025-035

[2] Sjöstrand, T., Ask, S., Christiansen, J. R., Corke, R., Desai, N., Ilten, P., Mrenna, S., Prestel, S., Rasmussen, C. O., & Skands, P. Z. (2015). An introduction to PYTHIA 8.2. Comput. Phys. Commun., 191, 159–177.

[3] Cacciari, M., Salam, G. P., & Soyez, G. (2008). The anti-kt jet clustering algorithm. JHEP, 04, 063.

[4] Cacciari, M., Salam, G. P., & Soyez, G. (2012). FastJet user manual (version 3.0.2). Eur. Phys. J. C, 72(3).

[5] CMS Collaboration. (2020). Pileup mitigation at CMS in 13 TeV data. JINST, 15, P09018.

[6] Qu, H., & Gouskos, L. (2020). Jet tagging via particle clouds. Phys. Rev. D, 101, 056019.

[7] CMS Collaboration. (2024). Run 3 commissioning results of heavy-flavor jet tagging at $\sqrt{s}$ = 13.6 TeV with CMS data using a modern framework for data processing. CMS-DP-2024-024.

[8] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph Attention Networks. In International Conference on Learning Representations (ICLR).

[9] CMS Collaboration. (2024). Jet fragmentation function and groomed substructure of bottom quark jets in proton-proton collisions at $\sqrt{s}$ = 5.02 TeV. CMS-PAS-HIN-24-005.