Summer stundent report

Particle tagging for $t\bar{t}$ production in heavy-ion collisions

Author: Danilo Drincic
Supervisors: Andre Govinda Ståhl, Pedro Silva and Sitian Qian
Date: July 27th, 2024.

1 Abstract

During the summer of 2024, I had the privilege of working as a summer student at CERN within the Experimental Physics Department, to be specific on CMS experiment. My project concentrated on advanced techniques for particle tagging, specifically b-tagging and W-tagging in hadronic events. B tagging is essential for identifying bottom quarks, and crucial for understanding Higgs boson decays and top quark production, while W-tagging is vital for singling out in a collision, the hadronic remnants of a W boson which has decayed into a pair of quarks. These hadronic remnants typically give rise to two jets which tend to merge to a single one if the initial momentum is sufficiently high. Through my work, I contributed to refining these tagging methods to improve the accuracy and efficiency of particle detection. This experience not only enhanced my technical skills but also deepened my understanding of particle physics.

2 Introduction

2.1 General Standard Model

The Standard Model of particle physics describes the fundamental particles (quarks, leptons, and bosons) and their interactions, crucial for understanding the universe at its smallest scales. Matter is composed of three kinds of elementary particles: leptons, quarks, and bosons. There are six leptons, classified by their charge (Q), electron number (L_e), muon number (L_μ), and tau number (L_τ), which fall into three generations as shown in Table 1.

Table 1: Lepton Classification

Generation	Particle	Q	L$_e$	L$_\mu$	L$_\tau$
First	ν_e	0	1	0	0
	e^-	-1	1	0	0
Second	ν_μ	0	0	1	0
	μ^-	-1	0	1	0
Third	ν_τ	0	0	0	1
	τ^-	-1	0	0	1

Similarly, there are six 'flavors' of quarks, classified by charge (Q), strangeness (S), charm (C), beauty (B), and truth (T), also categorized into three generations as illustrated in Table 2.

Table 2: Quark Classification

Generation	Particle	Q	S	C	B	T	Mass (MeV/c^2)
First	u	2/3	0	0	0	0	2.3
	d	-1/3	0	0	0	0	4.8
Second	c	2/3	0	1	0	0	1275
	s	-1/3	-1	0	0	0	95
Third	t	2/3	0	0	0	1	173100
	b	-1/3	0	0	-1	0	4180

Quarks come in three colors, leading to 36 different quark states when considering both quarks and antiquarks. Each interaction has its own mediator: the photon for the electromagnetic force, W and Z bosons for the weak force, gluons for the strong force, and presumably the graviton for gravity [1]. Figure 1 graphically depicts the Standard Model. The top quark (t), the heaviest known elementary particle, plays a crucial role in electroweak symmetry breaking and mass generation, with a mass around 173 GeV. It primarily decays into a W boson and a bottom quark ($t \to Wb$), with a short lifetime that prevents hadronization, allowing it to be studied through its decay products in high-energy colliders like the LHC. Precision measurements of top quark properties provide critical tests of the Standard Model and insights into potential new physics. In high-energy proton-proton (pp) collisions, top quarks are mainly produced via gluon-gluon fusion ($gg \to t\bar{t}$), as depicted in Figure 2, with a production cross-section in pp collisions at 13 TeV on the order of hundreds of picobarns [1]. In lead-lead ($PbPb$) collisions, the $t\bar{t}$ production cross-section decreases due to lower collision energy, but this is compensated by the combinatorics of nucleon pair collisions (enhancing the cross-section by a factor of $208^2 = 43264$). Top quarks provide a clean probe of the quark-gluon plasma (QGP) in heavy-ion collisions, as they decay before hadronization, offering direct insights into partonic interactions and energy loss mechanisms in the QGP, thus testing Quantum Chromodynamics (QCD) models under extreme conditions [16].

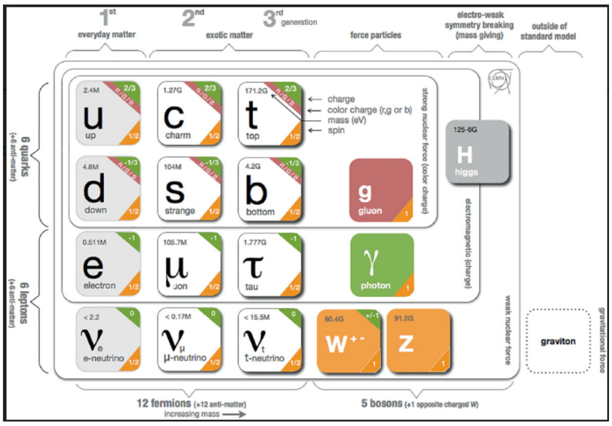


Figure 1: Graphical depiction of the Standard model [21].

2.2 The state-of-the-art

The state-of-the-art in measurements of proton-proton (pp) collisions represents the pinnacle of modern high- energy physics and experimental technology. At its core, this involves the use of highly advanced detectors such as ATLAS [14] and CMS [15], which are capable of tracking and measuring an astonishing variety of particles with incredible precision. These detectors incorporate sophisticated subsystems including silicon-based tracking detectors, calorimeters designed to measure energy by ab-

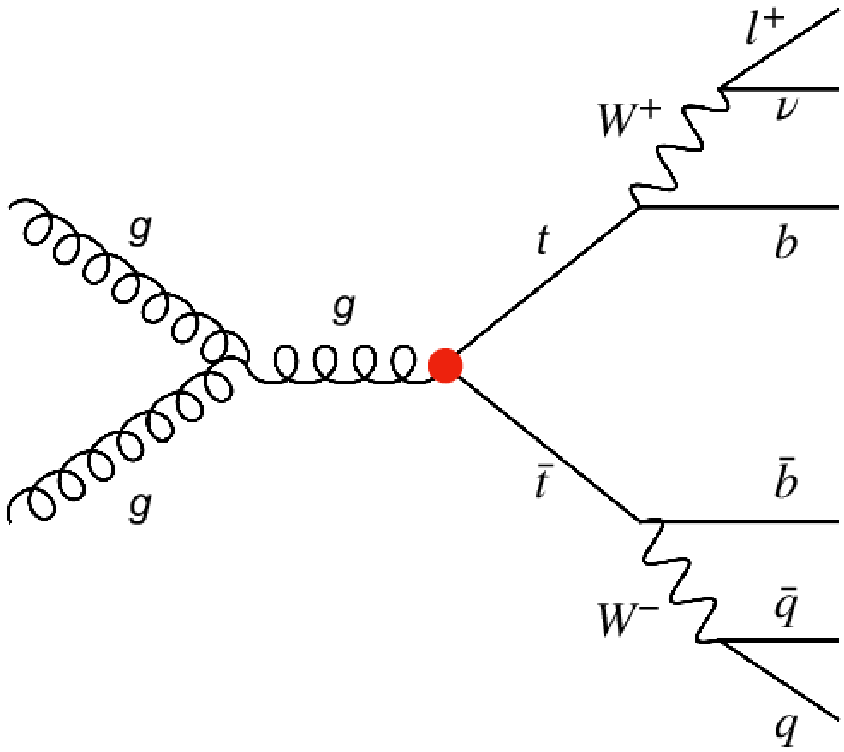


Figure 2: Feynman diagram for $t\bar{t}$ production [22].

sorbing particles, and specialized muon detectors, all working in concert to provide a comprehensive picture of each collision event. The real-time processing of data is managed by ultra-fast trigger systems that can discern and record the most promising collision events from millions occurring every second. Furthermore, the analysis of this data is enhanced by the application of advanced machine learning algorithms, which sift through massive datasets to identify meaningful patterns and signals amidst the noise. State-of-the-art also encompasses the use of detailed Monte Carlo simulations to model theoretical predictions and compare them with experimental outcomes, refining our understanding of particle interactions.

2.2.1 The state-of-the-art in lead-lead (PbPb) collisions

The CMS experiment at the LHC reported the first evidence of top quark production in heavy-ion collisions, specifically in lead-lead (PbPb) collisions at a center-of-mass energy of 5.02 TeV per nucleon pair [6]. This achievement was made using data from the 2018 heavy-ion run. The analysis employed two methods to identify top quark pair ($t\bar{t}$) events: a dilepton-only approach and a dilepton plus b-tagged jets approach. Both methods used a boosted decision tree (BDT) classifier to distinguish the signal from the background. The results showed a significance of 4.0 standard deviations for the observation of $t\bar{t}$ production, providing strong evidence for the presence of top quarks in this collision environment.

The state-of-the-art in lead-lead (PbPb) focuses on studying the quark-gluon plasma (QGP), a state of matter from the early universe, using advanced detectors such as ALICE [13]. These detectors handle the immense particle multiplicity and energy densities involved in PbPb collisions, providing detailed measurements of the QGP. However, these collisions face significant challenges, including complex data analysis due to overlapping particle tracks, energy loss mechanisms that obscure particle signals, and substantial background requiring precise calibrations and analysis techniques.

2.3 The anti-k_T algorithm

The anti-k_T algorithm is a widely used jet reconstruction method , designed to identify and cluster particles into jets from the complex environments of particle collisions [12] . This algorithm is a variant of the k_T clustering method and stands out for its effectiveness in handling dense environments and overlapping jets. The anti-k_T algorithm clusters particles by minimizing a distance measure that incorporates both the transverse momentum and spatial separation of particles. Specifically, the distance d_{ij} between two particles or clusters i and j is calculated using the formula

$$d_{ij} = \min(p_{T,i}^2, p_{T,j}^2) \cdot \frac{\Delta R_{ij}^2}{R^2},$$

where $p_{T,i}$ and $p_{T,j}$ are the transverse momenta of the particles or clusters, ΔR_{ij} is the distance in the $\eta - \phi$ plane (pseudorapidity and azimuthal angle), and R is a parameter that sets the jet size. The "anti" in anti-k_T signifies that this algorithm favors clustering that maintains larger distances between particles, leading to jets with clearer and more distinct boundaries. This property makes the anti-k_T algorithm particularly suited for high- density environments where multiple jets are closely packed.

For example, in proton-proton (pp) collisions, where high-energy jets are produced, the anti-k_T algorithm helps in reconstructing jets with clear separation, which is crucial for precise measurements and analyses. In lead-lead (PbPb) collisions, where the environment is even denser due to the large number of colliding nucleons, the anti-k_T algorithm excels in identifying individual jets despite the overlapping and merging of multiple jet structures. For instance, when studying jet quenching, a phenomenon observed in PbPb collisions where high-energy jets lose energy while traversing the quark- gluon plasma, the anti-k_T algorithm can effectively reconstruct the jets and measure their energy loss [7].

2.4 Jets classification

Jet classification is a crucial task that involves distinguishing between different types of jets based on their origins or specific characteristics. Jets are essentially collimated streams of particles resulting from the hadronization of high-energy quarks and gluons in particle collisions. Classifying these jets accurately requires analyzing a variety of features, such as jet energy, transverse momentum (p_T), mass, and the distribution of energy within the jet. Multivariable analysis plays a pivotal role in this process by leveraging multiple input features to build sophisticated models capable of differentiating between jets originating from different physical processes. By incorporating various kinematic and substructure variables into a machine learning model, scientists can train algorithms to recognize and classify these jets with high precision.

The training process involves using labeled datasets where jets are classified based on their known origins or types. Machine learning algorithms, such as decision trees, random forests, support vector machines, or neural networks, are then used to analyze these datasets and learn the patterns associated with different jet types. Once the model is trained, it is validated and tested on new data to ensure its accuracy and reliability. The validated model can then be applied to experimental data, enabling real-time classification of jets and facilitating the study of fundamental physical phenomena. For example multivariable analysis helps in distinguishing between jets from top quark decays and those from other processes [8].

2.5 Particle Transformer

Multivariable analysis (MVA) is a versatile tool applied across diverse fields beyond high-energy physics, offering insights by analyzing multiple variables simultaneously. Among the different MVAs, the Particle Transformer model [10], developed for use by the (CMS) experiment, represents a significant leap forward in particle physics data analysis through the application of advanced deep learning techniques. Inspired by transformer networks originally designed for natural language processing [11], the Particle Transformer model leverages the powerful attention mechanisms of these networks to analyze and interpret complex particle collision data. It processes the data by representing each particle or jet with a detailed feature vector that includes kinematic information such as transverse momentum and energy. The core innovation of the Particle Transformer lies in its self-attention layers, which allow the model to effectively weigh and prioritize different particles and their interactions, capturing intricate patterns and relationships that are critical for accurate particle and jet reconstruction. This approach is particularly valuable in high-energy collisions, where overlapping and merging jets make traditional reconstruction methods challenging.

In proton-proton (pp) collisions, the Particle Transformer model enhances the precision of jet reconstruction and particle identification by providing a sophisticated analysis of the particle data. This improvement is crucial for distinguishing between jets from different physical processes, such as those originating from top quark decays. By reducing uncertainties and improving measurement precision, the Particle Transformer model supports more reliable and detailed studies of high-energy collisions, thereby advancing the understanding of fundamental particles and interactions.

2.5.1 Implementation in ($PbPb$) collision measurements

The Particle Transformer model, while highly effective for jet reconstruction and particle classification in proton-proton (pp) collisions, encounters challenges in lead-lead (PbPb) collisions due to the extreme conditions of the QGP. In PbPb collisions, the QGP alters jet properties through phenomena such as jet quenching, which causes high-energy jets to lose energy and become softer and broader. This makes jet structures less distinct compared to those in pp collisions, challenging the Particle Transformer's ability to accurately capture and classify intricate patterns. Additionally, the increased particle density in PbPb collisions leads to severe jet overlaps and merging, complicating the reconstruction process and overwhelming the model's capacity to separate and identify individual jets amidst the high background noise. The combination of these factors—distorted jet profiles, high multiplicity, and complex interactions—reduces the effectiveness of the Particle Transformer model in these dense environments, highlighting the need for specialized techniques to handle the unique challenges of high-density collision scenarios.

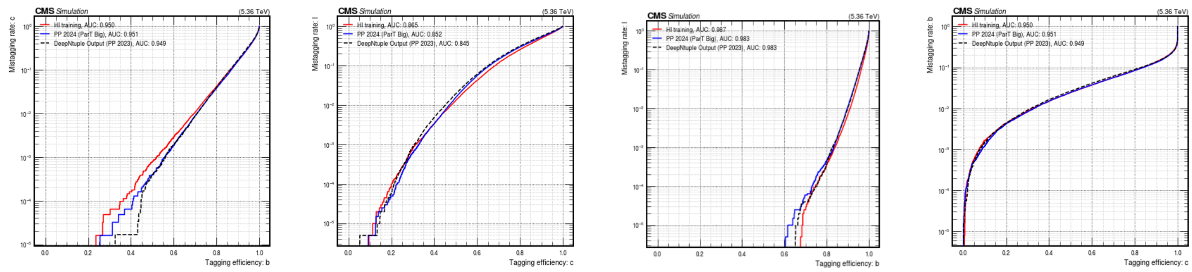
3 Results

This section outlines the results derived from simulations of $t\bar{t}$ production in PbPb collisions at 5.36 TeV. The $t\bar{t}$ production was simulated using the MadGraph [17] and Powheg [18]event generators, while the PbPb underlying events were generated using the Hydjet [19] event generator. Additionally, a separate dataset, representing signals simulated in pp collisions without pileup (PU), was utilized to provide a control sample for comparison. The CMS detector response was emulated using the Geant4 [20] framework, configured with the 2023 data-taking conditions.

The performance of the b-quark tagging, as compared to charm and light quarks, is detailed in Section 3.1 for the Particle Transformer model and in Section 3.2 for the DeepJet model. Both models were trained over 30 epochs to optimize their tagging efficiency and accuracy. Following this, Section 3.3 presents the results for the W-tagger, derived using a boosted decision tree (BDT) model.

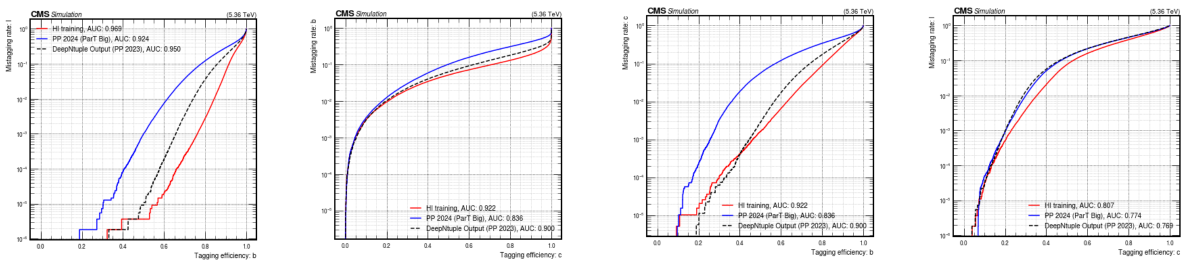
3.1 b tagging

If we look at 4 graphs in figure 3, where data corresponds to $t\bar{t}$ sample generated in proton-proton collisions, we see that all models give similar results and that is what was expected. This validates our heavy-ion (HI) training setup.



(a) Tagging efficiency b (b) Tagging efficiency c (c) Tagging efficiency b (d) Tagging efficiency c

Figure 3: Comparison of tagging efficiencies and mistagging rates for different particle types in pp simulations (5.36 TeV)

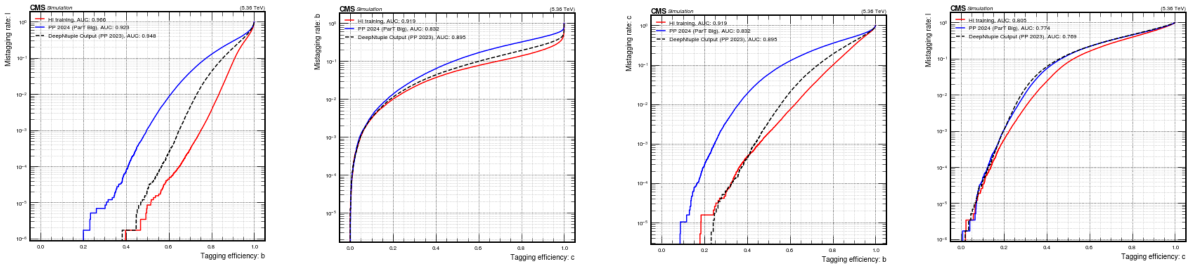


(a) Tagging efficiency b (b) Tagging efficiency c (c) Tagging efficiency b (d) Tagging efficiency c

Figure 4: Powheg embedded with Hydjet events in PbPb simulation (5.36 TeV)

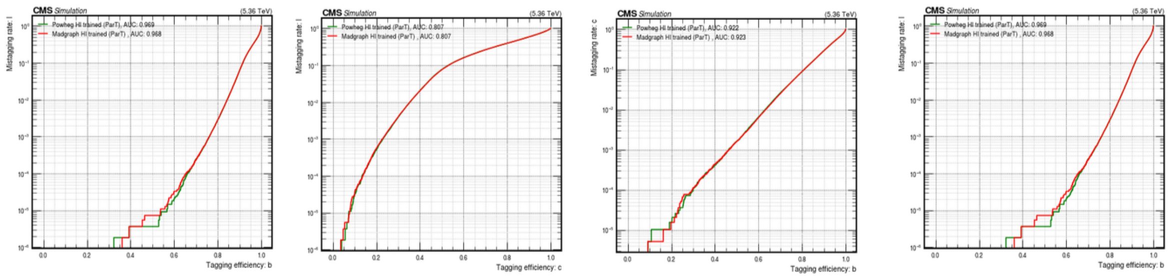
Before getting into interpretation of results, we should provide one reminder: a central PbPb collision corresponds approximately to 140 pileup events thus we expect degradation of pp algorithms. The red "line" (trained in PbPb simulation) provides a better performance than the ones trained in pp simulations (blue and black lines)

In figures 4 and 5 we see similiar (almost identical) plots for two different tools for generating singnal (Powheg and Madgraph), thus we conclude that the choice of tool is not significant. This can be further proved with next plots in figure 6.



(a) Tagging efficiency b (b) Tagging efficiency c (c) Tagging efficiency b (d) Tagging efficiency c

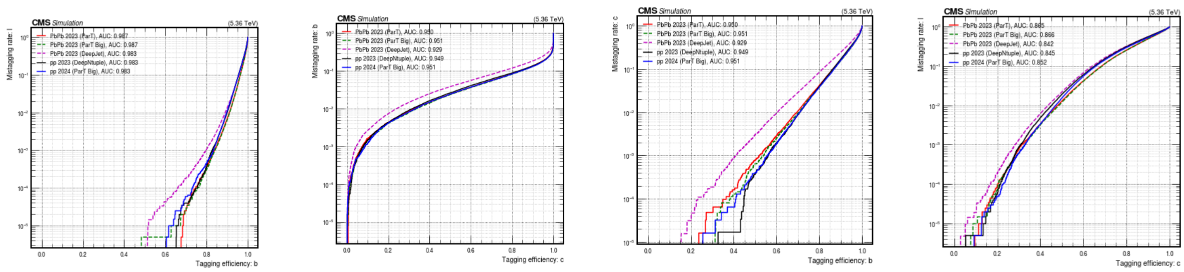
Figure 5: Madgraph embedded with Hydjet events in PbPb simulation (5.36 TeV)



(a) Tagging efficiency b (b) Tagging efficiency c (c) Tagging efficiency b (d) Tagging efficiency c

Figure 6: Madgraph trained model run on top of Powheg events

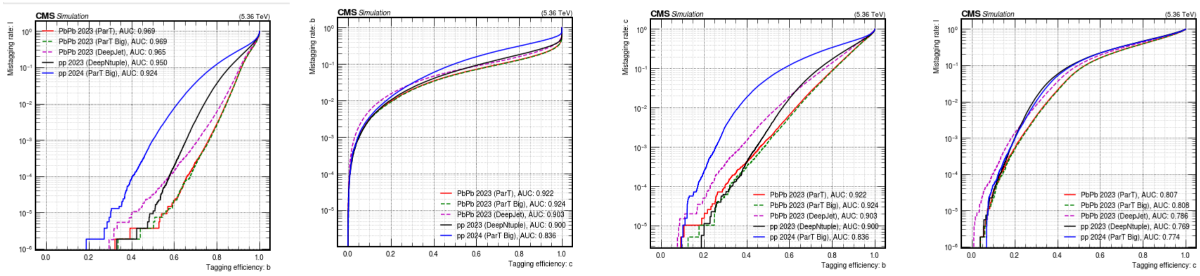
3.2 b tagging with DeepJet model



(a) tagging efficiency b (b) tagging efficiency c (c) tagging efficiency b (d) tagging efficiency b

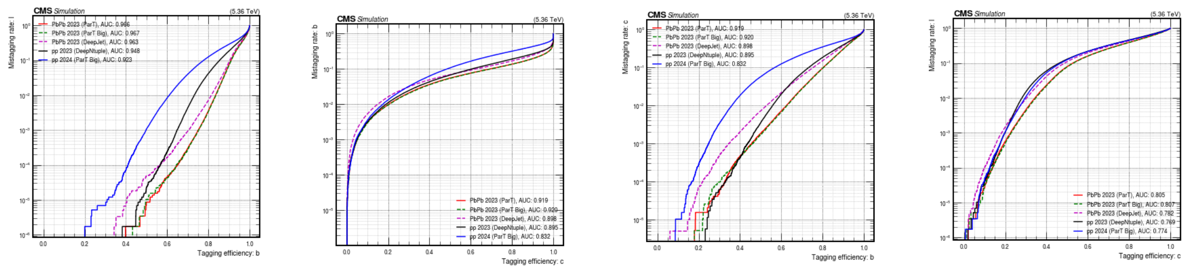
Figure 7: Comparison of tagging efficiencies and mistagging rates for different particle types in pp simulations (5.36 TeV) + DeePJet model

If we compere all those results with results in figures 3, 4 and 5, it can be seen that DeepJet model provides a worst performance than the Particle Transformer model, even if both models are trained on PbPb simulations.



(a) Tagging efficiency b (b) Tagging efficiency c (c) Tagging efficiency b (d) Tagging efficiency c

Figure 8: Powheg embedded with Hydjet events in PbPb simulation (5.36 TeV) + DeePJet model



(a) Tagging efficiency b (b) Tagging efficiency c (c) Tagging efficiency b (d) Tagging efficiency c

Figure 9: Madgraph embedded with Hydjet events in PbPb simulation (5.36 TeV) + DeePJet model

3.3 W tagging

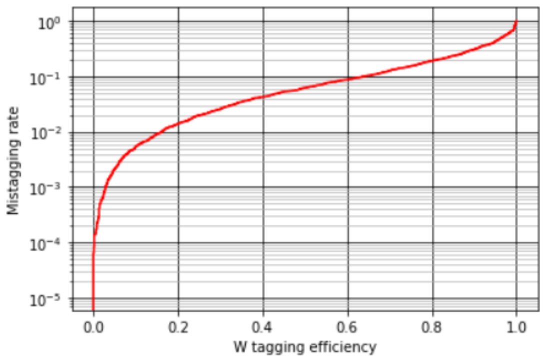
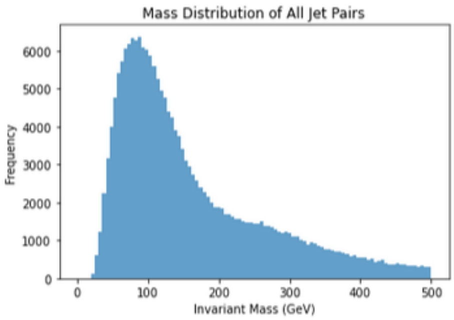
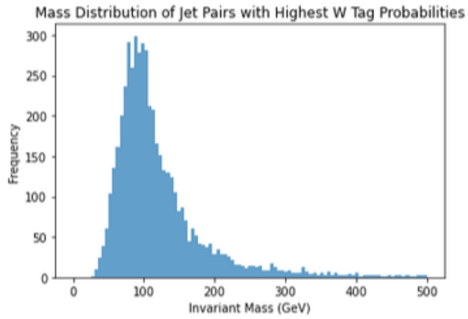


Figure 10: W tagging- ROC curve

The performance on the W tagger derived by training a BDT on the $t\bar{t}$ PbPb simulations is displayed in figure 10. Overall we can achieve 90% rejection of background with a 60% W tagging efficiency. The application of the W tagger is further checked in figure 11, where we can see the plot of the mass distribution of the pair of jets. In first case, we consider all jet pairs and in second only jet pairs with highest W tag probabilities. As we can tell there is a big difference and that significant difference is in signal vs background ratio. In 11b we have more W boson signals and in first we have all pairs not regarding their origin. Basically,the 11a has both the signal and the background, but the background dominates and that is why there is continuous distribution.



(a) All jet pairs



(b) Jet pairs with highest W tag probabilities

Figure 11: Mass distribution

4 Summary

In proton-proton (pp) collisions, all models exhibit similar tagging efficiencies for $t\bar{t}$ samples, confirming the validity of the heavy-ion (HI) training setup. This consistency is observed across different particle types, including b and c quarks, as well as in mistagging rates. In lead-lead (PbPb) collisions, simulations using both the Powheg and MadGraph generators yield nearly identical results, suggesting that the choice of simulation tool is not significant. Notably, the model trained on PbPb simulations outperforms models trained on pp simulations, as expected due to the higher complexity of HI collisions. Additionally, the W tagging performance shows that the mass distribution of jet pairs differs significantly between all jet pairs and those with the highest W tag probabilities, emphasizing the crucial role of accurate tagging and isolating the W boson signal from the background.

Bibliography

[1] David J. Griffiths, "Introduction to Elementary Particles," *Wiley-VCH*,1987.

[2] A. M. Sirunyan *et al.*, "Observation of Top Quark Production in Proton-Nucleus Collisions," *Physical Review Letters*, vol. 119, no. 23, p. 232001, 2017.

[3] A. M. Sirunyan *et al.*, "Evidence for Top Quark Production in Nucleus-Nucleus Collisions," *Physical Review Letters*, vol. 125, no. 4, p. 042001, 2020.

[4] The ATLAS Collaboration, "Observation of $\bar{t}t$ Production in Lepton+Jets and Dilepton Channels in p+Pb Collisions at $\sqrt{s_{NN}} = 8.16$ TeV with the ATLAS Detector," ATLAS Public Note, 2023.

[5] CMS Machine Learning Documentation, "ParticleNet," `https://cms-ml.github.io/documentation/inference/particlenet.html` .

[6] The CMS Collaboration, "Evidence for top quark production in nucleus-nucleus collisions," Phys. Rev. Lett. 125 (2020) 222001. `https://cms-results.web.cern.ch/cms-results/public-results/publications/HIN-19-001/index.html#Tab`

[7] A. M. Sirunyan *et al.*, "Measurement of the Nuclear Modification Factor of B-Jets in PbPb Collisions at $\sqrt{s_{NN}} = 5.02$ TeV," *Physical Review Letters*, vol. 120, no. 23, p. 232301, 2018.

[8] CMS Collaboration, "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV," *Journal of Instrumentation*, vol. 13, no. 05, P05011, 2018. Published by IOP Publishing for Sissa Medialab.

[9] E. Bols, J. Kieseler, M. Verzetti, M. Stoye, and A. Stakia, "Jet Flavour Classification Using DeepJet," *Journal of Instrumentation*, prepared for submission.

[10] H. Qu, C. Li, and S. Qian, "Particle Transformer for Jet Tagging," 2022. [Online]. Available: `https://arxiv.org/abs/2202.03772`.

[11] A. Vaswani *et al.*, "Attention Is All You Need," *arXiv preprint arXiv:1706.03762*, submitted on 12 Jun 2017 (v1), last revised 2 Aug 2023 (v7).

[12] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-k_T jet clustering algorithm," *Journal of High Energy Physics*, vol. 04, p. 063, 2008. Published by Institute of Physics Publishing for SISSA. Received: February 20, 2008; Accepted: April 4, 2008; Published: April 16, 2008.

[13] ALICE Collaboration, K. Aamodt *et al.*, "The ALICE experiment at the CERN LHC," *Journal of Instrumentation*, vol. 3, no. 08, p. S08002, 2008. DOI: 10.1088/1748-0221/3/08/S08002. Published by IOP Publishing for SISSA.

[14] ATLAS Collaboration, "The ATLAS Experiment at the CERN Large Hadron Collider," *Journal of Instrumentation*, vol. 3, no. 08, p. S08003, 2008. DOI: 10.1088/1748-0221/3/08/S08003. Published by IOP Publishing for SISSA.

[15] CMS Collaboration, "The CMS Experiment at the CERN LHC," *Journal of Instrumentation*, vol. 3, no. 08, p. S08004, Aug 2008. DOI: 10.1088/1748-0221/3/08/S08004. Published by IOP Publishing for SISSA.

[16] L. Apolinário, J. G. Milhano, G. P. Salam, and C. A. Salgado, "Probing the time structure of the quark-gluon plasma with top quarks," *Physical Review Letters*, vol. 120, no. 23, p. 232301, 2018. DOI: 10.1103/PhysRevLett.120.232301. e-Print: arXiv:1711.03105 [hep-ph]. Published: Jun 9, 2018. Report number: CERN-TH-2017-237.

[17] J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer, and T. Stelzer, "MadGraph 5: Going Beyond," *Journal of High Energy Physics*, vol. 06, p. 128, 2011, FERMILAB-PUB-11-448-T. DOI: 10.1007/JHEP06(2011)128. e-Print: 1106.0522 [hep-ph].

[18] S. Alioli, P. Nason, C. Oleari, and E. Re, "A General Framework for Implementing NLO Calculations in Shower Monte Carlo Programs: the POWHEG BOX," *Journal of High Energy Physics*, vol. 06, p. 043, 2010, DESY-10-018, SFB-CPP-10-22, IPPP-10-11, DCPT-10-22. DOI: 10.1007/JHEP06(2010)043. e-Print: 1002.2581 [hep-ph].

[19] I. P. Lokhtin and A. M. Snigirev, “A Model of Jet Quenching in Ultrarelativistic Heavy Ion Collisions and High-p(T) Hadron Spectra at RHIC,” *European Physical Journal C*, vol. 45, pp. 211-217, 2006. DOI: 10.1140/epjc/s2005-02426-3. e-Print: hep-ph/0506189 [hep-ph].

[20] S. Agostinelli *et al.* (GEANT4 Collaboration), “GEANT4—A Simulation Toolkit,” *Nuclear Instruments and Methods in Physics Research Section A*, vol. 506, pp. 250-303, 2003. DOI: 10.1016/S0168-9002(03)01368-8. Report numbers: SLAC-PUB-9350, FERMILAB-PUB-03-339, CERN-IT-2002-003.

[21] David Goldenberg, “Standard Model Standard Infographic,” *Medium*, `https://daveg.medium.com/standard-model-standard-infographic-fbf9ad981b30`.

[22] B. Işıldak, Alper Hayreter, M. Hüdaverdi, F. Ilgın, S. Salva, Ebru Simsek, S. Güyer, “Investigating the Violation of Charge-parity Symmetry Through Top-quark ChromoElectric Dipole Moments by Using Machine Learning Techniques,” *Acta Phys. Polon. B*, vol. 54, pp. 5-A4, 2023. DOI: 10.5506/APhysPolB.54.5-A4. e-Print: 2306.11683 [hep-ph].