PROCEEDINGS OF SCIENCE

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

Ming-Yan Lee[a,*] **for the CMS collaboration**

[a] *Physics Institute III A,*
Sommerfeldstraße 16, Aachen, Germany

E-mail: Ming-Yan.Lee@physik.rwth-aachen.de

The identification of jets arising from heavy-flavor (bottom or charm) quarks primarily relies on detector inputs from reconstructed charged particle tracks and information about secondary vertices contained within reconstructed jets. In Run 3, improved machine-learning techniques have been introduced to distinguish heavy-flavor jets from those originating from the hadronization of light-flavor (uds) quarks or gluons (g). Therefore, it is crucial to compare the distributions of data and simulations of input variables, tagging discriminants, and other pertinent kinematic observables between data and simulated events. In this proceeding, five selections to enriched different processes are presented: top quark-antiquark production ($t\bar{t}$) in the dileptonic final state (enriched in b jets), in the semileptonic final state (enriched in b and c jets), W boson plus charm production (enriched in c jets), Drell-Yan production, and QCD multijet production (enriched in light-flavor jets). These selections are shown with proton-proton collision data at $\sqrt{s}$ = 13.6 TeV corresponding to an integrated luminosity of 61.7 fb^{-1} and recorded by the CMS experiment in 2022 and 2023. These studies rely on a modern and fast framework. It has been developed and automated to produce the comparisons presented, along with its technical details.

12th Large Hadron Collider Physics Conference (LHCP2024)
3-7 June 2024
Boston, USA

*Speaker

PoS(LHCP2024)272

1. Introduction

The identification of jets arising from heavy-flavor (bottom or charm) quarks is crucial for several analyses in CMS [1–4], such as investigations of the properties of the Higgs boson, top quarks, and the search for new physics beyond the standard model. To distinguish heavy-flavor jets from light-flavor quarks or gluons jets, flavor tagging algorithms primarily rely on detector inputs from reconstructed charged particle tracks and information about secondary vertices contained within reconstructed jets. The information associated with jets is used as input to advanced machine-learning methods. Therefore, understanding the agreement of data and simulation is crucial to evaluate the performance and spot possible mismodeling in simulation. This proceeding (based on Ref. [5]) presents the data and simulation comparison using the proton-proton collision data collected by the CMS detector at $\sqrt{s}$ = 13.6 TeV during 2022-2023 data-taking periods, which correspond to 61.7fb^{-1}. The state of the art taggers used for heavy flavor tagging in CMS Run 3 – DeepJet [6, 7], ParticleNetAK4 [5, 8], and RobustParTAK4 [5, 9–11] – are shown with five selections of data, each enriched with b jets, c jets and light- (udsg) flavor jets in different kinematic regions. A modern framework with efficient and automatized workflows delivers this study.

2. Heavy flavor jet identification in CMS

The heavy flavor jet identification algorithms rely on the variables connected to the properties of heavy flavor hadrons present in jets, such as the presence of secondary vertices, higher track multiplicities, and more tracks with positive signed impact parameters. The collective behavior of the inputs allows us to achieve good discrimination of heavy flavor jets using advanced machine-learning methods developed at the beginning of Run 3. In the early Run 3 studies, two new taggers, **ParticleNetAK4** and **RobustParTAK4**, were introduced to improve the performance and robustness. **ParticleNetAK4** is customized for heavy-jet jet identification for AK4 jet classification where AK4 refers to jets reconstructed using the anti-k_t [12] clustering algorithm with a distance parameter of $\Delta R = 0.4$ using particle-flow [13] candidates. This algorithm incorporates heavy flavor and hadronic tau identification, along with flavor-aware jet energy correction and resolution, based on the ParticleNet [8] architecture. **RobustParTAK4** uses the ParticleTransformer [9, 10] architecture for AK4 jets with the pairwise interaction features between all input jet constituents and secondary vertices and implements an adversarial training [11] to enhance the robustness of the model against the mismodeling presented in our simulated events. To evaluate the performance on b jets and c jets, three discriminants, `BvsAll, CvL, CvB`, are defined using the probability of the neural network (`P(b)` and `P(c)`) of identifying b or c jets from other jets. `BvsAll` is defined as `BvAll = P(b)` which is used to evaluate the performance of b jet efficiency and light-flavor jet mistag rate. The CvL and CvB are the c jet discriminants which are defined as `CvL = P(c) / [P(c) + P(udsg)]` and `CvB = P(c) / [P(c) + P(b)]`.

3. Heavy flavor jet performance of Run 3 data

Jets used in this study are AK4 jets and the pileup mitigation is performed using the pileup-per-particle identification (PUPPI) algorithm [14, 15], which assigns a weight to every particle

depending on its probability of originating from either a pileup or the leading vertex. Dedicated jet energy corrections (JEC) derived from Run 3 data [16, 17] are applied to the jets. The jets are required to have $p_T > 20\,\text{GeV}$ with $|\eta| < 2.5$, fulfilling tight identification criteria, and separated from the selected prompt lepton by at least $\Delta R > 0.4$. The selections enriched with b, c, and light-flavor jets are summarized below:

- Dileptonic $t\bar{t}$ selection (**b jet enriched**): Dileptonically decaying $t\bar{t}$ events form a final state with a large b jet purity. This event topology is relevant for deriving calibration for b tagging [19]. Events are selected with a set of electron-muon (eμ) trigger paths. The electron and muon are required to fulfill $p_T > 30\,\text{GeV}$, $|\eta| < 2.5(2.4)$ and to pass tight identification and isolation requirements [4, 20]. Events with at least two jets, one electron, and one muon are selected.

- Semileptonic $t\bar{t}$ selection (**b jet and c jet enriched**): The hadronically decaying W boson from the semileptonic $t\bar{t}$ events provide a significant amount of c jets. This region can be used for calculating b tagging and c tagging efficiency scale factors [19, 21]. Events are selected using a single-muon trigger path. The selected muon is required to fulfill the same selection criteria as the one of the dileptonic $t\bar{t}$ phase space. Events are required to have $p_T^{miss} > 50\,\text{GeV}$ and at least four jets.

- W boson plus charm jet (W+c) selection (**c jet enriched**): This selection is largely enriched in c jets and is utilized for evaluating the c-tagging performance [21]. The signal process contains a leptonically decaying W boson produced along with a c jet. These c jet candidates are identified using the semileptonic decay of the c hadron, which can produce a soft lepton within the jet in the final state. The same trigger path and the same selection criteria for the isolated muon as for the semileptonic $t\bar{t}$ selection are required. At least one additional soft muon with a reduced p_T threshold and a relative isolation greater than 0.2 is selected and matched with one of the selected jets. The jet multiplicity is required to be at least one jet and up to three jets. Both opposite-sign (OS) and same-sign (SS) isolated muon and soft muon pairs are taken into account. Additional selection criteria to enrich the sample with W boson events and to suppress QCD multijet and Drell–Yan contributions are applied as well. To enrich the selected events with W bosons, the sum of p_T^{miss} and isolated muon four-vector is required to be $> 55\,\text{GeV}$.

- Drell–Yan (DY) plus jets selection (**light-flavor jet enriched**): This selection is enriched in light-flavor jets used for the calibration of light-mistagging efficiencies [21]. A di-muon trigger path is employed to select $Z\rightarrow\ \mu\mu$ events. The leading (subleading) muon has to fulfill a criteria of $p_T > 15(12)\,\text{GeV}$, where both muons are required to satisfy $|\eta| < 2.4$, as well as tight identification and isolation requirements [20]. The invariant di-muon mass is required to be $|m_Z - m_{\mu\mu}| < 15\,\text{GeV}$. At least one jet is required in each event.

- Inclusive QCD multijet selection (**light-flavor jet enriched**): This selection has a large fraction of light-flavor jets with higher statistics at high p_T region. Events are selected if they satisfy a trigger selection of at least one AK4 jet with $p_T > 180\,\text{GeV}$, $|\eta| < 2.4$. Due to the high event rates, only part of the events passing the trigger requirement are selected

(prescaled trigger). The fraction of accepted events depends on the prescale value, which varies during the data-taking period according to the instantaneous luminosity. The data are compared to multijet events simulated at leading order in QCD.

Although an overall agreement between data and simulation is observed (as shown in Figure 1), certain trends are observed in the distribution of the tagger scores. This indicates that a calibration of the tagger score is required. During the 2022 and 2023 data-taking period, two main detector issues occurred, which required dedicated simulations to mimic the effect. The water leakage issue of electromagnetic calorimeter (ECAL) [18] happened in late 2022 in the endcap region (2022 post-EE) and a region of barrel pixel (BPix) was deactivated in the late 2023 data-taking period (2023 post-BPix). The performance of heavy flavor jets using dileptonic $t\bar{t}$ phase space is stable between different periods among all the taggers as shown in Figure 2.

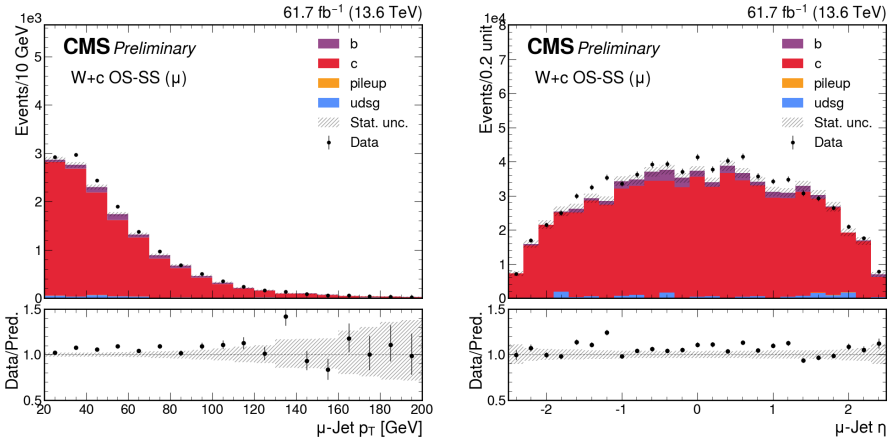

Figure 1: Jet p_T (left) and η (right) distributions in the W+c phase space.

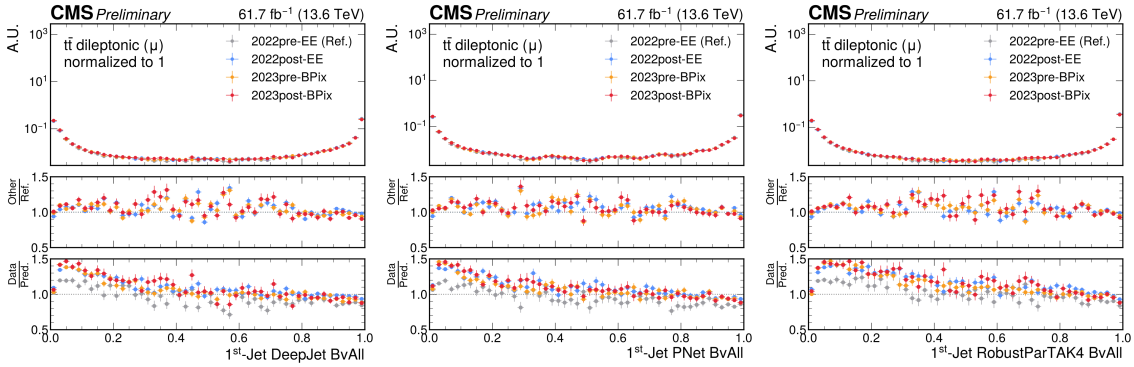

Figure 2: Tagger scores across different data-taking periods in the dileptonic $t\bar{t}$ phase space. The first panel shows the distribution normalized to 1. The second panel uses the 2022 pre-EE data-taking period as a reference and shows the ratio of the scores observed in other data-taking conditions versus to the one observed in the reference. The third panel shows the data and prediction ratios in each individual era.

4. Common BTV framework [24]

To deal with the growing data at the LHC, an efficient and reproducible framework is important for heavy flavor tagging studies. This framework is based on `coffea` [23] which uses customized NanoAOD [22] as input. The `coffea` framework reads the information as `awkward` [18] arrays (columnar data) and also includes tools to apply corrections and systematic variations. The Common BTV framework (`BTVNanoCommissioning`) [24] unifies the selections from different phase spaces and implements the corrections and systematic variations. The processed information can be further converted to arrays, histograms, and figures, and automatically transferred to the storage area (EOS) via GitLab CI pipeline.

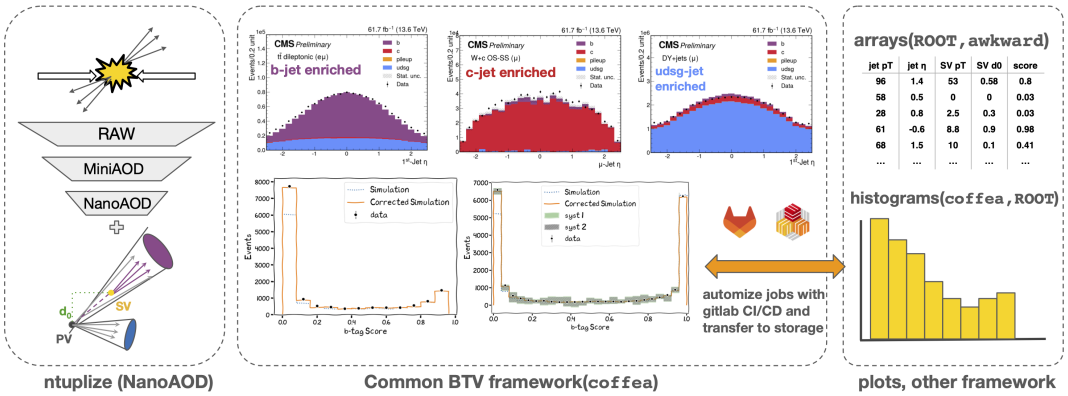

Figure 3: Workflow of the common BTV framework. It begins by adding customized flavor-tagging information to the NanoAOD files and creating flat ntuples. Next, it proceeds to the common BTV framework where events are selected for b, c, or light-flavor jet enriched regions, with the corrections and systematic variations applied on the fly. Finally, the information is stored either as histograms (`coffea` [23], `ROOT` [25]) or arrays (`awkward` [18], `ROOT` [25]), to be used for plotting or as input for other frameworks (e.g., scale factor derivation).

5. Summary

The study demonstrates the outputs of the new heavy flavor jet taggers evaluated on CMS Run 3 data and simulation, and the utility of the modern data processing framework in handling complex data. With the stability of different taggers and good agreement among different phase spaces, the results provide a foundation for ongoing and future physics analyses using the improved heavy flavor tagging capabilities of the CMS experiment.

References

[1] CMS Collaboration, *The CMS experiment at the CERN LHC* , *JINST* **3** (2008) S08004.

[2] CMS Collaboration, *Development of the CMS detector for the CERN LHC Run 3*, *JINST* **19** (2024) P05064.

[3] CMS Collaboration, *Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV*, *JINST* **8** (2018) P04013.

[4] CMS Collaboration, *Identification of b-quark jets with the CMS experiment*, *JINST* **13** (2018) P05011.

[5] CMS Collaboration, *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 Detector Performance Summary CMS-DP-2024-024.

[6] E. Bols, J. Kieseler, M. Verzetti, M. Stoye and A. Stakia, *Jet flavour classification using DeepJet*, *JINST* **15** (2020) P12012.

[7] CMS Collaboration, *Performance summary of AK4 jet b tagging with data from proton-proton collisions at 13 TeV with the CMS detector*,CMS Detector Performance Summary CMS-DP-2023-005.

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

[9] H. Qu, C. Li, S. Qian, *Particle Transformer for Jet Tagging*, arXiv:2202.03772.

[10] CMS Collaboration, *Transformer models for heavy-flavor jet identification*, CMS Detector Performance Summary CMS-DP-2022-050.

[11] A. Stein, X. Coubez, S. Mondal, A. Novak, A. Schmidt, *Improving Robustness of Jet Tagging Algorithms with Adversarial Training*, *Comput Softw Big Sci* **6** (2022) 15.

[12] M. Cacciari, G. P. Salam and G. Soyez, *The anti-kt jet clustering algorithm*, *JHEP* **0804** (2008) 063.

[13] CMS Collaboration, *Particle-flow reconstruction and global event description with the CMS detector*, *JINST* **12** (2017) 10, P10003.

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

[15] CMS Collaboration, *Pileup-per-particle identification: optimisation for Run 2 Legacy and beyond*, CMS Detector Performance Summary CMS-DP-2021-001.

[16] CMS Collaboration, *Jet Energy Scale and Resolution Measurements Using Prompt Run3 Data Collected by CMS in the First Months of 2022 at 13.6 TeV*, CMS Detector Performance Summary CMS-DP-2022-052.

PoS(LHCP2024)272

[17] CMS Collaboration, *Jet energy scale corrections and uncertainties derived from CMS Run 2 data*, arXiv:1706.02830.

[18] CMS Collaboration, *The water leak problem in the electromagnetic calorimeter endcaps*, CMS internal note, 2022.

[19] CMS Collaboration, *Performance of heavy-flavor identification with the CMS detector using 2017-2018 data*, CMS Detector Performance Summary CMS-DP-2020-010.

[20] CMS Collaboration, *Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at* $\sqrt{s} = 13$ *TeV*, *JINST* **17** (2022) P07025.

[21] CMS Collaboration, *Measurements of c-tagging efficiency using* $W + c$ *events at CMS*, CMS Physics Analysis Summary CMS-PAS-BTV-20-001.

[22] CMS Collaboration, *The CMS NanoAOD format for Run 2 legacy analyses*, CMS Detector Performance Summary CMS-DP-2019-016.

[23] L. Gray, J. Pata, N. Smith, A. Ustyuzhanin, *Coffea - Columnar Object Framework for Effective Analysis*, arXiv:2008.12712.

[24] cms-btv-pog, *BTVNanocommissioning*, https://github.com/cms-btv-pog/BTVNanoCommissioning.

[25] R. Brun and F. Rademakers, *ROOT — An object-oriented data analysis framework*, *Nucl. Instrum. Meth. A* **389** (1997) 81-86.

PoS(LHCP2024)272