Implementation of a lepton identification algorithm using machine learning for measurements of the tt+X associated production process in the CMS experiment at the LHC

Master Thesis

Katris Panagiotis

Three Member Examination Committee :

Konstantinos Vellidis , Associate Professor, Physics Department, NKUA- Main Supervisor

Paris Sphicas, Professor, Physics Department, NKUA

Fotios Diakonos, Professor, Physics Department, NKUA

National & Kapodistrian University of Athens

Department of Physics

June, 2024

Abstract

The goal of this thesis is the development of techniques for the identification and selection of leptons originating from top quark decays via a subsequent leptonic W boson decay. In the associated production of top quarks with vector bosons, prompt leptons can also emerge from the associated vector boson decay. Leptons from hadronic decays are considered as non-prompt and appear to be less isolated than the prompt ones. These non-prompt leptons are among the most important backgrounds in processes that have multilepton final states. This thesis focuses on the implementation of a machine learning algorithm for lepton discrimination in tt+X processes at CMS with emphasis to the $t\bar{t}H(c\bar{c})$, aiming to reduce the non-prompt lepton background and retaining at the same time maximum prompt lepton selection efficiency.

Περίληψη

Στόχος της παρούσας εργασίας είναι η ανάπτυξη τεχνικών για την ταυτοποίηση και την επιλογή λεπτονίων που προέρχονται από διασπάσεις top κουάρκ μέσω μιας επακόλουθης λεπτονικής διάσπασης μποζονίου W. Κατά τη συσχετισμένη παραγωγή top κουάρκ με διανυσματικά μποζόνια, prompt λεπτόνια μπορούν επίσης να προκύψουν από τη διάσπαση των συσχετισμένων μποζονίων. Τα λεπτόνια από αδρονικές διασπάσεις θεωρούνται non-prompt και εμφανίζονται λιγότερο απομονωμένα από τα prompt λεπτόνια. Αυτά τα non-prompt λεπτόνια είναι από τα πιο σημαντικά υπόβαθρα σε διεργασίες που έχουν πολυλεπτονικές τελικές καταστάσεις. Η παρούσα εργασία εστιάζει στην υλοποίηση ενός αλγορίθμου μηχανικής μάθησης για τη διάκριση λεπτονίων σε διεργασίες tt+X στο CMS με έμφαση στη $t\bar{t}H(c\bar{c})$ διεργασία, με στόχο τη μείωση του non-prompt λεπτονικού υποβάθρου, διατηρώντας ταυτόχρονα της μέγιστη απόδοση στην επιλογή των prompt λεπτονίων.

Contents

1 The Standard Model **5**

1.1 Particles of the Standard Model 5
1.2 Quantum Electrodynamics 7
1.3 Electroweak interaction 9
1.4 Quantum Chromodynamics (QCD) 12
1.5 The Higgs mechanism 14
1.6 Fermion mass and Yukawa interaction 18
1.7 Higgs Boson production modes 20
1.7.1 Gluon Fusion (ggH) 22
1.7.2 Vector Boson Fusion (VBF) 22
1.7.3 Vector-boson associated production (VH) 22
1.7.4 Top-Quark associated production ($t\bar{t}H$) 22
1.8 Higgs Boson Decay Modes 23
1.8.1 Decay into Heavy Fermion Pairs 23
1.8.2 Decay into Gauge Boson Pairs 24
1.8.3 Loop-Induced Decays 24
1.8.4 Higgs to charm coupling 25
1.9 T**he top quark** 26

2 The CMS Experiment **29**

2.1 The LHC 29
2.2 The CMS Detector 32
2.3 Coordinate System & Kinematics 33
2.4 The Structure of the CMS Detector 36
2.4.1 The Magnet 37
2.4.2 Inner Tracking System 39
2.4.3 Electromagnetic Calorimeter (ECAL) 42
2.4.4 Hadron Calorimeter (HCAL) 44
2.4.5 Muon Detectors 46
2.4.6 Forward Detectors 50
2.5 Trigger and DAQ system 50
2.6 Object reconstruction and identification 52
2.6.1 Electron reconstruction 52
2.6.2 Muon reconstruction 53
2.7 Trigger performance in 2023 55

3 $t\bar{t}H(c\bar{c})$ Analysis /Framework **60**
3.1 The ParticleNet Tagger . . . 61
3.2 Particle Transformer and Event Classifier . . . 64
3.3 Lepton Identification . . . 71
3.3.1 Muon identification . . . 72
3.3.2 Electron identification . . . 73

4 Boosted Decision trees **75**
4.1 Decision trees . . . 75
4.1.1 Algorithm . . . 75
4.1.2 Hyperparameters . . . 76
4.2 Boosting algorithm . . . 78
4.3 Gradient boosting . . . 79

5 Lepton MVA identification **81**
5.1 Input features . . . 81
5.2 ROC Curves . . . 88
5.3 Comparison with the existing ID . . . 89
5.3.1 Single-lepton channel . . . 91
5.3.2 Dilepton channel . . . 97

6 Results **104**
6.1 Implementation of the new lepton ID . . . 104
6.2 Comparison variables . . . 106
6.2.1 Dilepton Channel-$t\bar{t}$ . . . 107
6.2.2 Dilepton Channel-$t\bar{t}H(c\bar{c})$ . . . 108
6.2.3 Single-lepton Channel-$t\bar{t}$ . . . 108
6.2.4 Single-lepton Channel-$t\bar{t}H(c\bar{c})$ . . . 109
6.3 Data/MC plots . . . 111
6.3.1 Dilepton channel . . . 111
6.3.2 Single-lepton channel . . . 127
6.4 Sensitivity and upper limits . . . 135

7 Summary **137**

1 The Standard Model

1.1 Particles of the Standard Model

The Standard Model (SM) [1] stands out as the most comprehensive theory shedding light on the fundamental composition of matter. Every recognized particle engages in interactions through the four primary natural forces: electromagnetic, strong nuclear, weak nuclear, and gravitational. Notably, the electromagnetic and gravitational forces exhibit infinite ranges, whereas the strong and weak forces exert influence solely at the subatomic particle level. In the realm of particle physics, our primary focus lies in scrutinizing the impacts of electromagnetic, weak, and strong nuclear interactions. It's noteworthy that, as of now, gravity has not been successfully quantized in this field of study [2].

To quantify the potency of each force, a dimensionless constant, referred to as the coupling constant, is employed. The approximate magnitudes of these constants are conveniently outlined in Table 1.

Gravitational	$\alpha_g \sim \mathcal{O}(10^{-37} - 10^{-43})$
Weak	$\alpha_w \sim \mathcal{O}(10^{-6})$
Electromagnetic	$\alpha \sim \mathcal{O}(10^{-2})$
Strong	$\alpha_s \sim \mathcal{O}(1)$

Table 1: The order of magnitude for the coupling constants of the fundamental forces in nature.[3]

The gravitational force significantly lags behind in strength, spanning orders of magnitude less than other forces. As previously highlighted, physicists have not successfully incorporated gravity into the Standard Model. Consequently, the Standard Model encompasses the electromagnetic, strong, and weak forces, complete with their respective carrier particles. This framework adeptly elucidates the interactions of these forces with all matter particles.

Specifically, the elementary particles constituting observable matter in the known universe are fermions, characterized by a spin of $1/2$, and their corresponding antiparticles. The interaction between fermions is conducted by the exchange of gauge bosons, force-carrier particles endowed with integer spins.

In particle physics, particularly within the framework of the Standard Model, a cornerstone equation is the Dirac equation of relativistic quantum mechanics. This equation

intricately delineates the dynamics and interactions governing fermions and their corresponding antiparticles.

$$(i\gamma^{\mu}\partial_{\mu} - m)\psi = 0 \tag{1}$$

where ψ is the Dirac spinor of the fermion, m its mass and γ^{μ} (μ=1,2,3,4) are the four Dirac γ-matrices. With this equation we can really accurately describe the dynamics of all the known SM Fermions.

Fermions (particles that have half-integer spin [4]) are further divided into two categories; quarks and leptons. Elementary particles of the former type engage in interactions through all three forces defined within the Standard Model (SM). They serve as the building blocks for baryons, exemplified by protons and neutrons, as well as mesons. In contrast, leptons exclusively interact with the electromagnetic and weak forces, constituting particles like electrons and neutrinos.

Both categories encompass six particles, organized into three generations based on their masses. The 1st generation, being the lightest, comprises particles that collectively form all stable matter observed in the universe. As we get to higher generations the particle masses get progressively larger.To this point of time there are three known Fermions' generations. A detailed list of these particles is presented in Table 2.

Generation	**Particle**	**Symbol**	**Charge (e)**	**Mass (MeV/c^2)**
4*1^{st}	up quark	u	$+\frac{2}{3}$	2.2
	down quark	d	$-\frac{1}{3}$	4.7
	electron	e^-	-1	0.511
	electron neutrino	ν_e	0	< 0.000001
4*2^{nd}	charm quark	c	$+\frac{2}{3}$	1,280
	strange quark	s	$-\frac{1}{3}$	96
	muon	μ^-	-1	105.7
	muon neutrino	ν_μ	0	< 0.00019
4*3^{rd}	top quark	t	$+\frac{2}{3}$	173,100
	bottom quark	b	$-\frac{1}{3}$	4,180
	tau	τ^-	-1	1,776.9
	tau neutrino	ν_τ	0	< 0.018

Table 2: Quarks and Leptons in the Standard Model of elementary particles, including their charges and masses.

Concerning bosons, each fundamental force arises from the exchange of force-carrier particles. Specifically, the electromagnetic force is mediated by the neutral photon (γ) that does not interact with itself, the strong force by the gluon (g) - which interacts with itself (making the study of the strong force more complicated) , and the weak force by the charged bosons $W^\pm$ and the neutral Z. All these force carriers possess a spin of 1 [4].

1.2 Quantum Electrodynamics

In Quantum Electrodynamics (QED), all fundamental interactions among elementary particles stem from gauge transformations governed by specific gauge symmetries. In the context of QED, the Lagrangian describing electromagnetic interaction is formulated by ensuring that the Lagrangian governing a single fermion remains invariant under the U(1) electromagnetic gauge transformation.

The unified weak and electromagnetic interaction arises from the requirement of gauge invariance for the Lagrangian under the transformation of the $SU(2)_L \otimes U(1)_Y$ symmetry

group. Similarly, the strong interaction is accounted for by ensuring gauge invariance for the Lagrangian of an individual quark under the transformation of the $SU(3)_C$ symmetry group where the index c denotes the color charge. Consequently, within the Standard Model (SM), all three fundamental forces are encapsulated as its Lagrangian is constructed to maintain gauge invariance under the transformations of the $SU(3) \otimes SU(2)_L \otimes U(1)_Y$ symmetry groups.

Starting from the Eq 1:

$$\mathscr{L}_{\text{Dirac}} = \bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi = -m\bar{\psi}\psi + i\bar{\psi}\gamma^\mu \partial_\mu \psi \tag{2}$$

This Lagrangian remains invariant under global $U(1)$ transformations. However, a global transformation applied to the field at a single point in spacetime should not have universal implications, as this would violate special relativity where information cannot propagate faster than light. Thus, for local transformations, the field must be modified to consider its dependence on continuous spacetime coordinates x^μ:

$$\psi \rightarrow \psi' = e^{ia(x)Q}\psi \tag{3}$$

where $a(x)$ is an arbitrary function dependent on spacetime and Q is the electromagnetic charge operator. However, the resulting transformed Lagrangian is not locally $U(1)$ invariant. To rectify this, one introduces gauge invariance by appending an extra term to the Lagrangian. This is achieved by including an interaction term between a spin-1 field and the spin-$\frac{1}{2}$ field, given by:

$$\mathscr{L}_{\text{EM-int}} = QA_\mu \bar{\psi}\gamma^\mu \psi \tag{4}$$

where g is an arbitrary coupling constant determining the interaction strength between the two fields. The gauge transformation dictates that the spin-1 gauge field A_μ (representing the physical photon field) transforms as:

$$A_\mu \rightarrow A'_\mu = A_\mu + \partial_\mu a(x) \tag{5}$$

ensuring the resulting Lagrangian remains locally $U(1)$ invariant. This leads to the Maxwell Lagrangian:

$$\mathscr{L}_{\text{Maxwell}} = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} \tag{6}$$

where $F_{\mu\nu}$ is the electromagnetic field tensor.

The field strength tensor $F_{\mu\nu}$ is defined as:

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \tag{7}$$

The Maxwell Lagrangian is a special case of the Proca Lagrangian, neglecting the mass term $m^2 A_\mu A^\mu$, as a consequence of the gauge transformation in Eq. 5. By combining all the $U(1)$ gauge-invariant Lagrangian terms from Eqs. 2, 4, and 6, we obtain the QED Lagrangian describing the electromagnetic interaction between fermions and the photon field:

$$\mathscr{L}_{QED} = \mathscr{L}_{\text{Dirac}} + \mathscr{L}_{\text{EM-int}} + \mathscr{L}_{\text{Maxwell}} = \bar{\psi}(i\gamma^\mu D_\mu - m)\psi - g\bar{\psi}\gamma^\mu\psi A_\mu - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} \tag{8}$$

To simplify the notation, it is customary and mathematically motivated to introduce the covariant derivative to replace the ordinary derivative, given by:

$$\partial_\mu \rightarrow D_\mu \equiv \partial_\mu - igA_\mu \tag{9}$$

With this, Eq. 8 can be expressed as:

$$\mathscr{L}_{QED} = -m\bar{\psi}\psi + i\bar{\psi}\gamma^\mu D_\mu\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} \tag{10}$$

Noether's theorem establishes a connection between the transformation in Eq. 3 under the internal symmetry $U(1)$ and a conserved quantity. The Noether current corresponds to the electric four-current:

$$j^\mu_{EM} = -g\bar{\psi}\gamma^\mu\psi \tag{11}$$

And the conserved quantity, the electric charge Q, is determined by integrating j^0_{EM}:

$$Q = -g\int d^3x\bar{\psi}\gamma^0\psi \tag{12}$$

In the quantum framework, ψ is associated with the probability amplitude, requiring the total probability $\int d^3x\bar{\psi}\gamma^0\psi = 1$. Thus, Eq. 12 implies that the coupling strength $-g$ is indeed the conserved quantity, which for electromagnetism represents the electric charge.

1.3 Electroweak interaction

Inspired by the successful identification of a $U(1)$ internal symmetry in the Dirac Lagrangian, one seeks to determine if any other internal symmetries exist in the Lagrangian. It is found that an internal symmetry can be identified for two massless spin-$\frac{1}{2}$ fields by adding twice the Dirac Lagrangian shown in Eq. 2 without the mass term:

$$\mathscr{L}'_{D1+D2} = i\bar{\Psi}\gamma^\mu\partial_\mu\Psi \tag{13}$$

where the Lagrangian acts on the doublet Ψ consisting of two Dirac spinors:

$$\Psi = \begin{pmatrix}\psi_1\\ \psi_2\end{pmatrix} \tag{14}$$

Each spinor is a Lorentz invariant object constructed by a pair of left-handed and right-handed Weyl spinors, and the linear combination in terms of left-chiral and right-chiral states is given by:

$$\psi = \begin{pmatrix} \chi_L \\ \zeta_R \end{pmatrix} = \begin{pmatrix} \chi_L \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \zeta_R \end{pmatrix} \tag{15}$$

The left-chiral and right-chiral states can be projected out of the doublet Ψ using the projection operators $P_{L,R}$:

$$P_L\psi = P_L\begin{pmatrix} \chi_L \\ \zeta_R \end{pmatrix} = \begin{pmatrix} \chi_L \\ 0 \end{pmatrix} \equiv \psi_L \tag{16}$$

$$P_R\psi = P_R\begin{pmatrix} \chi_L \\ \zeta_R \end{pmatrix} = \begin{pmatrix} 0 \\ \zeta_R \end{pmatrix} \equiv \psi_R \tag{17}$$

where $P_{L/R} = \frac{I \mp \gamma_5}{2}$, γ_5 being the chirality operator. Madame Wu's experiment indicates that only left-handed fermions exclusively interact via weak interaction [5], implying that the Lagrangian in eq. 13 is not invariant under parity transformation. To accommodate the notion of parity violation in the Standard Model (SM), the left-chiral fields are described as $SU(2)_L$ doublets, as they transform and mix into each other. Similarly, the right-chiral fields are represented by $SU(2)$ singlets as they do not transform into each other. Thus, under the electroweak gauge symmetry $SU(2)_L \otimes U(1)_Y$, the left-handed fermion fields transform as a doublet, while the right-handed fields transform as a singlet:

$$\Psi_L \to \psi_L' = e^{i\tilde{\alpha}(x)I_i + i\beta(x)Y}\psi_L \quad \text{under } SU(2)_L \tag{18}$$

$$\psi_R \to \psi_R' = e^{i\beta(x)Y}\psi_R \quad \text{under } U(1)_Y \tag{19}$$

where the sum over the index i is implicitly assumed and $\tilde{\alpha}(x) = (\alpha_1(x), \alpha_2(x), \alpha_3(x))$, $\beta(x)$ are arbitrary real functions of the space-time coordinates. There are 3 generators of the $SU(2)_L$ group, referred to as isospin vector operators, denoted as $I_i = \frac{\sigma_i}{2}$ where σ_i are the Pauli spin matrices:

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \qquad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{20}$$

The parameter Y in 19 represents the weak hypercharge operator, relating the two conserved quantities Q from Eq. 12 and I_3 according to the Gell-Mann-Nishijima relation:

$$Y = 2(Q - I_3) \tag{21}$$

The weak hypercharge for fermion fields, considering their electric charges and isospin charges, is summarized in Table 3. The weak isospin doublet is defined by the eigenstate

of $\frac{1}{2}\sigma_3$ generator, with corresponding eigenvalues of isospin charge $+\frac{1}{2}$ and $-\frac{1}{2}$. Hence, the following notation is adopted to distinguish between the handedness for three generations:

- × The left-handed lepton doublet: $l_i^L = (\nu_i, e_i)_L^T, \quad i = 1,2,3$
- × The left-handed quark doublet: $q_i^L = (u_i, d_i')_L^T, \quad i = 1,2,3$
- × The right-handed lepton singlet: down-type lepton $e_i^R, \quad i = 1,2,3$
- × The right-handed quark singlet: up-type quark $u_i^R, \quad i = 1,2,3$; down-type quark $d_i^R, \quad i = 1,2,3$

with the subscript i running over the different fermion families. The down-type quark mass eigenstate in the doublet Eq. 21 is expressed in terms of eigenstates of the weak interaction, $d_i' = \sum_j V_{ij} d_j$, where V is the Cabibbo-Kobayashi-Maskawa (CKM) mixing matrix [6], to account for the experimental observation of quark mixing.

Field	e_L	ν_e	u_L	d_L	e_R	ν_e	u_R	d_R
Hypercharge Y	-1	-1	$\frac{1}{3}$	$-\frac{2}{3}$	-2	0	$\frac{4}{3}$	$-\frac{1}{3}$

Table 3: The weak hypercharge for left-handed and right-handed fermion fields.

The Lagrangian Eq. 13 is made gauge invariant under the transformation in Eq. 18 and Eq. 19 by introducing gauge vector fields in covariant derivative form:

$$D_\mu = \partial_\mu - igB_\mu Y - ig' W_\mu \tag{22}$$

The W_μ gauge fields, defined as $W_\mu \equiv \frac{\sigma_i}{2} W_\mu^i$, where $i = 1,2,3$, are associated with the $SU(2)_L$ group, and the B_μ gauge field is associated with the $U(1)_Y$ group; g and g' are their respective gauge coupling constants. The gauge fields transform as:

$$B_\mu \to B_\mu' = B_\mu - \frac{1}{g}\partial_\mu \beta(x) \quad \text{under } U(1)_Y \tag{23}$$

$$W_\mu \to W_\mu' = W_\mu - \frac{1}{g'}\partial_\mu \alpha_i(x) - \epsilon_{ij}^k \alpha_j(x) W_\mu^k \quad \text{under } SU(2)_L \tag{24}$$

where ϵ_{ijk} is the structure constant of $SU(2)$ group given by $[\sigma_i, \sigma_j] = i\epsilon_{ijk}\sigma_k$. The respective transformations result in a locally invariant Lagrangian for each gauge field:

$$\mathscr{L}_{gauge} = -\frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} W_{\mu\nu} W^{\mu\nu} \tag{25}$$

with the field strength tensors $B_{\mu\nu}$ and $W_{\mu\nu}$ for gauge fields defined as:

$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu \tag{26}$$

$$W_{\mu\nu} = \partial_\mu W_\nu - \partial_\nu W_\mu - ig[W_\mu, W_\nu] \tag{27}$$

The last term in Eq. 27 arises from the non-Abelian nature of the $SU(2)$ group, as the generators σ_i $(i = 1, 2, 3)$ do not commute with each other. By summing the contributions from the Lagrangians of Eq. 13 and Eq. 25, and replacing the ordinary derivative ∂_μ with the covariant derivative Eq. 22, the Lagrangian can be written compactly as:

$$\mathscr{L} = i\bar{\Psi}_L \gamma^\mu D_\mu \Psi_L - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} W_{\mu\nu} W^{\mu\nu} \tag{28}$$

From the unification of two forces, we obtain the relation described in Eq. 21, implying that the A_μ field associated with the generator Q of $U(1)_Y$ must be a combination of gauge fields W_3^μ and B^μ, which are associated with the generators I_3 and Y respectively. Consequently, the mixing between the two gauge fields W_3^μ and B^μ defines the neutral weak boson Z^μ and the photon field A^μ. On the other hand, the charged weak bosons $W^\pm$ are given as a complex combination of the two gauge fields W_1^μ and W_2^μ. In summary, we have:

$$\text{the photon } A^\mu = \frac{1}{\sqrt{g'^2 + g^2}} \left(g^2 W_3^\mu + g' B^\mu \right) \tag{29}$$

$$\text{the neutral weak boson } Z^\mu = \frac{1}{\sqrt{g'^2 + g^2}} \left(g'^2 W_3^\mu - g B^\mu \right) \tag{30}$$

$$\text{the charged weak bosons } W_\pm^\mu = \frac{1}{\sqrt{2}} \left(W_1^\mu \mp i W_2^\mu \right) \tag{31}$$

and the two coupling strengths g and g' are related by the weak mixing angle or Weinberg angle, $\theta_W = \tan^{-1}(g/g')$.

The absence of mass terms for both boson and fermion fields in the form $\frac{1}{2} m^2 W^\mu W_\mu$ and $\frac{1}{2} m^2 \left(\bar{\psi}_L \psi_R + \bar{\psi}_R \psi_L \right)$ are justifiable as these terms would destroy the gauge invariance of the Lagrangian. Therefore, the Lagrangian Eq. 28 describes only a system of massless fermions and bosons. However, experimental results show that all gauge bosons of weak interactions are massive. To include the mass terms for $W^\pm$ and Z without breaking the local gauge invariance, we need the "Higgs mechanism" to account for the mass of those bosons and subsequently for the masses of the other particles, as it will be discussed in a following section.

1.4 Quantum Chromodynamics (QCD)

Quantum Chromodynamics (QCD) is the theory that describes the strong interaction among particles carrying color charge. It is formulated based on the SU(3) gauge group,

governing the interactions of quarks and gluons.

The dynamics of three massless spin-1/2 fields in QCD are described by the free Lorentz-invariant Lagrangian:

$$\mathscr{L}_{\text{quarks}} = \bar{\Psi} q(i\gamma^{\mu}\partial_{\mu} - m)\Psi_q, \tag{32}$$

where Ψ_q is a triplet of spin-1/2 fields given by

$$\Psi_q = \begin{pmatrix} \psi_{q1} \\ \psi_{q2} \\ \psi_{q3} \end{pmatrix}. \tag{33}$$

The symmetry group SU(3) describes transformations with 3x3 matrices of unit determinant. The generators of this group, denoted by T_A ($A = 1,...,8$), are Hermitian and traceless. They are related to the Gell-Mann matrices λ_A.

The addition of eight spin-1 gauge fields, G_{μ}, is introduced through the covariant derivative:

$$D_{\mu} = \partial_{\mu} - ig_s G_{\mu}, \tag{34}$$

where g_s is the gauge coupling constant. The gluon fields G_{μ} transform under SU(3) gauge transformations.

The field strength tensor $G_{\mu\nu}$ is defined as

$$G_{\mu\nu} = \partial_{\mu} G_{\nu} - \partial_{\nu} G_{\mu} - ig_s[G_{\mu}, G_{\nu}]. \tag{35}$$

The dynamical contribution from the Lagrangian of gluon fields is given by

$$\mathscr{L}_{\text{gauge}} = -\frac{1}{4}(G_{\mu\nu})^A (G^{\mu\nu})_A. \tag{36}$$

The total QCD Lagrangian is the sum of the Lorentz-invariant term and the gauge contribution:

$$\mathscr{L}_{\text{QCD}} = \mathscr{L}_{\text{quarks}} + \mathscr{L}_{\text{gauge}} = \bar{Q}(i\gamma^{\mu} D\mu - m)Q - \tfrac{1}{4}(G_{\mu\nu})^A (G^{\mu\nu})_A. \tag{37}$$

The strength of the strong force exhibits a unique characteristic: it increases as the distance between interacting particles grows, and conversely weakens as they approach each other. This behavior is described by the Quantum Chromodynamics (QCD) effective coupling constant, denoted as α_s, which quantifies the strength of the interaction between quarks and gluons. Responisble for this behaviour is the gluon-self interaction term $ig_s[G_{\mu}, G_{\nu}]$ in Eq. 35, absent in QED.

The effective coupling constant α_s is expressed as a function of the momentum transfer $(Q^2)^2$ between the quarks involved in the interaction. It follows a logarithmic dependence given by the formula:

$$\alpha_s(Q^2) = \frac{1}{b_0 \ln\left(\frac{Q^2}{\Lambda^2_{\mathrm{QCD}}}\right)}$$

Here, b_0 represents the leading term in the perturbative QCD expression, and Λ_{QCD} defines the cutoff scale of perturbative QCD validity. Empirically, Λ_{QCD} is estimated to be around 100-400 MeV, corresponding to a distance scale of approximately 1 femtometer (the scale of size of the hadrons).When the momentum transfer $(Q^2)^2$ is below the scale Λ^2_{QCD}, the theory enters the non-perturbative regime dominated by soft parton, constituent particle of a hadron (such as a proton or neutron), interactions. In contrast, at higher momentum transfer scales relevant to Large Hadron Collider (LHC) collisions (where Q^2 is much larger than Λ^2_{QCD}), the effective coupling constant α_s becomes small, indicating that quarks and gluons behave approximately as free particles. This phenomenon, known as "asymptotic freedom," allows perturbative methods to be applied effectively in computing quantitative predictions for hard scattering processes involving partons.

Confinement is a fundamental aspect of quantum chromodynamics (QCD), where the strong force between quarks and gluons becomes increasingly stronger as the distance between them increases. In response to this increasing energy, it becomes energetically favorable for new quark-antiquark ($q\bar{q}$) pairs to be created until all colored particles are ultimately confined into color-neutral states. As a result of this phenomenon, individual bare quarks and gluons cannot be observed in isolation. This process of converting colored quarks and gluons into color-neutral bound states is known as hadronization. During hadronization, the newly created $q\bar{q}$ pairs combine with existing quarks and gluons to form color-neutral composite particles called hadrons.

1.5 The Higgs mechanism

The requirement of massless gauge fields in the electroweak Lagrangian Eq. 28 contradicts experimental observations, which show that weak interactions are mediated by massive W^+, W^-, and Z gauge bosons. To preserve gauge invariance while allowing for massive gauge bosons, a new scalar potential known as the Higgs potential is introduced in the Lagrangian. By spontaneously breaking symmetry, particle masses are naturally generated through interaction with a new scalar field called the Higgs field. This collective process that generates particle masses in the Standard Model is termed the "Higgs mechanism".

To begin, we consider a locally $SU(2)$ and $U(1)$ invariant Lagrangian for scalar fields

written as:

$$\mathscr{L}_{Higgs} = (D_\mu \Phi)^\dagger (D^\mu \Phi) - V(\Phi) \tag{38}$$

where the complex scalar doublet Φ needed to generate masses for the electroweak gauge bosons consists of a neutral scalar field ϕ_0 and a charged field ϕ^+:

$$\Phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} \tag{39}$$

This minimal Higgs model features two complex scalar fields forming a weak isospin doublet, where the upper and lower components of the doublet differ by one unit of charge. The weak isospin (T, T_3) and hypercharge (Y) quantum numbers of the Higgs doublet are summarized in Table 4.

Field	Isospin T	Isospin T_3	Hypercharge Y
ϕ^+	1/2	1/2	1
ϕ^0	1/2	−1/2	0

Table 4: Properties of the Higgs complex scalar doublet in the minimal Higgs model.

The Higgs doublet Φ is transformed under the symmetry as:

$$\Phi' = e^{ib_i(x)\sigma_i/2}\Phi \tag{40}$$

where the gauge fields W_μ and B_μ in Eq. 38 are transformed according to Eqs. 23 and 24.

The Higgs potential $V(\Phi)$ is an interaction energy characterized by two independent parameters μ and λ:

$$V(\Phi) = \mu^2 \Phi^\dagger \Phi + \lambda (\Phi^\dagger \Phi)^2 \tag{41}$$

with $\mu^2 < 0$ and $\lambda > 0$, ensuring that the potential has a non-zero minimum value and is bounded from below. The quadratic nature of the potential implies that the minimized $V(\Phi)$ corresponds to a non-zero minimum field value for every value of ϕ:

$$\phi_{\text{min}} = \pm \frac{1}{\sqrt{2}} v e^{i\phi}; \quad v = \sqrt{\frac{-\mu^2}{\lambda}} \tag{42}$$

where v is the vacuum expectation value (VEV) of the Higgs field, approximated as 246 GeV. These minima lie on a circle with radius $\frac{1}{\sqrt{2}}v$, as shown in Figure 1. By choosing a specific vacuum state for the doublet from infinite possible states, the Lagrangian loses its gauge symmetry and undergoes symmetry breaking.

A convenient choice for the minimum state is:

$$\Phi_{\text{min}} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_0^+ \\ \phi_0^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix} \tag{43}$$

The fluctuations around the minimum energy can be calculated using perturbation theory by expanding the field $\phi(x)$ around the chosen vacuum state (Eq. 43). The excited state of the Higgs doublet, at first order in the series expansion, is given by:

$$\Phi(x) = \frac{1}{\sqrt{2}}(\phi_1^r(x) + i\phi_1^c(x), v + \phi_2^r(x) + i\phi_2^c(x))^T \approx e^{ib_i(x)\sigma_i/2}\frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v + h(x) \end{pmatrix} \tag{44}$$

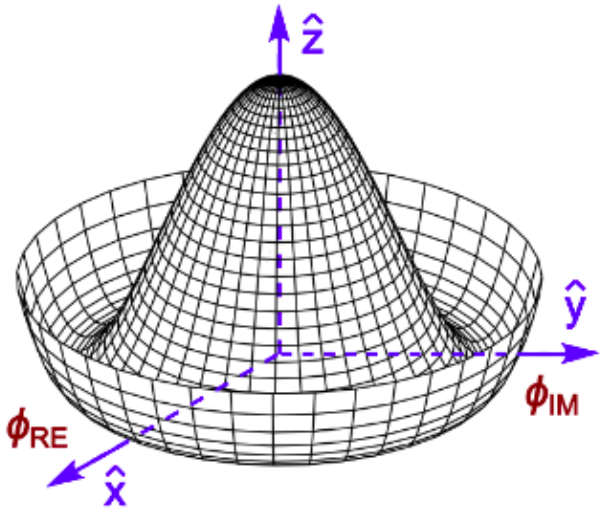


Figure 1: The Higgs potential for the complex scalar doublet field Φ describes the interaction energy along the $\hat{z}$-axis, where ϕ_{RE} and ϕ_{IM} denote the real and imaginary parts of the doublet. [7]

Parameterized by the four field components $\phi_{1r}, \phi_{1c}, \phi_{2r}, \phi_{2c}$, the exponent term in Eq. 44 can be eliminated due to gauge invariance, following the transformation given by Eq.44. By gauging away the three gauge fields $\phi_{1r}, \phi_{1c}, \phi_{2c}$ and choosing only one gauge field, the excited state in 44 is rewritten as:

$$\Phi_{un}(x) = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v + h(x) \end{pmatrix} \tag{45}$$

The unitary gauge approach implies that out of the four gauge fields in the complex scalar doublet of Eq.44, three have been transformed into the longitudinal components of the gauge bosons $W^{\pm}, Z^0$, while the remaining component, which is neutral and scalar, becomes the physical field h called the Higgs field. The excitation of the Higgs field describing the particle state above the chosen vacuum state gives rise to a new particle, the Higgs boson, which is experimentally detectable to verify the theory.

Under the framework of spontaneous symmetry breaking, the Higgs Lagrangian given by Eq.38 can be expressed in terms of the vacuum expectation value (VEV) v and the

Higgs field $h(x)$, by adding a constant term $-\frac{1}{4}\mu^2 v^2$ to the Higgs potential of Eq.41 and inserting Eq.45 into Eq.38:

$$\mathscr{L}_{\text{Higgs}} = \frac{1}{2}(\partial_\mu h)(\partial^\mu h) + \frac{1}{8}g'^2(W_\mu^1)^2 + (W_\mu^2)^2(v+h)^2 + \frac{1}{8}g'^2 W_\mu^3 - gB_\mu{}^2(v+h)^2 - V(h) \tag{46}$$

where the Higgs potential is given by:

$$V(\Phi) \rightarrow V(h) = -\mu^2 h^2 - \frac{\mu^2}{v}h^3 - \frac{\mu^2}{4v^2}h^4 \tag{47}$$

Note that choosing a particular state, Eq.43, renders the Lagrangian no longer invariant under the $SU(2)_L \otimes U(1)_Y$ gauge transformation. However, the assignment of VEV to the neutral component of the Higgs doublet ensures the conservation of electric charge and the $U(1)_y$ symmetry remains unbroken.

Finally, the full physics content of the Lagrangian is obtained by replacing the gauge fields with their respective physical fields via substitution of Eq.29, Eq.30, Eq.31 into Eq.46, adding the dynamical contribution of the gauge field Eq.25:

$$\mathscr{L} = \mathscr{L}_{\text{Higgs}} + \mathscr{L}_{\text{gauge}} \tag{48}$$

$$\begin{aligned}
\mathscr{L} &= (D_\mu \Phi_{\text{un}})^\dagger (D^\mu \Phi_{\text{un}}) - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}W_{\mu\nu}W^{\mu\nu} \\
&= \frac{1}{2}(\partial_\mu h)(\partial^\mu h) + \mu^2 h^2 + \frac{\mu^2}{v}h^3 + \frac{\mu^2}{4v^2}h^4 - \frac{1}{2}(W^+_{\mu\nu})(W^-)^{\mu\nu} + \frac{1}{4}g'^2 v^2 (W^+_\mu)(W^-)^\mu \\
&\quad - (\frac{h^2}{4} + \frac{vh}{2})g'^2(W^+_\mu)(W^-)^\mu - \frac{1}{4}Z_{\mu\nu}Z^{\mu\nu} - \frac{1}{8}v^2(g^2+g'^2)Z_\mu Z^\mu \\
&\quad - (\frac{h^2}{8} + \frac{vh}{4})(g^2+g'^2)Z_\mu Z^\mu - \frac{1}{4}A_{\mu\nu}A^{\mu\nu}
\end{aligned} \tag{49}$$

Eq.49 describes the Lagrangian for a free massive neutral scalar boson field $h(x)$, a free massive neutral vector boson field $Z_\mu(x)$, a pair of massive charged vector boson fields $W^\pm_\mu(x)$, and a massless photon field $A_\mu(x)$. From the expected mass term in the form of $\frac{1}{2}m^2\phi^\dagger\phi$, their respective masses can be identified from Eq.49 as:

$$m_h = v\sqrt{2\lambda} \tag{50}$$

$$m_W = \frac{g'v}{2} \tag{51}$$

$$m_Z = \frac{v}{2}\sqrt{g^2+g'^2} \tag{52}$$

$$m_\gamma = 0 \tag{53}$$

The electroweak bosons acquire mass through the VEV of the Higgs field and their respective gauge interaction coupling constants g, g'. The combination corresponding to the Z boson in Eq.30, which is associated with the neutral Goldstone boson of the broken symmetry, has acquired mass through the Higgs mechanism and the field corresponding to the photon has remained massless. Hence, the resulting mass $m_\gamma = 0$ is used as a consistency check for the VEV assignment in the Higgs doublet.

Concerning the Higgs mass m_h, the parameter v is fixed but λ is not experimentally constrained in the Standard Model (SM); the mass of the Higgs boson is not predicted by the theory. Therefore, at the fundamental level, the existence of the Higgs boson is indispensable to the foundation of SM.

1.6 Fermion mass and Yukawa interaction

The Lorentz invariant mass term is usually expressed as a combination of left-handed and right-handed fields:

$$\bar{\psi}\psi = \bar{\psi}_L\psi_R + \bar{\psi}_R\psi_L \tag{54}$$

However, Eq.54 is not invariant according to their respective transformation properties Eq.18, Eq.19 under $SU(2)$ due to handedness. Nevertheless, a $SU(2)\times U(1)$ and Lorentz invariant term that couples a spin 0 doublet and spin $\frac{1}{2}$ spinor together with the equally allowed Hermitian conjugate term can be constructed as:

$$\mathscr{L} = -\lambda(\bar{\Psi}_L\Phi\psi_R + \bar{\psi}_R\bar{\Phi}\Psi_L) \tag{55}$$

The coupling constant λ is called a Yukawa coupling describing not only the interaction between the fermions and the Higgs field, but also leading to mass terms for the spin $\frac{1}{2}$ fields after the symmetry breaking. Since there is a second field in the doublet Ψ, two representations of the Higgs field are needed to give masses to the down quarks and electrons, and to the up quarks.

Therefore, the non-invariance can be fixed by introducing two representations of the Higgs field into the Lagrangian Eq.54, Φ and its hermitian conjugate version given as $\tilde{\Phi}_i = \epsilon_{ij}\Phi_j^*$:

$$\Phi = \begin{pmatrix}\phi^+\\ \phi^0\end{pmatrix} \equiv \begin{pmatrix}0\\ \frac{v+h}{\sqrt{2}}\end{pmatrix} \quad \text{with} \quad Y(\Phi) = \frac{1}{2} \tag{56}$$

$$\tilde{\Phi} = \begin{pmatrix}\phi^0\\ \phi^-\end{pmatrix} \equiv \begin{pmatrix}\frac{v+h}{\sqrt{2}}\\ 0\end{pmatrix} \quad \text{with} \quad Y(\tilde{\Phi}) = -\frac{1}{2} \tag{57}$$

As a result, by using the usual notation for leptons and for quarks, the Yukawa Lagrangian describing the interaction between the Higgs and fermion fields is given by:

$$\mathscr{L}_{\text{Yukawa}} = \mathscr{L}_{\text{lepton Yukawa}} + \mathscr{L}_{\text{quark Yukawa}} \tag{58}$$

$$\begin{aligned}\mathscr{L}_{\text{lepton Yukawa}} = &-\lambda_1^i(\bar{l}_L^i \Phi e_R^i + \bar{e}_R^i \Phi^\dagger l_L^i) \\ &- \lambda_2^i(\bar{l}_L^i \tilde{\Phi} \nu_R^i + \bar{\nu}_R^i \tilde{\Phi}^\dagger l_L^i)\end{aligned} \tag{59}$$

$$\begin{aligned}\mathscr{L}_{\text{quark Yukawa}} = &-\lambda_d^i(\bar{q}_L^i \Phi d_R^i + \bar{d}_R^i \Phi^\dagger q_L^i) \\ &- \lambda_u^i(\bar{q}_L^i \tilde{\Phi} u_R^i + \bar{u}_R^i \tilde{\Phi}^\dagger q_L^i)\end{aligned} \tag{60}$$

where the Yukawa couplings, λ_i with $i = 1, 2, 3$, correspond to the flavor interaction eigenstates. Note that the neutrino has no right-handed partner in the Standard Model (SM), it will not acquire a mass term through the Yukawa coupling, and thus the terms are omitted. By choosing the Vacuum Expectation Value (VEV) for the spin-0 field (breaking the symmetry), which is equivalent to inserting Eq. 45 into the Lagrangian Eq. 58,after expanding the fields at VEV, the lepton sector reads:

$$\mathscr{L}_{\text{lepton Yukawa}} = -\lambda_1^i(\bar{l}_L^i \Phi_{un} e_R^i) + \text{h.c.} = -\lambda_1^i \frac{v}{\sqrt{2}} \bar{e}_L^i e_R^i - \lambda_1^i \frac{h}{\sqrt{2}} \bar{e}_L^i e_R^i + \text{h.c.} \tag{61}$$

Analogously, the quark sector becomes:

$$\mathscr{L}_{\text{quark Yukawa}} = -\lambda_d^i \bar{q}_L^i \Phi d_R^i - \lambda_u^i \bar{q}_L^i \tilde{\Phi} u_R^i + \text{h.c.} = -\lambda_d^i \frac{v}{\sqrt{2}} \bar{d}_L^i d_R^i - \lambda_u^i \frac{v}{\sqrt{2}} \bar{u}_L^i u_R^i - \lambda_d^i \frac{h}{\sqrt{2}} \bar{d}_L^i d_R^i - \lambda_u^i \frac{h}{\sqrt{2}} \bar{u}_L^i u_R^i + \text{h.c.} \tag{62}$$

Remarkably, by the same fashion, the lepton mass terms can be deduced from Eq. 61 as:

$$m_e = \lambda_e \frac{v}{\sqrt{2}}; \quad m_\mu = \lambda_\mu \frac{v}{\sqrt{2}}; \quad m_\tau = \lambda_\tau \frac{v}{\sqrt{2}} \tag{63}$$

where $\lambda_e, \lambda_\mu, \lambda_\tau$ are dimensionless Yukawa couplings for the electron, muon, and tau lepton, respectively. While for the 6 flavor quark mass terms are inferred from Eq. 62 as:

$$m_u = \lambda_u \frac{v}{\sqrt{2}}; \quad m_c = \lambda_c \frac{v}{\sqrt{2}}; \quad m_t = \lambda_t \frac{v}{\sqrt{2}}; \quad m_d = \lambda_d \frac{v}{\sqrt{2}}; \quad m_s = \lambda_s \frac{v}{\sqrt{2}}; \quad m_b = \lambda_b \frac{v}{\sqrt{2}} \tag{64}$$

Through the Yukawa Lagrangians Eq. 61 and Eq. 62, the charged leptons and quarks not only acquire mass from the VEV of the Higgs field; in addition, an interaction term is introduced between the fermion fields and the Higgs fields, with a coupling strength proportional to the fermion mass given as:

$$\lambda = \frac{\sqrt{2} m_f}{v} \tag{65}$$

In summary, the possible Higgs boson interactions with Standard Model particles under the symmetry group $SU(3)_C \times SU(2)_L \times U(1)_Y$ are characterized by the Feynman diagrams shown in Figure 2.

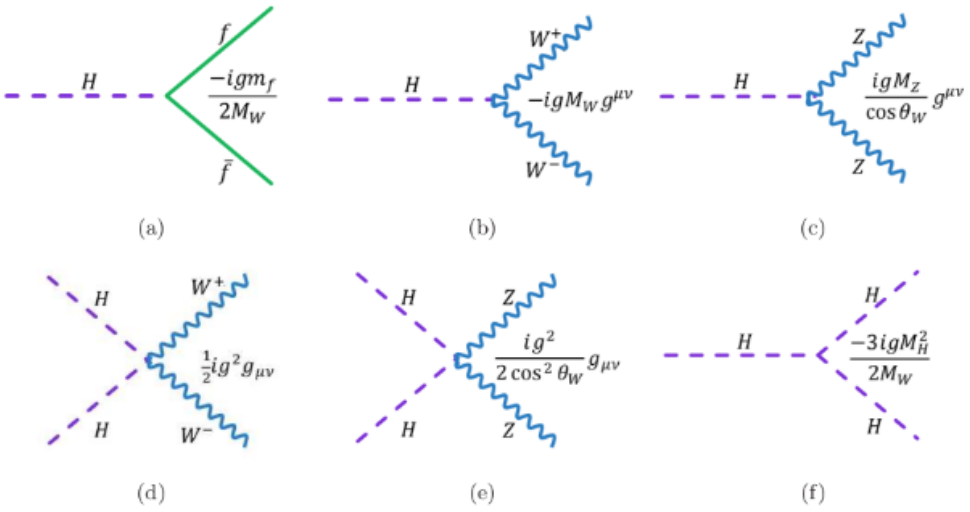


Figure 2: The Feynman diagrams for Higgs bosons interaction with itself, and to fermions and gauge bosons. [8]

The signatures for the Standard Model (SM) Higgs boson detection in collider experiments depend on the production process and the decay pattern. Both production and decay contribute to specific kinematic features that can be used to distinguish signal from background events. Numerous experimental searches for the Higgs boson were conducted during the first run period of the Large Hadron Collider (LHC) with proton-proton collisions at $\sqrt{s}$ = 7 and 8 TeV. On July 4, 2012, the observation of a new particle with Higgs boson-like properties at a mass of approximately 125 GeV was reported by the ATLAS [9] and CMS [10],[11] Collaborations, with integrated luminosity of 5.1 (for 7 TeV) and 5.3 (for 8 TeV) inverse femtobarns (fb^{-1}). Subsequent measurements of the new particle properties, such as spin, parity, and coupling strength to SM particles, are consistent within experimental uncertainties with the expectation for the SM Higgs boson [12],[13],[14],[15]. The full Run-II dataset collected during LHC proton-proton collisions at a higher energy $\sqrt{s}$ = 13 TeV provides more statistical power to further constrain the Higgs boson mass measurement and elucidate other production and decay modes.

1.7 Higgs Boson production modes

The production of the Higgs boson in proton-proton (pp) collisions occurs through four main mechanisms, with corresponding Feynman diagrams shown in Figure 3 and

their production cross-section presented in Figure 4.

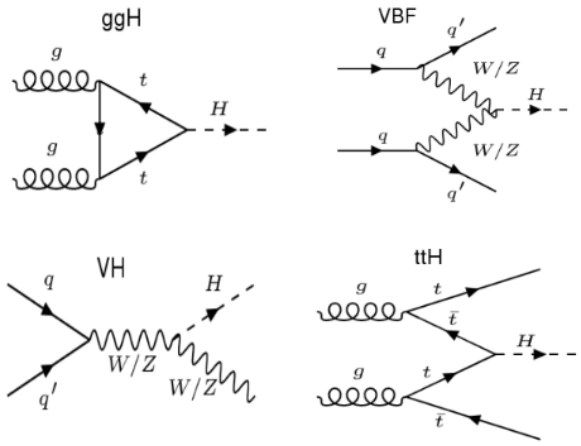


Figure 3: The leading order Feynman diagrams contributing to the Higgs boson production in pp interaction: gluon fusion (ggH), vector-boson fusion (VBF), vector boson associated production (VH), and top quark associated production (ttH).

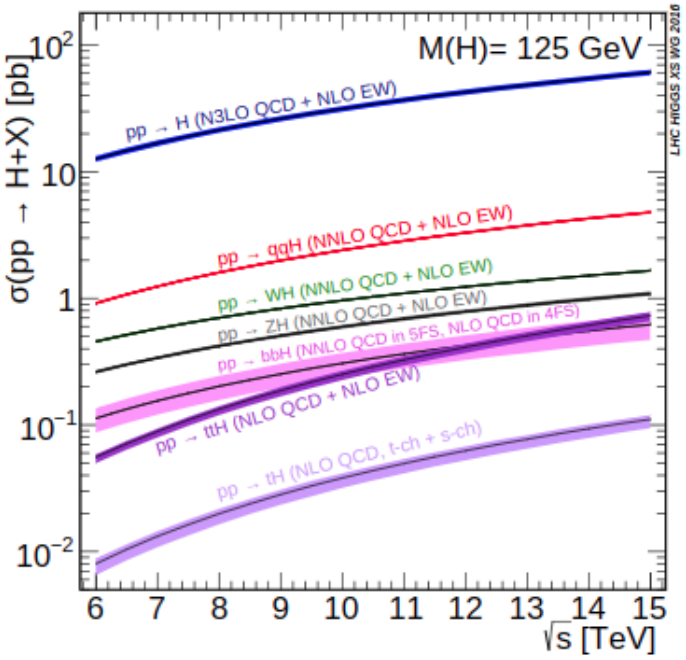


Figure 4: Theoretical prediction of SM Higgs boson production cross-section as a function of $\sqrt{s}$ for proton-proton collisions, showing major production modes and associated uncertainties.[16]

1.7.1 Gluon Fusion (ggH)

At tree level, the Higgs boson hardly couples to light-flavor quarks due to the proportionality of coupling strength to their masses, and it does not couple directly to gluons because of the unbroken $SU(3)_C$ symmetry of Quantum Chromodynamics (QCD). However, the dominant production mechanism for the Higgs boson at the LHC is gluon fusion (ggH). This is enabled by loop-induced processes involving gluons, with the top quark typically running in the loop due to its strong coupling to gluons and its large Yukawa coupling to the Higgs boson [16]. The production mode can be symbolically represented as $pp \to H$.

1.7.2 Vector Boson Fusion (VBF)

The SM Higgs boson has significant couplings to W and Z bosons and to top quarks, which allow it to be produced via their gauge boson interactions. In the VBF process, the two incoming quarks radiate W or Z bosons, which then merge to produce a Higgs boson. This mode is characterized by the second-largest production cross-section at the LHC. Final states in VBF typically exhibit two jets with high rapidity in the forward and backward regions of the detector, with suppressed QCD radiation between the jets. The production mode can be symbolically represented as $qq \to qqH$.

1.7.3 Vector-boson associated production (VH)

Associated production processes (VH) involve the production of a Higgs boson together with a W or Z boson. Although the production cross-section for VH processes is smaller than VBF, it is experimentally feasible due to the identifiable leptons and neutrinos from W or Z boson decays, which aid in selecting signal events. The production mode can be symbolically represented as $qq \to VH$ with $V = W, Z$.

1.7.4 Top-Quark associated production ($t\bar{t}H$)

The $t\bar{t}H$ production mode allows for a direct measurement of the Yukawa coupling between the top quark and the Higgs boson. Despite its relatively low production cross-section and complex final state involving a large number of hadronic jets, advancements in multivariate analysis and machine learning techniques have enabled the observation of this channel with important statistical significance [17]. The production mode can be symbolically represented as $qq \to t\bar{t}H$. This process also includes the single-top associated production (tH)

1.8 Higgs Boson Decay Modes

The Higgs boson is an unstable particle with a very short lifetime, which means its presence is detected through its decay products. Like other unstable particles, the Higgs boson can be described by a Breit-Wigner resonance.

At tree level, all decay rates of the Higgs boson are determined by its couplings to Standard Model (SM) particles, which are constrained by unitarity. The couplings of the Higgs field to the SM particles are proportional to the masses (see Fig. 2). Therefore, the Higgs boson predominantly decays into the heaviest state allowed by the phase space. The branching ratios for different decay modes as predicted by the SM are shown in Figure 5.

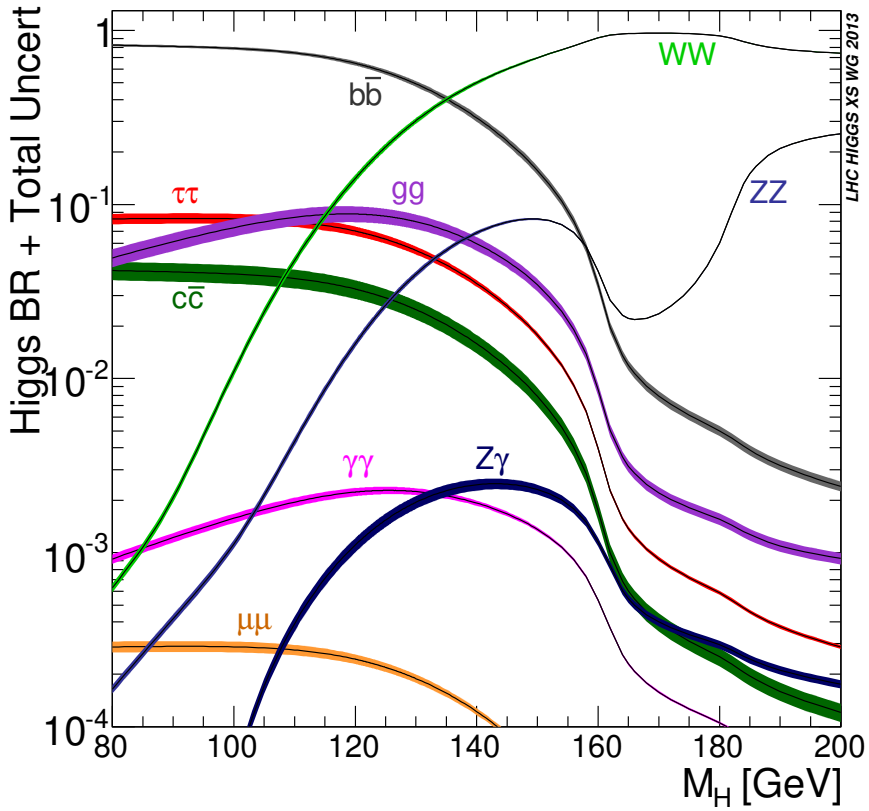


Figure 5: Higgs boson branching ratios with uncertainty widths as a function of the Higgs boson mas. The physical branching ratios correspond to a mass of 125 Gev[18].

1.8.1 Decay into Heavy Fermion Pairs

At leading order in SM couplings, the Higgs boson can decay into pairs of heavy fermions through Yukawa interactions, as shown in Figure 6(a). The partial widths of these decays are proportional to the square of the ratios between the fermion mass and the Higgs boson vacuum expectation value (Eq.65). For Higgs boson masses below the top

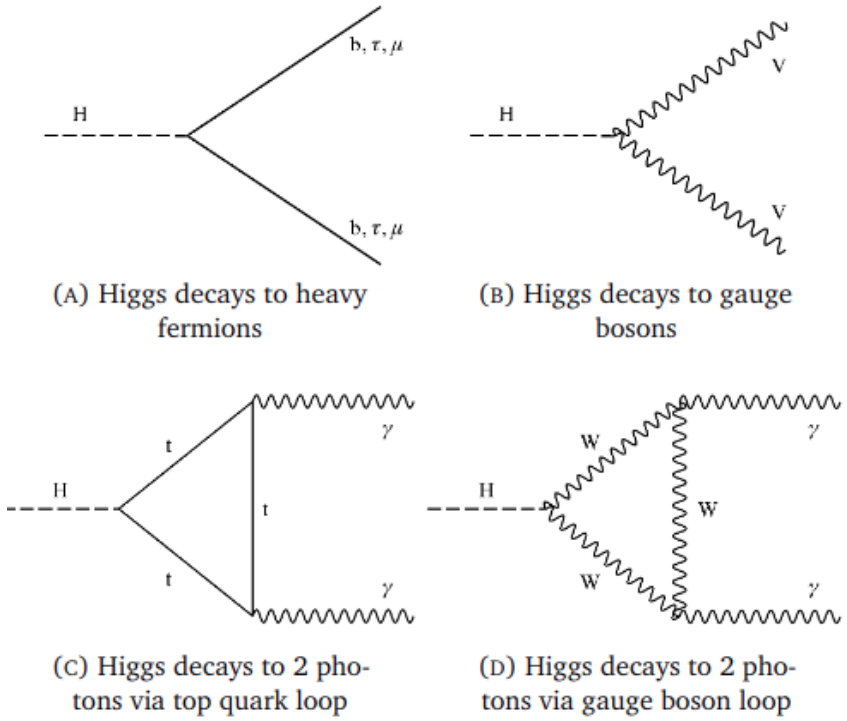
quark threshold, the dominant decay is into $b\bar{b}$ pairs, with smaller contributions from $c\bar{c}$ pairs, $\tau^+\tau^-$ pairs, and negligible contributions from other fermion pairs.

1.8.2 Decay into Gauge Boson Pairs

The Higgs boson can also decay into pairs of W and Z bosons through $SU(2)_L$ interactions, as depicted in Figure 6(b). This decay becomes dominant for Higgs boson masses above twice the mass of the W or Z bosons.

1.8.3 Loop-Induced Decays

Loop-induced decays of the Higgs boson into pairs of photons and gluons are also important. The decay into gluons is a mirror process of the gluon fusion production, occurring through a quark loop.The large Yukawa coupling of the top quark and the eight-fold color multiplicity of the final state allow this loop decay to be competitive to the tree-level decay into $\tau^+\tau^-$ pairs. Higgs boson decays into diphotons can happen both through a fermion loop (Fig.6(c)), dominated by the top quark contribution, and through a W boson loop with two WWγ vertices (Fig.6(d)). The partial width of the $H \to \gamma\gamma$ decay is about a factor 40 smaller than $H \to gg$, due to smaller electroweak couplings. Nevertheless, this channel has several unique advantages, playing a major role both in the discovery of the Higgs boson and in the study of its properties.



(A) Higgs decays to heavy fermions

(B) Higgs decays to gauge bosons

(C) Higgs decays to 2 photons via top quark loop

(D) Higgs decays to 2 photons via gauge boson loop

Figure 6: Leading order Feynman diagrams for Higgs boson decays[19].

1.8.4 Higgs to charm coupling

Both the ATLAS and CMS experiments have observed interactions between the Higgs boson and the third-generation charged fermions, with CMS also providing evidence of its decay into a pair of muons, while ATLAS reported a 2σ excess compared to background predictions. Despite extensive efforts by both collaborations, direct searches for Higgs boson decays into a charm quark-antiquark pair, $H \to c\bar{c}$, decays into an electron-positron pair, exclusive decays into mesons, and reinterpretations of the Higgs p_T spectrum have not yielded experimental evidence for Higgs boson couplings to first-generation fermions or second-generation quarks. These findings align with the SM's prediction that the coupling strength of the Higgs boson to each fermion scales proportionally to the fermion's mass. The SM predicts a branching fraction of 2.89% for $H \to c\bar{c}$, approximately 20 times smaller than the branching fraction for the Higgs boson to decay into a bottom quark-antiquark pair, $H \to b\bar{b}$. However, physics beyond the Standard Model could either enhance or diminish the coupling of the Higgs boson to the charm quark, consequently affecting the $H \to c\bar{c}$ branching fraction. This thesis aims to help a general search of this coupling with the production of the Higgs boson in association with a top quark–antiquark pair ($t\bar{t}H$).

1.9 The top quark

The initial detection of the top quark occurred in 1995 at the Tevatron by both the CDF and DØ collaborations [21]. Its discovery was delayed compared to other quarks due to its substantial mass, approximately 172 GeV, which earlier colliders lacked the energy to produce. Top quarks are generated at hadron colliders through two primary mechanisms: strong interactions (pair production) or electroweak interactions (single top production). Despite its significantly larger mass compared to other quarks, the existence of the top quark had been anticipated since the discovery of the bottom quark in 1977[22]. The identification of the bottom quark affirmed the existence of a third quark generation, as proposed by Kobayashi and Maskawa in 1973 to account for CP violation in the Standard Model [23]. The top and bottom quarks constitute the third and heaviest generation of quarks. The top quark, a 1/2 spin particle, possesses an electric charge of +2/3e and exhibits weak isospin characteristics when paired with the bottom quark.

Top quarks are produced through both strong and weak interactions. In hadron collisions, the dominant production mechanism is via strong interactions, with the cross section for top quark pair production being expressed as a sum over partons and integrated over momentum fractions. The top quark then decays predominantly into a W boson and a b-quark, with the decay width determined by various parameters including the Fermi coupling constant, the masses of the top quark and W boson, and the CKM matrix element. The different decay modes of the top quark pair include all-hadronic, semi-leptonic, and dileptonic channels, each characterized by the decay modes of the 2 W bosons and exhibiting distinct branching fractions.

The top quark's large mass confers unique properties, including its near-unity Yukawa coupling constant to the Standard Model Higgs boson and its relatively short lifetime compared to its hadronization time. Consequently, the top quark cannot form bound states, allowing for the study of its bare properties. Measurements of top quark polarization and spin correlations are achieved by analyzing angular distributions of various decay products.

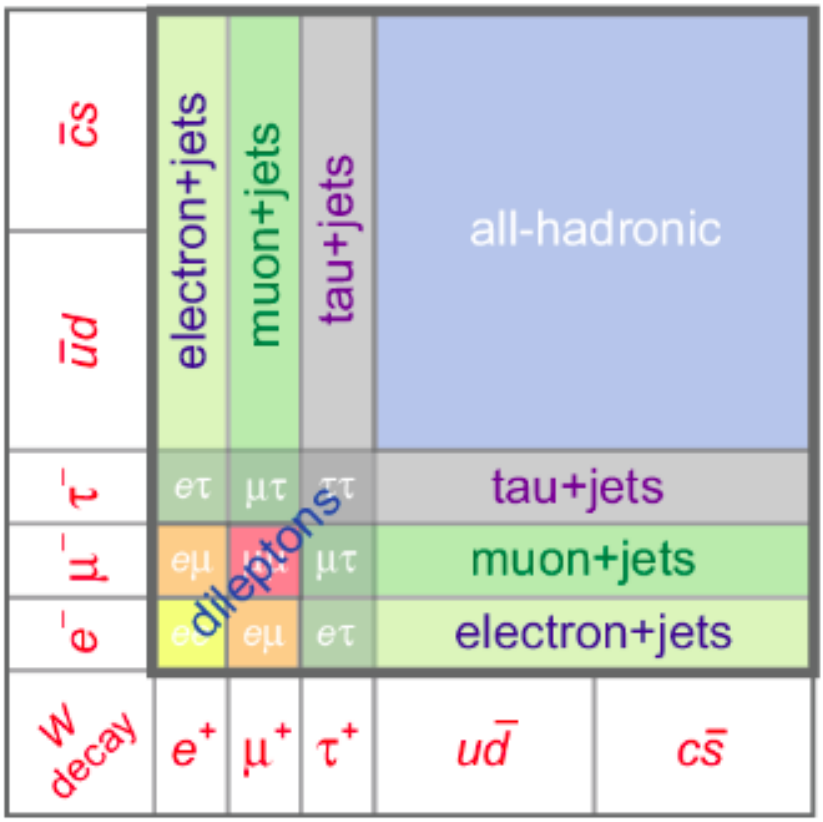


Figure 7: Decay modes of top quark pairs [24].

The final states for the leading pair-production process can be divided into three classes:

A. $t\bar{t} \to W^+ b\, W^- \bar{b} \to q q' b\, q'' q''' \bar{b}, \quad (45.7\%)$

B. $t\bar{t} \to W^+ b\, W^- \bar{b} \to q q' b\, \ell^- \nu_\ell \bar{b} + \ell^+ \nu_\ell b\, q'' q''' \bar{b}, \quad (43.8\%)$

C. $t\bar{t} \to W^+ b\, W^- \bar{b} \to \ell^+ \nu_\ell b\, \ell'^- \nu_{\ell'} \bar{b}. \quad (10.5\%)$

where an example of $t\bar{t}$ production and decay in the semi-leptonic channel can be shown in Figure 8

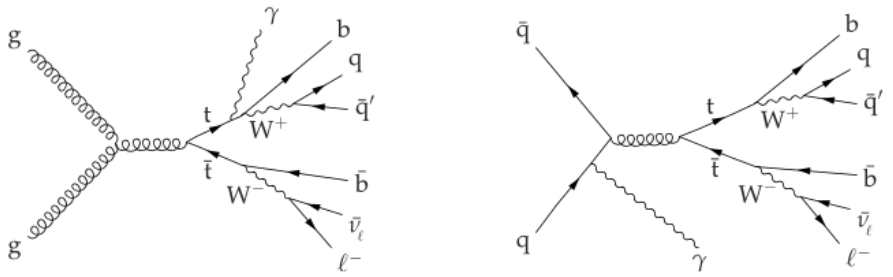


Figure 8: Examples of $t\bar{t}$ production and their following decays in the semi-leptonic channel. [25].

2 The CMS Experiment

2.1 The LHC

The LHC is a particle collider situated on the border between Switzerland and France at the European Organization for Nuclear Research (CERN). The LHC comprises a series of machines that progressively accelerate proton particles to higher energies for collisions. Figure 9 provides a schematic overview of CERN's accelerator complex.

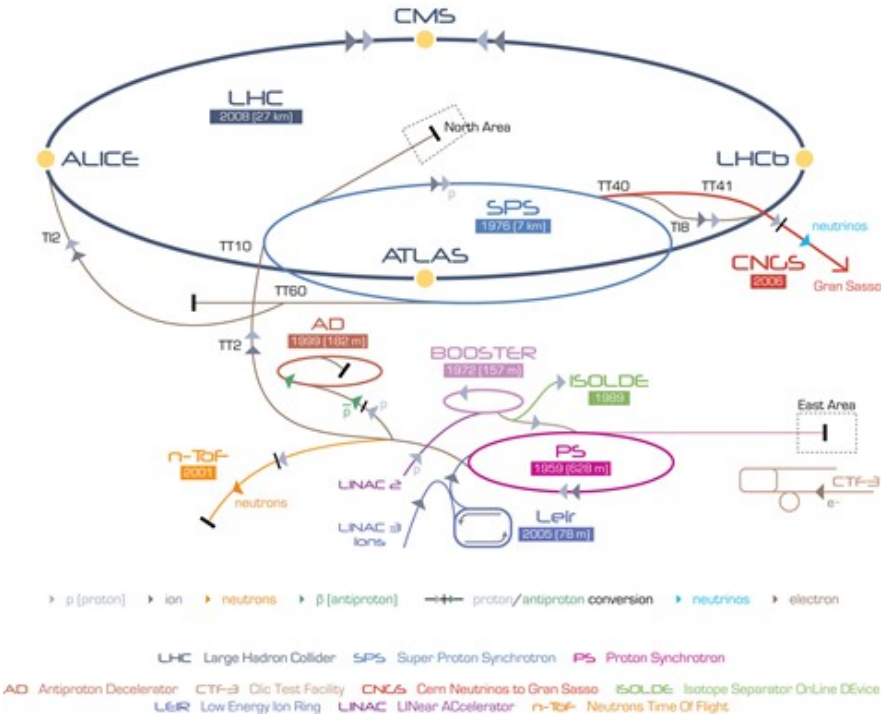


Figure 9: The CERN accelerator complex serving 4 major experiments [27]

The LHC hosts four major experiments positioned at different points along its ring where protons collide:

- A Toroidal Large Hadron Collider Apparatus (ATLAS) [28]
- Compact Muon Solenoid (CMS) [29]
- Large Hadron Collider beauty (LHCb) [30]
- A Large Ion Collider Experiment (ALICE) [31]

ATLAS and CMS are general-purpose detectors designed to explore a broad spectrum of physics phenomena independently. LHCb focuses on heavy flavor physics. ALICE is specifically designed to investigate heavy ion collisions and the production of quark-gluon plasma.

Two critical factors for the LHC's operation are its center-of-mass energy and luminosity, which monitors the collision rate. The instantaneous luminosity (L_{inst}) is a key parameter representing the collision rate at the LHC. It is calculated using the formula:

$$L_{\text{inst}} = f \eta_1 \eta_2 \frac{N^2}{4\pi\sigma_x\sigma_y} \tag{66}$$

where f is the collision frequency, η_1 and η_2 are the numbers of particles in the colliding bunches, and σ_x and σ_y are the transverse beam sizes.

The integrated luminosity (L) is defined as:

$$L = \int L_{\text{inst}}\, dt \tag{67}$$

For a specific process, the number of observed events (N_{obs}) is related to the integrated luminosity (L) by:

$$N_{\text{obs}} = \sigma \times \epsilon \times L \times A \tag{68}$$

where σ is the cross section of the process and ϵ is the detection efficiency optimized by experimentalists ans A is the detector acceptance.

The Large Hadron Collider (LHC) has operated at three different proton-proton (pp) center-of-mass energies: 7 TeV, 8 TeV, and 13 TeV, and also conducted various heavy ion collision scenarios involving lead or xenon nuclei. Operations at 7 TeV and 8 TeV occurred from 2010 to 2012, referred to as Run-I, while the 13 TeV operation extended from 2015 to 2018, known as Run-II. Run-III has begun in 2021 at a design luminosity of 13.6 TeV. Subsequently, a significant upgrade named the High Luminosity (HL) upgrade is planned to achieve a targeted integrated luminosity of 3000 fb^{-1}, representing an almost tenfold increase in instantaneous luminosity compared to Run-III. The LHC program timeline from 2011 to 2040 is depicted in Figure 10.

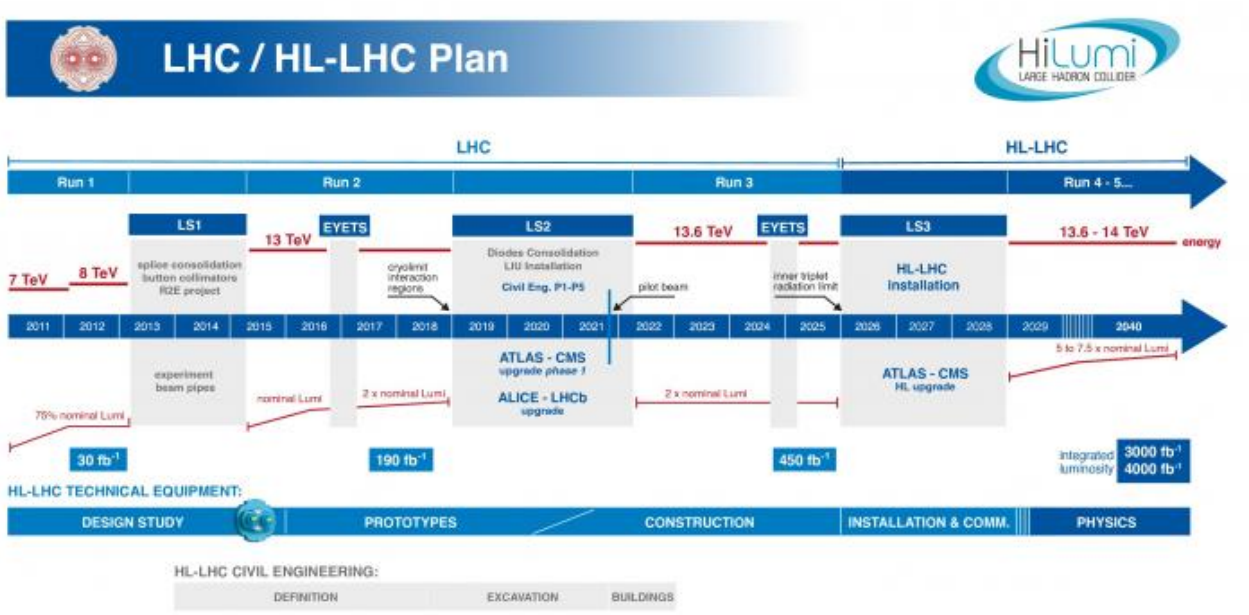


Figure 10: The roadmap sketch for LHC programme up to the high-luminosity LHC, orcalled HL-LHC[32]

At the current design luminosity, the collisions result in approximately $< \mu > \approx 25$ interactions per bunch crossing. Additional interactions from other protons in the same or nearby bunches, known as pile-up interactions, accompany the event of interest triggered by a hard scattering process. There are two types of pile-up: in-time pile-up occurs during the same bunch-crossing as the collision of interest, while out-of-time pile-up occurs in preceding or subsequent bunch crossings.

As instantageous luminosity increases, the probability of multiple proton-proton interactions within a single bunch crossing rises. Therefore, mitigating the pile-up effect involves improving the identification and reconstruction of a single primary collision where the physics event of interest occurs. By the end of Run-II, the CMS detector had recorded a dataset corresponding to an integrated luminosity of 140 fb^{-1} at Run II. The cumulative luminosity delivered by the LHC and recorded by CMS during the Run-II data-taking periods is depicted in Figure 11.

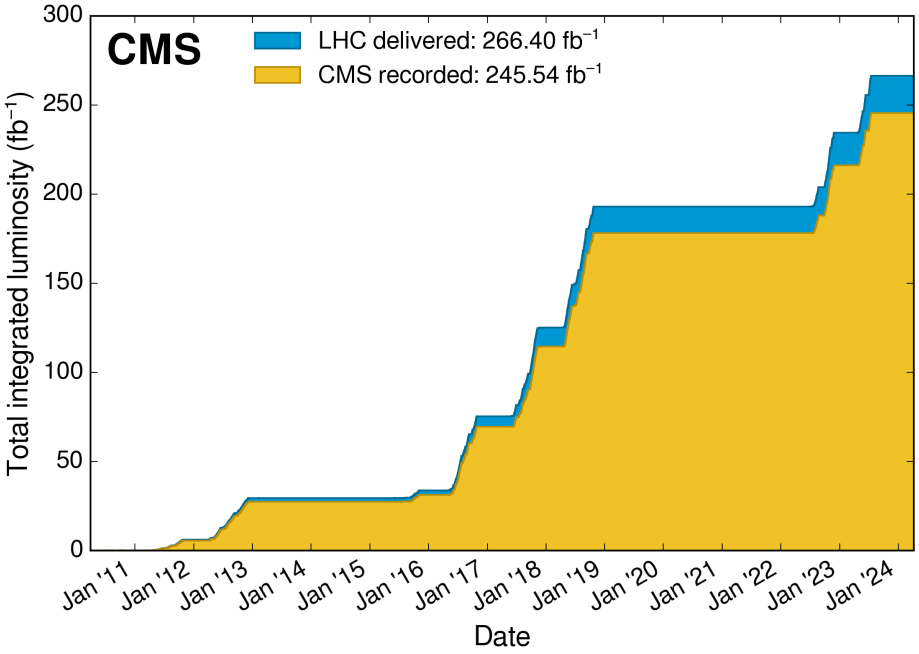


Figure 11: Cumulative delivered and recorded luminosity versus time for 2010-2012, 2015-2018, and 2022-2023 (pp data only). [33]

2.2 The CMS Detector

Our focus lies on the CMS (Compact Muon Solenoid) experiment, designed for general purposes, including refining measurements within the standard model and probing potential indicators of new physics beyond its constraints. The upcoming sections serve to introduce the detector's structure, its constituent layers, the employed coordinate system in our experiments, and subsequently, the data format at our disposal. Figure 12 illustrates the detector's configuration. Progressing from innermost to outermost layers, we encounter the tracker, responsible for tracking the trajectories of charged particles, followed by the Electromagnetic Calorimeter (ECAL), the Hadronic Calorimeter (HCAL), Solenoid Magnets, and ultimately, the Muon Chambers, which lend their name to our detector. [34]

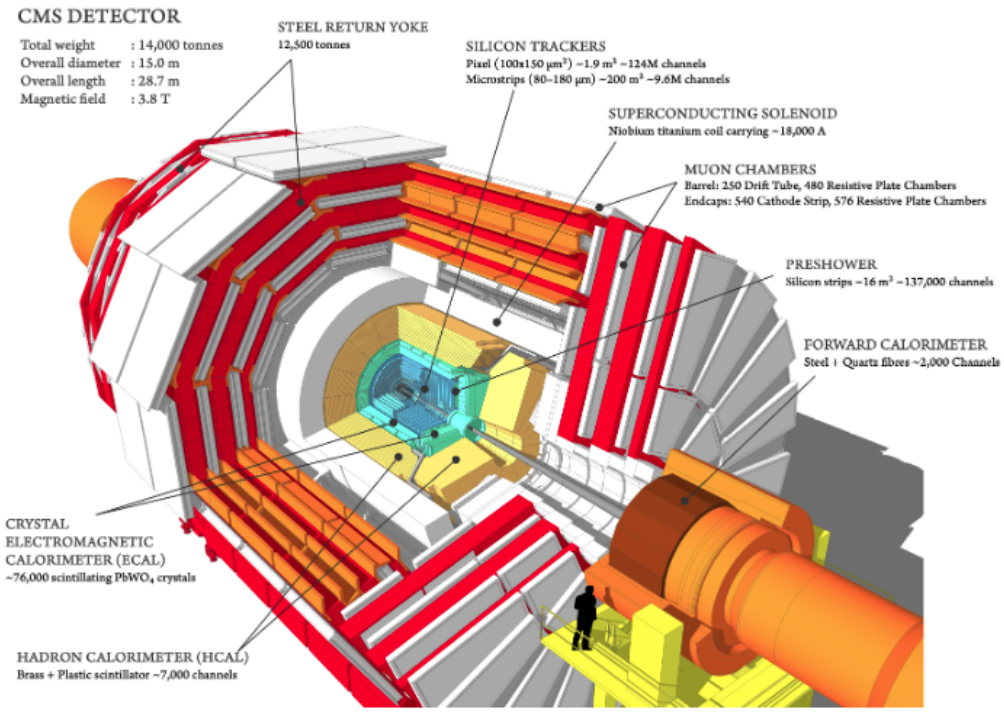


Figure 12: Schematic View of the CMS Detector [35]

The next section will delve into the coordinate system employed within the CMS experiment and provide insights into the layout of the CMS detector.

2.3 Coordinate System & Kinematics

The CMS experiment utilizes a specific coordinate system, depicted in Figure 13, with its origin situated at the collision point within the detector. The z-axis is oriented along the axis of pp collisions, directed towards the Jura mountains. The x-axis points towards the center of the LHC machine, and the y-axis extends upward, establishing a right-handed coordinate system.

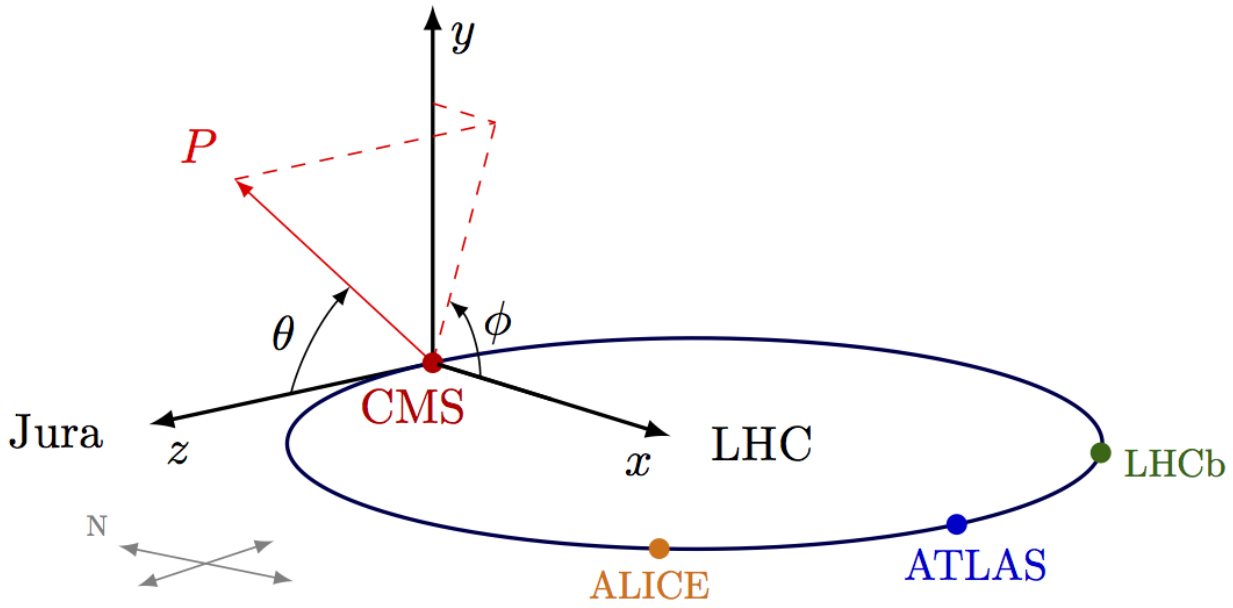


Figure 13: The coordinate system of CMS. [36]

In the framework of special relativity, the four-vector characterizing a particle produced in a proton-proton collision and measured by the detector is denoted as follows:

$$\mathcal{P}^{\mu} = (E, p_x, p_y, p_z) \tag{69}$$

Here, E represents the particle's energy, and p_i denotes the momenta along each axis. Throughout our equations, natural units are employed, signifying that the speed of light (c) and Planck's constant ($\hbar$) are both set to 1.

While the transverse components p_x and p_y remain Lorentz invariant under a boost along the collision axis, the other two components, E and p_z, do not share this invariance. To address this, we introduce new kinematic and coordinate variables. Consequently, we opt for three invariant components: the azimuthian angle ϕ, measured from the x-axis in the x-y plane; the total transverse momentum p_T, lying in the x-y plane and formed by the vector addition of p_x and p_y; and the particle's mass m, which is inherently invariant. These quantities can be expressed as functions of the components of $\mathcal{P}^{\mu}$:

$$P_t = |\vec{p}| sin\theta \tag{70}$$

$$\phi = tan^{-1}(\frac{y}{x}) \tag{71}$$

$$m = \sqrt{E^2 - |\vec{p}|^2} \tag{72}$$

As a fourth component we introduce a new quantity , the rapidity. Rapidity is defined as :

$$y = \frac{1}{2}\ln\left(\frac{E+p_z}{E-p_z}\right) \tag{73}$$

This quantity is not Lorenz Invariant , so if we consider a boost of velocity β along the beam axis (z-axis) we get the following :

1) The transformation of E $\rightarrow E^{'}$

$$E^{'} = \gamma(E+\beta p_z) \tag{74}$$

2) The invariance of the transverse momentum components :

$$p_x^{'} = p_x \tag{75}$$

$$p_y^{'} = p_y \tag{76}$$

3) The transformation of the z-component of momentum :

$$p_z^{'} = \gamma(p_z+\beta E) \tag{77}$$

Given the above we can write the boosted version of 73 as:

$$y' = \frac{1}{2}\ln\left(\frac{E'+p_z'}{E'-p_z'}\right) = \frac{1}{2}\ln\left(\frac{(E+p_z)(1+\beta)}{(E-p_z)(1-\beta)}\right) = y + \ln(\gamma(1+\beta)) \tag{78}$$

Where the term $ln[\gamma(1+\beta)]$ is a constant. So by making a boost along the z-axis we see that the rapidity is not Lorenz invariant. To further simplify the definition 73 we can write $p_z = |\vec{p}|cos\theta$ and also $E = \frac{|\vec{P}|}{\beta}$. So we can rewrite eq. 73 as :

$$y = \frac{1}{2}\ln\left(\frac{1+\beta\cos\theta}{1-\beta\cos\theta}\right) \tag{79}$$

For a massless particle we can also define the pseudo-rapidity from the above equation. Namely we set $\beta = 1$, so $\eta = y(\theta, \beta = 1)$. As a result we can calculate the pseudo rapidity as:

$$\eta = \frac{1}{2}\ln\left(\frac{1+\cos\theta}{1-\cos\theta}\right) = \ln\left(\frac{\cos\frac{\theta}{2}}{\sin\frac{\theta}{2}}\right) = -\ln\tan\frac{\theta}{2} \tag{80}$$

Considering the energies and transverse momentum ranges typical at CERN (where $E, p_T >> 1$, implying $\beta \rightarrow 1$), we find that $y \approx \eta$ holds true for stable particles. Thus, we can reliably substitute rapidity with pseudo-rapidity.

When examining pseudo-significance values, it becomes apparent that at $\eta = 0$, we reference the transverse plane, while the limit $\eta \rightarrow \infty$ corresponds to the collision axis. Given the practical constraints imposed by the detector's geometry, particularly at CMS, our exploration is typically confined to $|\eta| < 5$. Consequently, based on these considerations, we can express equation 69 as:

$$\mathcal{P}^{\mu} = (p_t, \eta, \phi, m) \tag{81}$$

2.4 The Structure of the CMS Detector

From the innermost to the outermost regions, the CMS (Compact Muon Solenoid) detector is composed of several sub-detectors:

× **Silicon Tracker**: Consists of an inner silicon pixel vertex detector and an array of silicon microstrip detectors covering a total active area of about 215 m^2. Situated in the region $|\eta| < 2.5$ and radial distance $r < 1.2$ m, it reconstructs charged particle tracks and vertices.

× **Electromagnetic Calorimeter (ECAL)**: Constructed with scintillating lead tungstate ($PbWO_4$) crystals, located in the region $|\eta| < 3$ and 1.2 m $< r <$ 1.8 m. The ECAL measures the trajectory and energy of passing electrons and photons.

× **Hadronic Calorimeter (HCAL)**: Made up of brass layers alternated with plastic scintillators, situated in the region $|\eta| < 5$ and 1.8 m $< r <$ 2.9 m. It measures the direction and energy deposited by passing hadrons.

× **Superconducting Solenoid Magnet**: Generates a uniform magnetic field in its internal region necessary for bending the trajectories of charged particles. Housed within a 21 m long iron yoke with a diameter of 14 m, it is located in the region $|\eta| < 1.5$ and 2.9 m $< r <$ 3.8 m. The magnet enables measurement of charged particle momenta through curvature observed in the tracking system.

× **Muon System**: The outermost detection system, accommodated in the region $|\eta| <$ 2.4 and 4 m $< r <$ 7.4 m. It is comprising Drift Tube Chambers (DT) in the barrel region and Cathode Strip Chambers (CSC) in the endcaps, with Resistive Plate Chambers (RPC) for triggering purposes. It is used to reconstruct tracks of traversing muons.

Proton-proton collisions occur in the beam pipe region, immediately surrounded by the tracking system. Charged particles produced in collisions traverse outward, leaving signals or hits in the silicon tracker. Hits are used to trace particle trajectories and reconstruct tracks, back-tracking to the primary interaction point.

Electrons and photons are absorbed by the ECAL, where their electromagnetic showers are detected for energy and direction determination. Hadrons may also initiate showers in the ECAL, fully absorbed in the HCAL, providing energy and direction estimates. Muons and neutrinos largely pass through the calorimeters with minimal interaction. Muons produce hits in the muon detectors, located outside the calorimeters, aiding in muon track reconstruction.

As, it can be seen from Figure 12, all subdetectors are divided into a "barrel" compartment placed parallel to the beam axis and two "endcap" compartmentd placed vertical to the beam axis, thus making CMS nearly hermetic (i.e. "π" detector)

2.4.1 The Magnet

The "S" in the CMS abbreviation refers to the Solenoid, which is the central device around which the entire experiment is built. The "M" stands for Muons, and the "C" represents Compact, reflecting how tightly integrated the sub-detectors are within the detector structure. Vertical slices of the detector show minimal empty space between the sub-detectors, with only a thin layer separating the solenoid [38] and the muon chambers. The solenoid is positioned immediately after the hadronic calorimeter.

The solenoid's coils are composed of superconducting niobium-titanium, capable of generating a powerful magnetic field of 3.8 T with a current of 13,000 A. This magnetic field is approximately 100,000 times stronger than the Earth's magnetic field. The solenoid measures 13 meters in length and 6 meters in diameter. Initially designed for a 4 T field, the operating amplitude was adjusted to 3.8 T to ensure longevity and prevent damage.

The inductance of the magnet is 14 H, and the nominal current for 4 T is 19,500 A, giving a total stored energy of 2.66 GJ, equivalent to about half a tonne of TNT. Dump circuits are in place to safely dissipate this energy should the magnet quench.

The primary function of the large magnet is to bend the paths of all charged particles resulting from high-energy collisions in the LHC. By tracking a particle's path, its momentum can be measured, as higher-momentum particles exhibit less curvature in the magnetic field. The combination of the magnetic field with high-precision position measurements from the tracking system and muon chambers enables accurate momentum measurement, even for high-energy particles.

The muon detectors are interleaved within a 12-sided iron structure that not only surrounds the magnet coils but also guides the magnetic field. Comprising three layers, this "return yoke" spans 14 meters in diameter and acts as a filter, allowing through only muons and weakly interacting particles like neutrinos. The enormous magnet also provides much of the experiment's structural support, requiring significant strength to withstand its own magnetic forces.

The iron-made structural supports and return yokes contribute significantly to the total weight of the CMS detector, which amounts to 15,000 tonnes. The return yokes help maintain a nearly homogeneous magnetic field throughout the detector volume outside the Solenoid.

The yoke is composed of common steel and forms five three-layered dodecagonal barrel wheels and three endcap disks at each end. In the barrel region, the innermost yoke layer is 295 mm thick, and each of the two outermost ones is 630 mm thick. The yoke contributes to only 8% of the central magnetic flux density. Its main role is to increase the field homogeneity in the tracker volume and to reduce the stray field by returning the magnetic flux of the solenoid. Additionally, the steel plates serve as absorbers for the four interleaved layers ("stations") of muon chambers, which provide for a measurement of the muon momentum independent of the inner tracking system.

The mapped magnetic flux density on a longitudinal section of the CMS detector is shown in Figure 14. Approximately two thirds of the magnetic flux return through the barrel yoke,half of which enters directly into the barrel without passing through the endcap disks. One third of the total flux escapes radially, returning outside the steel yoke.

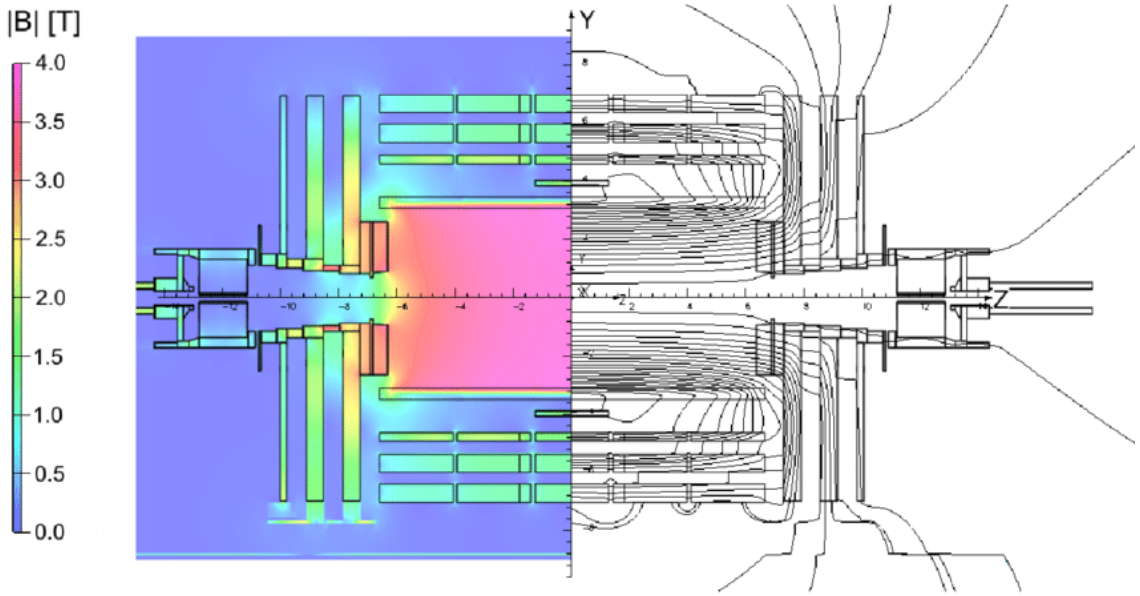


Figure 14: Value of magnetic field (left) and field lines (right) predicted on a longitudinal section of the CMS detector.[39]

The Magnetic Field along the beam axis is parametrized as:

$$B_z(0,z) = \frac{1}{2} B_0 \sqrt{1+\bar{a}}[f(u)+f(v)] \tag{82}$$

where

$$u = \frac{1-\bar{z}}{\bar{a}}$$
$$v = \frac{1+\bar{z}}{\bar{a}}$$
$$f(x) = \frac{x}{\sqrt{1+x^2}}$$
$$\bar{z} = \frac{2z}{L}$$
$$\bar{a} = \frac{2a}{L}$$

Inside the solenoid region (orange-pink region), the two components of the magnetic field (along the z axis and the radial one respectively) are parametrized as:

$$B_z(r,z) = \sum_{\nu=0}^{\infty} \frac{(-1)^\nu}{(\nu!)^2} (\frac{r}{2})^{2\nu} \frac{\partial^{2\nu}}{\partial z^{2\nu}} B_z(0,z) \tag{83}$$

$$B_r(r,z) = \sum_{\nu=0}^{\infty} (-1)^\nu \frac{1}{(\nu-1)!\nu!} \frac{\partial^{2\nu-1}}{\partial z^{2\nu-1}} B_z(0,z) (\frac{r}{2})^{2\nu-1} \tag{84}$$

2.4.2 Inner Tracking System

The momentum of particles is determined by tracking their paths through a 4T magnetic field, which causes curved trajectories inversely related to particle momentum. The tracker records particle paths with precision, using minimal interference, achieved through accurate position measurements. The tracker, located closest to the collision point, is radiation-resistant. The CMS Tracker is entirely silicon-based. It operates by detecting ionization in doped semiconductor wafers, where drifting electron-hole pairs are collected by p-n junctions configured as strips or pixels. This technology allows reconstructing particle trajectories from "hits" left in silicon sensors. The pixel tracker, central to the detector, handles high particle intensity and is surrounded by silicon microstrip detectors. As particles pass through, pixels and microstrips generate electric signals that are amplified and detected. The tracker's main role is to record charged particle paths and measure their momentum based on trajectory curvature in the magnetic field. After reconstructing particle tracks, the tracker identifies primary (interaction points of protons) and secondary vertices (decay points of particles) based on associated track numbers. The CMS tracking system reconstructs tracks within the $|\eta| < 2.4$ range, corresponding to an angle of nearly 80 degrees from the xy-plane as shown in Figure 15. It consists of five parts: A) Pixel Detector, B) Tracker Inner Barrel (TIB), C) Tracker Inner Disks (TID), D) Tracker Outer Barrel (TOB), and E) Tracker End Caps (TEC), predominantly using silicon microstrips.

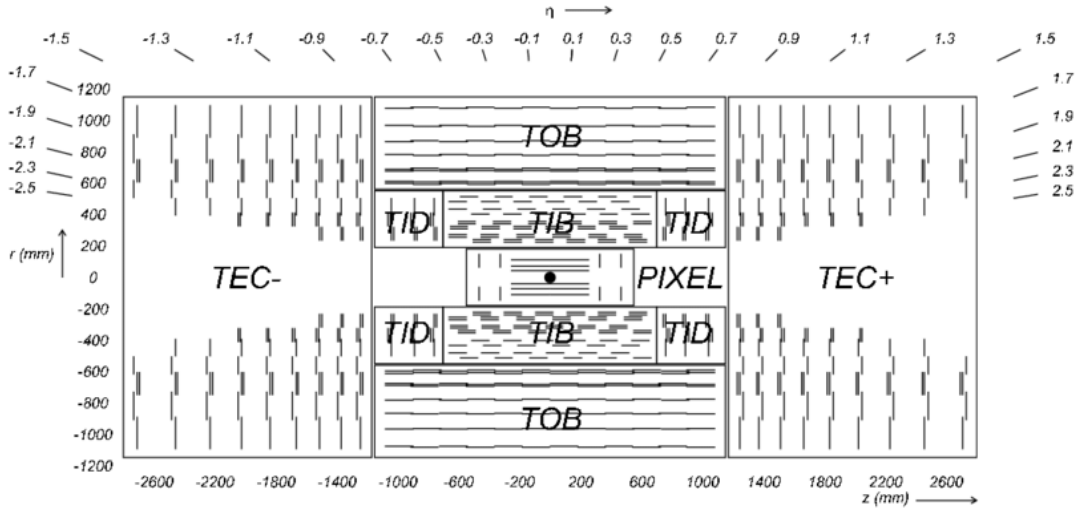


Figure 15: The CMS tracking system.[40]

The Pixel Detector [41] contains 124 million pixels and offers precise track and vertex reconstruction. It is situated closest to the beam pipe and has been upgraded twice since 2010. In 2018, it consisted of four cylindrical layers and six endcap disks, achieving a forward acceptance of $|\eta| < 2.5$.Charged particles passing through pixels generate electron-hole pairs. Each pixel collects charges as small electric signals, amplified by electronics using bump bonding techniques. By analyzing which pixels are activated, the 3D trajectory of particles can be reconstructed. To prevent overheating, the pixels are mounted on cooling tubes. The detector is structured into ladders (r-φ plane) and rings (z axis), forming modules with Readout Chips (ROCs). Each ROC reads out pixel data and is evaluated based on hit efficiency within a 500μm radius of a reconstructed trajectory (see Figure 16). Systematic uncertainty is estimated by comparing measurements with "ideal tracks," yielding an uncertainty of approximately 0.3%. For a 100 GeV muon, the Pixel Detector achieves a 1-2% transverse momentum resolution up to $|\eta| = 1.6$ and around 10μm transverse impact parameter resolution within $|\eta| < 2.5$.

The Strip Silicon Detectors [42] cover a total area of 223 m^2 over an axial range of 20-120 cm and contain 10 million detectors organized into 15200 modules read by 80,000 microelectronic chips. Each module comprises sensors, mechanical support, and readout electronics.The tracker and its electronics endure radiation but operate at -20°C to withstand it.The strip tracker records rough two-dimensional trajectory information across multiple surfaces to obtain a robust lever arm for better momentum measurement. In contrast, pixels record limited but accurate three-dimensional trajectory information near the interaction point. The strip tracker consists of 4 sub-modules creating a ten-layered

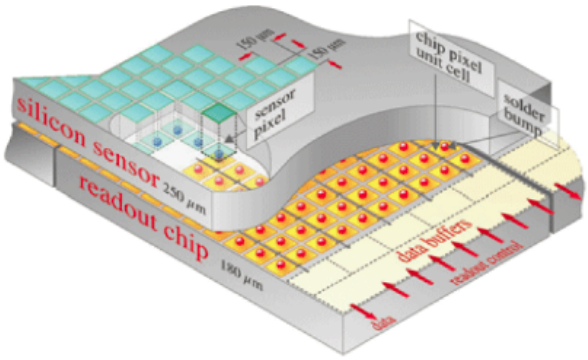


Figure 16: Sketch of a typical CMS pixel sensor[41]

silicon strip detector, extending to a radius of 130 centimeters.Silicon sensors detect particles similarly to pixels: as charged particles cross the material, they create electron-hole pairs generating a small current pulse. APV25 chips amplify these charges to reconstruct particle paths (see Figure 17) . The charge on each microstrip is read out and amplified by APV25 chips housed within hybrids containing electronics for monitoring sensor information and timing hits with collisions. The processed signals are converted into infrared pulses and transmitted over a 100-meter fiber optic cable for radiation-free analysis. The tracker uses 40,000 such links for signal transmission.Strip detectors are made of n-type silicon with heavily doped p+ implants on one side to provide one-dimensional information about traversing particles.The CMS sensor design uses AC coupling to bypass DC leakage current over polyresistors and detect the AC part over capacitors.Each strip is connected to its readout channel and amplifier using a 25μm thin wire welded onto the corresponding AC pad via ultrasonic bonding wedges.

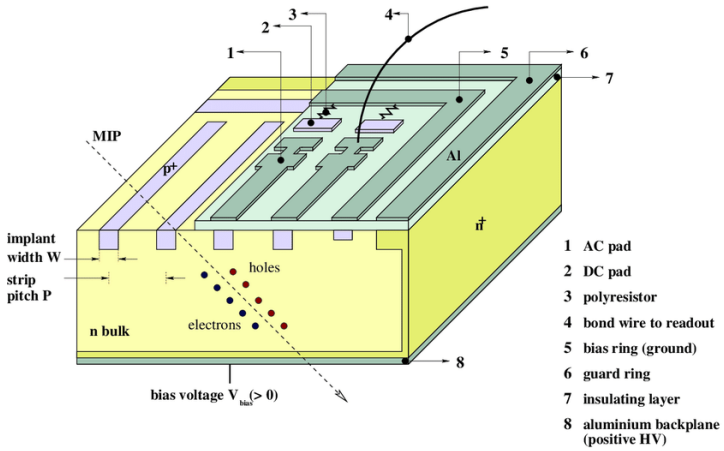


Figure 17: A
CMS silicon microstrip sensor.[43]

2.4.3 Electromagnetic Calorimeter (ECAL)

The electromagnetic calorimeter (ECAL) used in the CMS detector is designed to measure the energy of electrons and photons using 75,848 lead tungstate ($PbWO_4$) crystals [44] (see Figure 18). This calorimeter, with dimensions of 7.9 m in length and 3.6 m in diameter, completely surrounds the tracker detector and weighs approximately 90 tons. The primary purpose of an electromagnetic calorimeter is to measure the energy of particles that interact mainly via electromagnetic processes, such as electrons and photons. Electrons emit Bremsstrahlung photons, while photons undergo interactions like the photoelectric effect, Compton scattering, and pair production within the active material of the calorimeter. When a high-energy particle interacts within the calorimeter, it produces a cascade of lower-energy particles through successive interactions until the energy is sufficiently depleted and absorbed by the material. The Molière radius characterizes the transverse size of these electromagnetic showers, and the radiation length (X_0) quantifies the material's ability to interact with incoming particles. [45]

Lead tungstate ($PbWO_4$) crystals are chosen for the ECAL due to their high density (8.28 g/cm^3), short radiation length (0.89 cm), and small Molière radius (2.2 cm). These crystals respond rapidly, with about 80% of scintillation light emitted within 25 nanoseconds, matching the timing of LHC bunch crossings. They are also radiation-resistant, maintaining performance over the course of LHC operations.

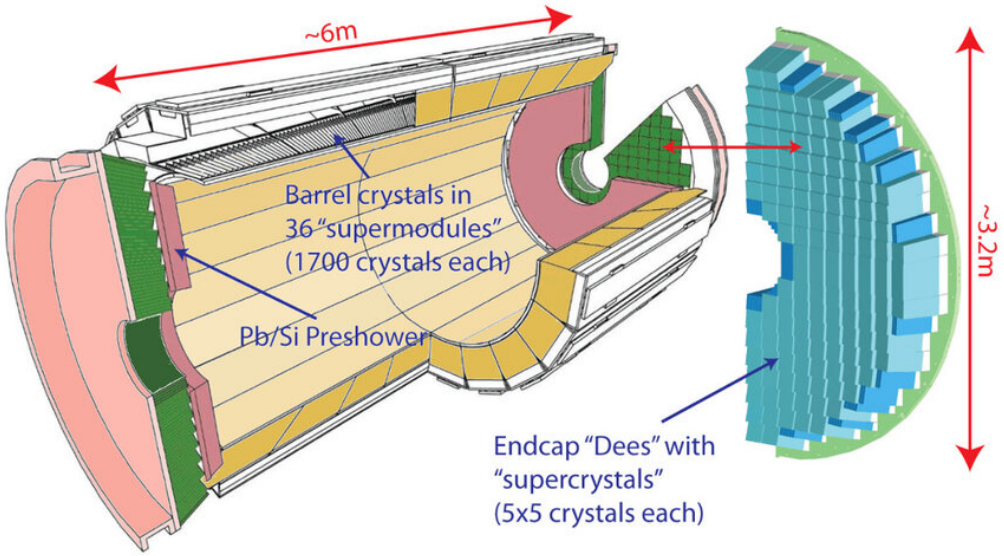


Figure 18: A schematic view of the CMS Electromagnetic calorimeters. [46]

In the barrel region, the $PbWO_4$ crystals are approximately 2.2×2.2×23 cm^3 in size.The electromagnetic calorimeter (ECAL) in the CMS detector comprises 75,848 lead tungstate ($PbWO_4$) crystals organized into 36 supermodules, each covering a 20-degree segment

in ϕ. The crystal length corresponds to approximately 26 radiation lengths (X_0). In the endcap electromagnetic (EE) calorimeter, the crystals are about $2.2 \times 2.2 \times 22$ cm^3 in size, corresponding to approximately 25 X_0, grouped into 5x5 supercrystal units arranged in a regular grid.All crystals are oriented towards the beamspot and are designed to operate in a 4 Tesla magnetic field. The barrel electromagnetic (BE) calorimeter covers a pseudorapidity range of $|\eta| < 1.479$, while the EE covers $1.479 < |\eta| < 3.0$.The preshower detector, located in front of the EE calorimeters, consists of lead layers and silicon strip sensors. It is used for identifying neutral pions and improving electron identification against minimum ionizing particles.

The energy resolution measured in the ECAL barrel (EB) [47] is characterized by:

$$\left(\frac{\sigma}{E}\right)^2 = \left(\frac{2.8\%}{\sqrt{E}}\right)^2 + \left(\frac{0.12\%}{E}\right)^2 + \left(0.3\%\right)^2 \tag{85}$$

where E is measured in GeV. The terms represent the stochastic term, noise term, and constant term, respectively. The homogeneous nature of the ECAL crystals minimizes scintillation photon loss in the absorber, reducing the stochastic term.

Despite emitting a relatively low light yield (~ 50 photons/MeV) that varies with temperature ($-2\%/°C$ at 18°C), overheating is mitigated by a water cooling system. To monitor crystal performance and prevent issues like low transparency or self-annealing due to irradiation, a precise monitoring system is employed.The crystals are read out by two types of photodetectors: silicon-based avalanche photodiodes (APD) in the barrel and vacuum phototriodes in the endcaps, both operated at high voltages. Collected data are transferred to off-detector electronics for further processing. Light loss due to irradiation during CMS operation and subsequent recovery is assessed using a laser light-injection system.

In the barrel region of the CMS ECAL, avalanche photodiodes (APDs) with 75% quantum efficiency and an excess noise factor of 2.1 at an operating gain of 50 are used. These APDs are insensitive to shower leakage particles passing through them. In the Endcaps, where radiation levels are higher and the magnetic field direction is within 25 degrees of the crystal axes, vacuum phototriodes (VPTs) are employed. VPTs are photomultipliers with a single gain stage developed specifically for CMS, featuring 20% quantum efficiency and a radiation-hard UV glass window. They typically operate at a gain of 10 in a 4 Tesla magnetic field.The data from 5x5 crystal arrays and their corresponding photodetectors are processed by multiple amplifiers, shaped, and digitized by trigger tower electronics at a 40 MHz rate. Subsequently, this information is sent to the level 1 trigger, and triggered data are forwarded to the off-detector electronics. The noise per channel, measured on completed supermodules, is 40 MeV with an rms spread of 3 MeV. The off-detector elec-

tronics of the ECAL handle both the Data Acquisition (DAQ) and Trigger paths. The DAQ path manages readout and data reduction, while in the Trigger path, Trigger primitives received from on-detector electronics are synchronized by Trigger Concentrator Cards and transmitted to the Regional Calorimeter Trigger. To monitor changes in crystal light yield under irradiation, such as variations in scintillation light reaching the photodetectors, blue (440 nm) and infrared (796 nm) wavelengths are utilized. Light pulses are distributed to the crystals via an optical fiber system. Additionally, the ECAL detector control system monitors parameters such as temperature, humidity, water cooling system status, water flow, and voltage supplies.

2.4.4 Hadron Calorimeter (HCAL)

The CMS hadron calorimeter (HCAL), together with the ECAL, forms a comprehensive calorimetry system for measuring jets and missing transverse energy (MET) [48]. The accurate measurement of jets and MET is essential for identifying various Standard Model processes such as QCD multi-jets, top, W+jets, Z+jets, and new physics signatures. Unlike electromagnetic cascades, the physical processes that initiate a hadron shower are distinct. Hadron production, nuclear de-excitation, and meson decays are primary contributors to these showers. It's estimated that about one-third of produced pions are neutral pions, which dissipate their energy in the form of electromagnetic showers. Hadronic showers often exhibit an electromagnetic component that can be slightly displaced from the hadronic part. Another notable characteristic of hadronic showers is their longer development time compared to electromagnetic showers. This is evident when comparing the particle content versus depth for pion and electron initiated showers. The longitudinal development of hadronic showers is governed by the interaction length, necessitating larger and longer hadronic calorimeters to encompass the entire cascade and minimize energy and particle losses from the rear layers of alternating active material and absorber plates. Consequently, the larger size of hadronic calorimeters contributes to challenges in achieving homogeneity.

The CMS HCAL (see Figure 19) comprises a brass/steel sampling hadron calorimeter, where the material generating the particle shower is distinct from the material that measures the deposited energy . It is designed to measure the energy of charged and neutral hadrons and is positioned outside the ECAL. The CMS HCAL consists of four sub-detectors to cover a wide pseudorapidity range up to $|\eta| = 5.2$, which includes the barrel and end-cap (HB, HE) covering $|\eta| < 1.3$ and $1.3 < |\eta| < 3$ respecritvely, the forward calorimeter (HF), and the outer calorimeter (HO).The HCAL barrel (HB) and end-cap (HE) detectors surround the electromagnetic calorimeter and are entirely within the

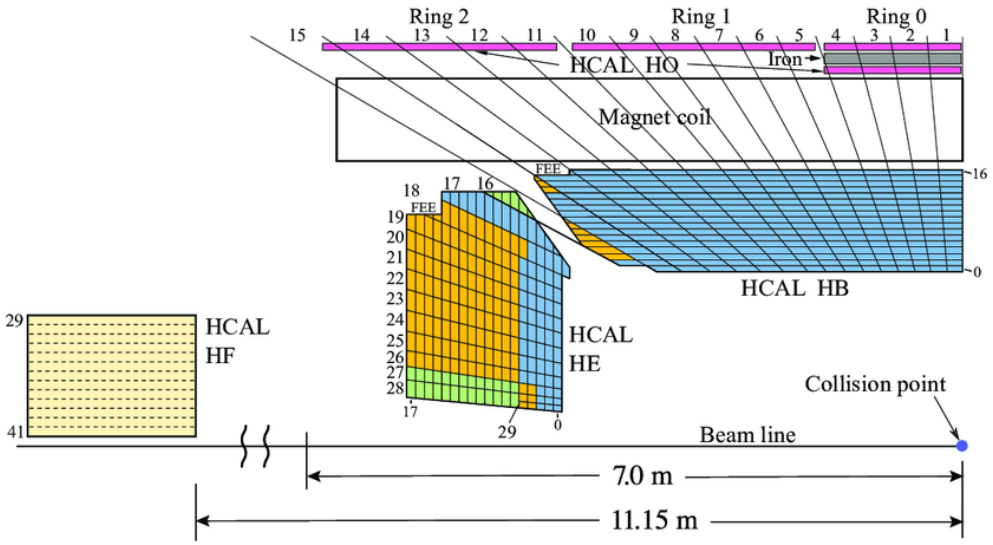


Figure 19: A quadrant of the CMS hadronic calorimeter.[50]

high magnetic field region of the solenoid [49]. The effective HCAL thickness in the $|\eta|$ < 1.3 region is extended by the addition of the outer barrel (HO) scintillators outside the magnet cryostat. Each subdetector covers the full range of the azimuthal angle φ.The HB and HE subdetectors consist of layers of plastic scintillator within a brass/stainless steel absorber, segmented into readout channels. In regions where $|\eta|$ is greater than 1.74, the φ segmentation is coarsely granulated. Scintillation light produced in the plastic scintillators is detected by hybrid photodiodes (HPDs).The interaction length λ, which is the mean distance traveled by a hadronic particle before undergoing an inelastic nuclear interaction, is equal to 5.82 at $\eta = 0$ and increases to 10.6λ at $|\eta| = 1.3$. The ECAL in front of HB contributes approximately 1.1λ. The scintillators are arranged in trays of tiles called megatiles, with a granularity of $\Delta\eta\Delta\phi = 0.087 \times 0.087$ for $|\eta| < 1.6$ and $\Delta\eta\Delta\phi = 0.17 \times 0.17$ for higher $|\eta| > 1.6$.The HO utilizes the solenoid coil as an additional absorber to act as a tail catcher for late showers, compensating for the smaller radiation lengths at low η.

During the several testing phases of the detector , using beams of pions , the combined resolution of the CMS ECAL and HCAL system is :

$$\left(\frac{\sigma}{E}\right)^2 = \left(\frac{84.7\%}{\sqrt{E}}\right)^2 + (7.4\%)^2 \tag{86}$$

The resolution is indeed significantly worse compared to Eq. 85 , making the energy measurement of the hadron showers less accurate than the electromagnetic ones.

2.4.5 Muon Detectors

The CMS collaboration emphasizes precise reconstruction and identification of high-energy muons as a key design goal. The primary objective is achieving approximately 1% resolution in dimuon mass for 100 GeV muons and determining muon charge up to 1 TeV. Muons play a vital role in various analyses due to their detection and identification potential, leading to processes like $H \rightarrow 4\mu$ and $H \rightarrow 2\mu$, and potentially revealing supersymmetric decay channels where muons are key final products. Muons are straightforward to reconstruct due to clear track signatures and are commonly used as triggering objects in experiments. Due to their substantial mass (about 207 times that of electrons), muons emit negligible Bremsstrahlung radiation of energy up to $\approx 1TeV$, resulting in minimal energy deposition in earlier detector layers. A muon typically registers hits in four stations within the muon chambers, and these hits are combined to form high-purity muon tracks. In contrast, electrons and other particles are typically stopped in the electromagnetic and hadronic calorimeters. Muon detectors [51] are strategically placed outside the solenoid and magnetic field, more than 3 meters away from collision points, covering the entire detector body up to $|\eta| < 2.4$ (see Figure 20). To efficiently reconstruct, identify, and measure muon momentum, three types of muon chambers are employed within the flux-return yoke of the magnetic field. These sub-detectors utilize different techniques based on gas ionization chambers: Drift Tubes (DTs) cover the barrel region, Cathode Strip Chambers (CSCs) cover the endcaps, and Resistive Plate Chambers (RPCs) provide additional coverage in overlapping areas with fast, independent, and highly segmented detectors for triggering along with the addition of the state of the art Gas Electron Multipiers (GEMs) that will complement the existing detectors in the endcaps. The muon system achieves spatial resolution of about 100 μm and momentum resolution better than 2% for muons with transverse momentum (p_T) up to about 100 GeV.

Drift Tubes

The DT system [53] covers the region up to $|\eta| < 0.8$ and consists of four stations. Each chamber's fundamental component is a drift tube cell, featuring an aluminum cathode with a gold-plated anode wire at its center. These cells are filled with a gas mixture of 85% Ar and 15% CO_2. The electric field in the drift tube cell, shaped by aluminum strip electrodes, ensures a constant drift velocity of approximately 55 μm/ns for free electrons created by ionizing radiation. The central wire operates at a high voltage (3600 V), while the surrounding cathodes are at -1800 V, and the upper and lower electrodes are at +1800 V. These details can be seen in Figure 21

Electrons drifting towards the anode wire trigger a signal after a maximum drift time of 400 ns. The resulting electric pulse is amplified, digitized using a time-to-digital con-

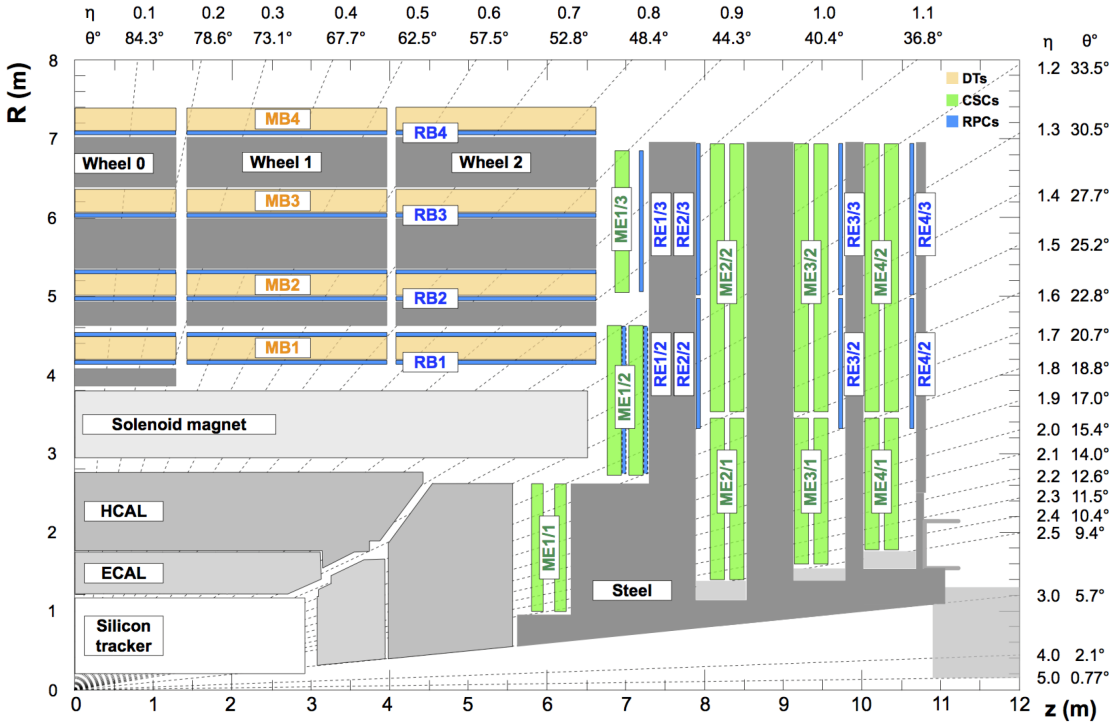


Figure 20: A schematic longitudinal view of one quarter of the CMS detector.[52]

verter (TDC), and then read out. The positional resolution in the DTs is around 100 μm per station, and the timing resolution in each superlayer is a few nanoseconds.

A DT chamber consists of twelve layers of parallel drift cells, grouped into three superlayers based on gas circulation, high voltage distribution, and front-end amplifiers (FEB). The layers are arranged with eight tubes oriented in the ϕ direction and four orthogonal to them (in the Z direction). Structural elements like honeycomb spacers separate these layers. The DTs form a twelve-sector ring along the transverse direction, with each sector housing four concentric chambers labeled MB1, MB2, MB3, and MB4 (Muon Barrel). Sectors 13 and 14 at the top and bottom of CMS contain additional chambers. The outer MB4 stations do not feature Z superlayers.

Resistive Plate Chambers

Resistive Plate Chambers (RPCs) [51] are utilized in both the barrel and endcap regions of the CMS muon detection system to provide rapid timing signals and enhanced sensitivity to background events. Their primary purpose is to furnish timing information for muon triggers with lower transverse momentum (p_T) thresholds across a broad pseudorapidity range ($|\eta| < 1.9$). An RPC consists of two parallel plates (an anode and a cathode) made of high-resistivity plastic material with a 2mm gap filled by a gas mixture (see Figure 22). When a charged particle traverses the RPC, it ionizes the gas, generating avalanches of electrons toward the positively charged side. These electrical signals are de-

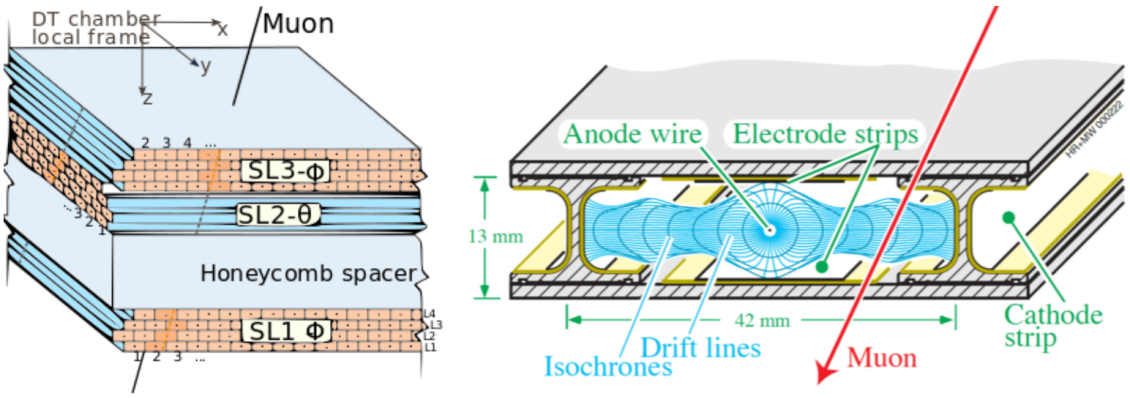


Figure 21: Left: A DT Chamber with three superlayers. Right: A drift cell showing the electric field lines in the gas volume.[54]

tected by external metallic strips. RPCs are integrated into the CMS subdetector system to improve beam crossing time measurement accuracy under high LHC luminosity conditions. They operate by detecting electron avalanches induced by charged particles passing through the RPC gap, allowing precise timing measurements on external strip readout planes with nanosecond resolution. The RPCs are arranged in stations aligned with the DT and CSC systems. In the barrel, there are four stations (RB1-RB4), including dual-layer RPCs adjacent to DT chambers, while the endcaps feature three stations (RE1-RE3), predominantly single-layered. RPC strips are oriented for measurement in the bending plane, parallel to DT wires in the barrel and CSC strips in the endcaps.Overall, RPCs provide fast response times at high event rates, offering exceptional time resolution (1 ns) within the LHC's 25 ns bunch separation. They contribute significantly to muon trigger efficiency and track reconstruction, particularly in resolving multi-hit track ambiguities.

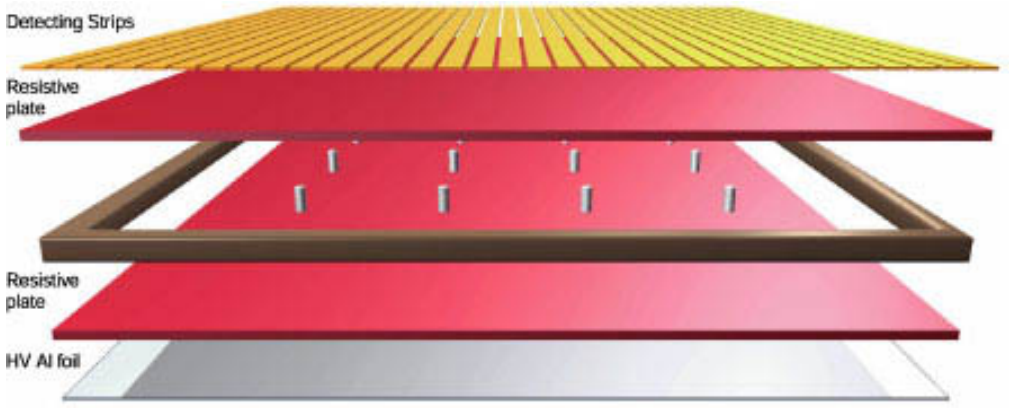


Figure 22: A Muon resistive plate chamber.[55]

Cathode Strip Chambers

Cathode Strip Chambers (CSCs) are located exclusively in the endcap regions covering

the η region from 0.9 to 2.4 due to higher muon rates, elevated background levels, and a strong, non-uniform magnetic field. CSCs offer a position resolution of approximately 100 μm and a timing resolution of 7 ns, attributed to their short drift path, making them suitable for fast triggering in the η region.The fundamental unit of CSCs comprises multi-wire proportional chambers with an array of anode wires arranged at an angle to cathode strips, filled with a gas mixture of 50% CO2, 40% Ar, and 10% CF4. CSCs are trapezoidal in shape and arranged in concentric rings centered on the beam line, each composed of six staggered layers between two aluminum cathode planes measuring muon position in radial (R) and φ coordinates (see Figure 23).

Electrons freed by incoming particles' ionization are collected by wires, while positive ions drift to strips, where the intersection of cathode strip and anode wire signals determines hit positions. CSCs operate similarly to drift chambers but with a stronger electric field and higher radiation resistance. They are segmented into strips and wires perpendicular to each other to measure φ and R coordinates, respectively.CSCs in the endcap region are grouped into four stations (ME1/1 to Me4/2) separated by return iron yokes perpendicular to the beam axis. ME1/1 chambers operate at 2900 V due to proximity to collision points and a stronger magnetic field, while others operate at 3600 V. ME1/1 chambers play a critical role in resolution and alignment, particularly in measuring the sagitta—distance from the center of the bending arc to its base—using narrower strips for enhanced precision.

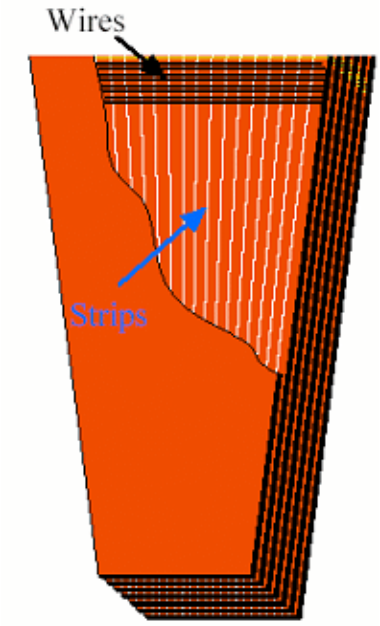


Figure 23: A Cathode Strip Chamber[57]

Gas Electron Multipiers

The Gas Electron Multiplier (GEM) detectors are integrated into the CMS muon system for improved detection in forward regions near the beam pipe, crucial for Phase-2 of the LHC due to increased radiation and event rates. GEMs excel in high-rate environments, enhancing muon system trigger capabilities and extending coverage into very forward regions. GEMs feature a key component called the GEM foil, consisting of a 50-micrometer-thick insulating polymer (polyimide) with copper conductors and microscopic holes in a hexagonal pattern. CMS GEM chambers utilize two PCBs enclosing a gas volume with three GEM foils, filled with an Ar/CO2 gas mixture. Incident ionizing particles create ionization, and a potential difference across the foils induces electron multiplication, generating readout signals on closely spaced strips.CMS GEMs are the largest ever deployed, with an area of about 0.5 m^2. Initial installation includes 144 chambers in each endcap, ready for Run-3 of the LHC. Expansion plans involve adding two more disks of GEM chambers per endcap between 2024 and 2026 for Phase-2 of the LHC. [58]

2.4.6 Forward Detectors

The CASTOR (Centauro And Strange Object Research) Calorimeter [59] at CMS is a sampling calorimeter located outside the main body of the detector at z = 14.4 m from the interaction point, covering the pseudorapidity region $5.2 < |\eta| < 6.6$. It uses Cherenkov sampling technology with quartz fibers and tungsten absorbing plates. The detector is segmented into 16 transverse and 14 longitudinal sections, with "long" and "short" fibers used to distinguish between electromagnetic and hadronic showers. Photomultiplier tubes (PMTs) connected via light guides convert detected light into electrical signals.

The Zero Degree Calorimeter (ZDC) [60] detects neutral particles in the $|\eta| > 8.5$ region, located at z = ±140 m from the interaction point. The CMS ZDC, similar in technology to CASTOR, uses quartz fibers and tungsten as effective material. It can measure spectator neutron multiplicity distribution and has been cross-calibrated to the 2010 dataset.

2.5 Trigger and DAQ system

During LHC operation, 40 million proton-proton collisions occur every second, triggering subdetectors. Due to data volume constraints, only a small fraction of these collisions, particularly those involving deep inelastic scattering leading to the production of new massive particles like c,b,t quarks, W,Z,H bosons are recorded. Proton bunches

collide every 25 nanoseconds, generating hundreds of particles. The high collision rate requires data storage pipelines capable of handling multiple interactions simultaneously. Detectors must have excellent time resolution to distinguish between events occurring within nanoseconds of each other. Data recording times for subsystems range from approximately 18 minutes to 2 hours.

To manage the high data rate, a trigger system is used to select interesting events for storage and analysis. This process, known as triggering, involves two levels: Level 1 (L1) triggering and High-Level Triggering (HLT), aiming to reduce data recording rates by a factor of 10^5. The CMS trigger system decides in real-time which data subset to read out and archive for offline analysis [61]. The L1 trigger [62] uses calorimeter and muon system data to make rapid decisions about event storage (Figure 24). The event rate is reduced from 40 MHz to 100 kHz at the L1 stage. The global muon trigger combines information from RPCs, CSCs,DTs to identify muon candidates. The L1 trigger decides in 3.2μs if the event is to be stored or not by searching for jets, muons, electrons and photons above p_T or η thresholds. The calorimeter trigger makes use of information from all calorimeters to reconstruct candidate jets, electrons or photons, and τ leptons, as well as the MissingM Energy Transverse (MET) in the event.

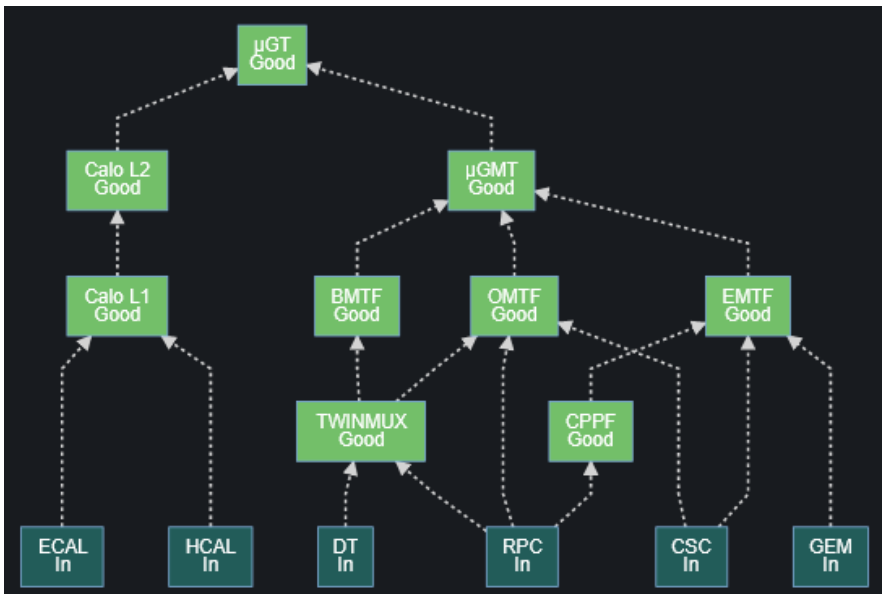


Figure 24: The overview of CMS L1 trigger system. Images obtained from grafana.

Events passing these selection criteria are transferred to the HLT for further offline analysis, reducing the event rate to approximately 100 events per second. During offline reconstruction, variables such as particle momentum may be recalculated based on the HLT outcome.The HLT [63] software consists of a stream-lined version of the offline reconstruction algorithms, optimised to comply with the strict time requirements of the on-

line selection. Particle flow objects, objects that are reconstructed using signals from all detectors like τ leptons and jets, are used in the HLT system. The use of PF reconstruction algorithm in the HLT system improves the energy resolution of trigger objects, increasing their efficiency with respect to offline selection, and provides more refined methods for pile-up mitigation.

2.6 Object reconstruction and identification

To reconstruct particles in the CMS detector, the particle flow (PF) algorithm [64] is utilized. This algorithm facilitates the reconstruction of electrons, muons, neutral hadrons, charged hadrons, and photons. It amalgamates information from all sub-detectors to discern the type of particles by exploiting compatible detector signals. Subsequently, this data is employed for subsequent reconstructions and identifications. The reason behind the focus in the reconstruction is because the thesis is specified in lepton identification techniques and it is of crucial importance to know which variables to use based on this knowledge.

2.6.1 Electron reconstruction

Electron reconstruction is based on the combination of measurements from the ECAL with the inner tracking system [65]. Because of the relatively long distance between the collision point and the ECAL, some of the electron's energy is lost through bremsstrahlung. It was shown with a test beam about 97% of its energy were deposited in a 5×5 crystal array. Overall about one third to 86% of the electron's energy is lost before it reaches the calorimeter, depending on the amount of material it passes through. To measure the initial energy accurately, it is therefore critical to collect all the radiated photons too.In the barrel region, this is done via the hybrid algorithm. It uses a seed crystal, the crystal with the most energy deposited in the considered region that is larger than a predefined minimum $E_{T,seed} > E_{T,seed}^{min}$. Along the transverse directions of the seed crystal arrays of 5×1 crystal are added in a range of N_{steps}, as long as their energy exceeds a minimum energy of E_{min} array. These clusters are then collected into a final global cluster called super cluster.In the endcaps the multi-5×5 algorithm is used. The seed crystal is the crystal with the local maximal energy of its four direct neighbours that again fulfils the requirement of $E_{T,seed} > E_{T,seed}^{min}$. Around these seed crystals, the energy is collected in 5×5 clusters, which may partly overlap.These clusters are then collected together to superclusters, if their energy exceeds a minimum [65]. Due to the large radiation losses of the electrons when they are curved in the magnetic field, the standard procedure of the Kalman filter [67] track

reconstruction is not used, but instead a special algorithm is deployed. This algorithm starts by seeding. It selects the first two or three hits in the tracker, from where the track can be initiated. Because of the importance of seeding for the reconstruction efficiency, two complementary algorithms are used for this. One starts from the supercluster energy and position in the ECAL and estimates the electron track to select the electron seed from all reconstructed seeds. The other algorithm relies on tracks, that have been reconstructed using the general algorithm for charged tracks, extrapolates them towards the ECAL and matches them to a supercluster. Further steps are then taken to increase the efficiency, such as using a matching-momentum criterion. These selected electron seeds are then used to build the electron tracks. This is done iteratively - layer by layer - with the energy loss taken into account. In case where several hits in a layer might be compatible with those predicted, several possible trajectory candidates are created. Over the whole track, only one missing hit is allowed. This procedure provides tracks up to the ECAL, where the fraction of momentum lost inside the tracker is calculated. Then, the tracks and the ECAL clusters are being associated with each other. There are criterions in place to obtain the highest reconstruction efficency while minimizing false positives. This leads to an overall efficiency for electron reconstruction for electrons from Z decay of $\approx 93\%$.

2.6.2 Muon reconstruction

To reconstruct muons, the tracks from the muon system and from the inner silicon tracker are treated separately at first [66]. The inner tracks are reconstructed by an algorithm based on Kalman filters [67]. Contrary to electrons, this works for muons because of their much higher mass and their therefore lower radiation losses through bremsstrahlung. So-called "standalone-muon tracks" are built by using the information from the muon system. Then the tracker muon tracks, that satisfy a transverse momentum of $p_T > 0.5$ GeV and a total momentum of $p > 2.5$ GeV, are reconstructed by extrapolating tracker tracks from the tracker to the muon system. To be classified as a tracker muon, at least one muon segment has to match the extrapolated track. Global muons are muons that are built from the outside of the detector to the inside through the matching of standalone muon tracks of the muon system to inner tracks. Because of the high efficiency of these reconstructions, about 99% of muons can be reconstructed as either tracker or global muons, or both. Generally, muons with a low p_T are identified as tracker muons, as they often only reach the innermost muon segment. Late showering hadrons that pass the HCAL might also be detected by the first muon station. Higher p_T muons usually pass through the whole muon system and are reconstructed as global muons. These muons are then passed to the particle-flow algorithm [64]. There all the information from all sub-

detectors is combined. It applies a set of selection criteria to reconstructed candidates. These criteria are based on the quality parameters of the reconstruction. To calculate the momentum of the muons, the Tune-P algorithm is used [68]. It selects the p_T measurement from a refit to reduce tails in the momentum resolution distribution caused by bad fits. The inner track is used if the p_T is smaller than 200 GeV; otherwise, the track with the lowest χ^2 of its fit is chosen. A muon that passes the particle-flow algorithm is excluded from being possibly reconstructed as another particle.

2.7 Trigger performance in 2023

As stated in section 2.5 The Level-1 (L1) trigger system consists of custom hardware processors that receive data from calorimeter and muon systems and reduce the event rate from 40 MHz to about 100 kHz. In Run 2, the L1 muon trigger system has undergone significant upgrades. It now integrates inputs from all available subdetectors and comprises three distinct regional track finders: the barrel muon track finder (BMTF) covering the range of $0 < |\eta| < 0.83$, the overlap muon track finder (OMTF) covering $0.83 < |\eta| < 1.24$, and the endcap muon track finder (EMTF) covering $1.24 < |\eta| < 2.40$. Each track finder receives "trigger primitives" (TPs) from individual subdetectors, which include position coordinates (θ and ϕ), direction, and timing information relative to a collision bunch crossing.

In contrast to Run 1, where adjacent RPC hits were processed independently, in Run 2, they are clustered into TPs before being utilized in track reconstruction. Additionally, RPC TPs are combined with nearby DT segments in the barrel region to enhance overall efficiency and timing for the BMTF. Each track finder employs processor boards with field-programmable gate arrays (FPGAs) to reconstruct muons, following a standardized sequence. TPs aligned in θ and ϕ are grouped to form tracks, which traverse the subdetector stations. The angular deflection between stations aids in assigning a transverse momentum (p_T) value, primarily based on the ϕ coordinate, as the magnetic field influences the curvature of charged-particle tracks in the transverse plane.

However, owing to variances in subdetector characteristics and magnetic field strength, each track finder has its unique track-building and p_T assignment logic. The Global Muon Trigger collects reconstructed muons from the three track finders, eliminating geometrically overlapping tracks based on p_T and quality criteria. The remaining tracks are then forwarded to the L1 Global Trigger, which determines whether to proceed with event processing by the HLT. Quality assessment is based on the number and location of TPs along a given track, with tracks meeting stringent quality requirements exhibiting the best p_T resolution and being selected for single-muon L1 seeds. In contrast, criteria are relaxed for multi-muon L1 seeds to enhance efficiency, allowing tracks with fewer TPs to be considered.

The following work was done by the muonDPG group of which was a part of so the results were conducted from me as well.

The efficiency of the Level-1 (L1) muon trigger is determined using the Tag-and-Probe (T&P) technique, a widely adopted method in high-energy physics for measuring trigger and reconstruction efficiencies. This method provides a data-driven approach, reducing reliance on simulation and allowing for more accurate assessments.

In this study, the **numerator** of the efficiency calculation consists of all probe muons that meet specific criteria. These probe muons must be matched within a distance $\Delta R < 0.1$ to an L1 trigger with transverse momentum $p_T > 22$ GeV, satisfying the stringent L1 quality criteria. These criteria are enforced to select L1 muons, referred to as L1 seeds, which initiate single or double muon reconstruction at the High-Level Trigger (HLT). The p_T thresholds of 22 GeV correspond to the minimum thresholds utilized in unprescaled single-muon and double-muon L1 seeds during Run 2, respectively.

Tag Muons

- × Pseudorapidity $|\eta| < 2.4$
- × Tight ID criteria
- × High-Level Trigger (HLT) isolated muon
- × Transverse momentum $p_T \geq 27$ GeV
- × Relative isolation < 0.15

Probe Muons

- × Pseudorapidity $|\eta| < 2.4$
- × Tight ID criteria
- × Matched within $\Delta R < 0.2$ to an offline reconstructed muon
- × Relative isolation < 0.15

Tag-and-Probe Pair

- × Separation $\Delta R(\text{tag}, \text{probe}) > 0.4$

The tag muon serves as a well-identified and isolated reference particle, ensuring a clean and unbiased sample. The probe muon is the particle for which the efficiency is being measured, and it is subject to looser selection criteria initially to allow an unbiased assessment of the efficiency of the tighter criteria.

The tag and probe method is employed because it allows for an unbiased and direct measurement of the efficiency of a particular selection criterion or trigger. By using a

well-identified tag particle, we can ensure that the event sample is pure and predominantly signal. The method also allows for the determination of efficiencies in real data, accounting for detector effects and other real-world conditions that may not be perfectly modeled in simulations. Furthermore, the T&P method helps in subtracting background contributions by fitting the invariant mass distribution of the tag and probe pairs, which ensures that the efficiency measurement is as accurate and precise as possible.

The L1 efficiency in data is depicted in Figure 25 for 2023, showing variations as a function of the offline p_T and η of the muon, as well in phi along with the 2D plot in η and ϕ grid.

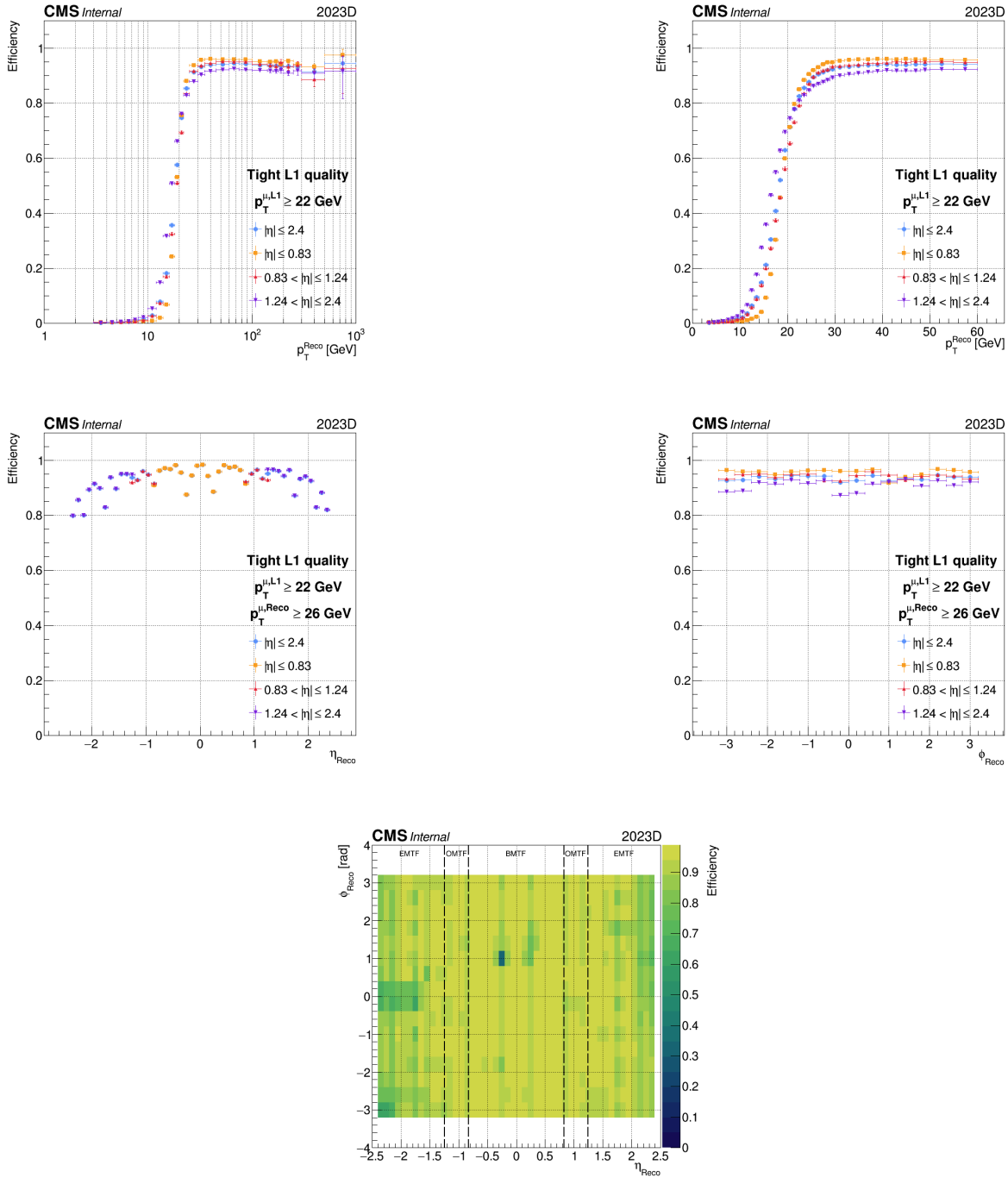


Figure 25: Efficiency for 2023 data.

Another important variable is also the charge misidentification which is a metric of how good the recontruction is between the L1 and offline muons. The results can be seen in Figure 26.

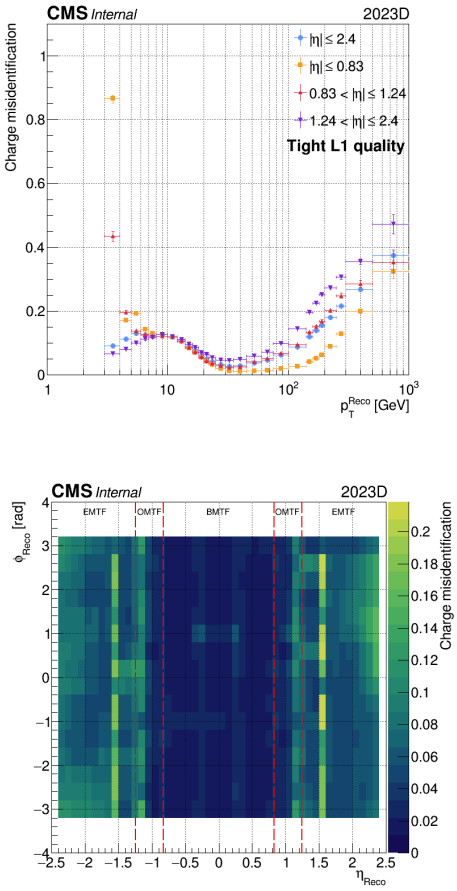


Figure 26: Charge misidentification for 2023 data.

The performance is very good in general with the best performance coming in $|\eta| < 0.83$ (BMTF region) in efficiency where after 30 GeV the efficiency is close to 1. As for the charge misidentification is close to 0 in all η regions except with a small discreptancy coming in the "chimneys" in the EMTF region.

3 $t\bar{t}H(c\bar{c})$ Analysis /Framework

The discovery of the Higgs (H) boson using LHC Run 1 data by the ATLAS and CMS experiments in 2012 [69] marked a pivotal moment in our understanding of the electroweak (EW) symmetry breaking mechanism. The measured Higgs boson mass is 125.38 ± 0.14 GeV [70]. The observed interactions with gauge bosons and third-generation fermions, alongside all measured properties, are consistent with the predictions of the standard model (SM). One of the next major physics priorities of the LHC is to establish interactions with second-generation fermions. Recently, the CMS Collaboration reported the first evidence of Higgs boson decays to muons [71]. A critical upcoming milestone is observing its coupling to second-generation quarks.

To this end, our focus is on the search for H boson decay to a charm quark-antiquark pair ($c\bar{c}$). The corresponding Yukawa coupling, y_c, can be significantly modified by physics beyond the SM (BSM). However, the SM-predicted branching fraction is small, around 3%, and the high production rate of quark and gluon jets at the LHC, coupled with the challenges in identifying charm jets in a hadronic environment due to their properties lying between lighter quark and bottom quark jets, makes this a challenging measurement.

The initial $H \to c\bar{c}$ searches conducted by the ATLAS [72] and CMS [73]collaborations target the associated production of a Higgs boson with a vector boson (W or Z). Despite the low cross section, this production mode, focusing on the leptonic decays of the W or Z boson, effectively suppresses to negligible levels the overwhelming QCD multijet background. The CMS analysis, using 138 fb^{-1} of LHC Run 2 data and developing novel Deep Learning (DL) techniques, provides the most stringent constraints on y_c to date [74].

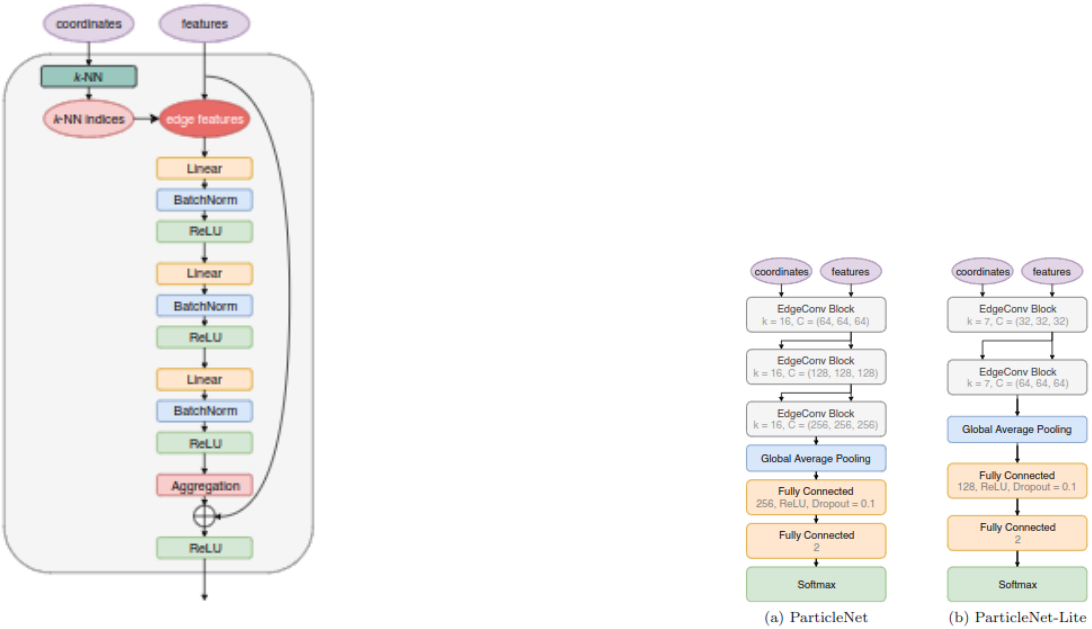
Despite these advancements, projections for the High-Luminosity LHC (HL-LHC) show that the current CMS VH analysis sensitivity will not be sufficient to achieve observation of $H \to c\bar{c}$ production. Therefore, we present a search for $H \to c\bar{c}$ targeting the production mode where the H boson is produced in association with a top-antitop pair ($t\bar{t}H$). Although $t\bar{t}H$ production has a smaller cross-section than VH, the presence of top quarks in the decay chain can, when efficiently utilized, offer powerful discriminating factors to suppress SM background processes. Achieving this requires innovation in all aspects of the analysis.

One of the primary hurdles in the search for $H \to c\bar{c}$ is effectively reconstructing the pair of c quarks originating from the Higgs decay, while simultaneously achieving substantial rejection of light quarks (u, d, s) and gluons, along with b quarks, which also contribute to the background in this quest. The process of identifying a jet arising from the hadronization of a c quark, termed "charm tagging," relies on several characteristics

including the extended lifetime of hadrons containing a c quark, displaced vertices and tracks, and potentially the presence of non-isolated electrons and muons. Conventional charm tagging algorithms utilize variables tailored to these properties and employ multivariate techniques (e.g., boosted decision trees) to enhance background rejection. However, owing to the shorter lifetime of c hadrons compared to that of b hadrons, and their similarity to hadrons produced with light quarks, the performance of traditional charm tagging algorithms is suboptimal. In the 2016 VH ($H \to c\bar{c}$) analysis [74], the deep neural network–based "DeepAK8" algorithm, where AK8,AK4 stands for anti-k_T jet recontruction [75] with jet radius 0.4, 0.8 respectively [76] was adapted to the radius 1.5 jets for the identification of the c quark pairs and showed very high performance. For this analysis, a more advanced $c\bar{c}$-tagging algorithm based on graph neural networks is adopted, leading to further improvement in the tagging performance compared to the DeepAK8 algorithm.

3.1 The ParticleNet Tagger

In this analysis, the ParticleNet [77] algorithm is adopted for the reconstruction and identification of the $H \to c\bar{c}$ candidates. At its core, the ParticleNet algorithm treats a jet as an unordered set of its constituents and builds a permutation-invariant graph neural network to effectively exploit the correlation between the jet constituents.



(a) ParticleNet (b) ParticleNet-Lite

Figure 27: **On the right**: The architectures of the ParticleNet network. **On the left**: The structure of the EdgeConv block

The ParticleNet architecture incorporates EdgeConv, a new neural-network module suitable for CNN-based high-level tasks on point clouds [78], operations extensively and adopts the dynamic graph update approach. However, several design modifications are implemented in ParticleNet compared to the original DGCNN [79] to better align with the jet tagging task, including adjustments to the number of neighbors, the configuration of the Multi Layer Perceptron (MLP) in EdgeConv, and the incorporation of shortcut connections. Figure 1 depicts the structure of the EdgeConv block utilized in this study. Initially, the EdgeConv block identifies the k nearest neighboring particles for each particle by utilizing the "coordinates" input to compute distances. Subsequently, the "edge features" input is constructed from the "features" input using the indices of the k nearest neighboring particles for the EdgeConv operation. The EdgeConv operation comprises a three-layer MLP, with each layer consisting of a linear transformation followed by batch normalization and a rectified linear unit (ReLU). Inspired by ResNet, a shortcut connection running parallel to the EdgeConv operation is included in each block to facilitate the direct passage of input features. Each EdgeConv block is characterized by two hyperparameters: the number of neighbors k and the number of channels C = (C1, C2, C3), corresponding to the number of units in each linear transformation layer.

The ParticleNet architecture utilized in this study is illustrated in Figure **??**, comprising three EdgeConv blocks. The first EdgeConv block utilizes the spatial coordinates of the particles in the pseudorapidity-azimuth space to compute distances, while subsequent blocks utilize learned feature vectors as coordinates. The number of nearest neighbors k is set to 8 for all three blocks, and the number of channels C for each EdgeConv block is (64, 64, 64), (96, 96, 96), and (128, 128, 128), respectively. Following the EdgeConv blocks, a channel-wise global average pooling operation is applied to aggregate the learned features over all particles in the cloud. This is followed by a fully connected layer with 128 units and Rectified Linear Unit (ReLU) activation. To prevent overfitting, a dropout layer with a drop probability of 0.1 is included. Subsequently, A fully connected layer with two units, followed by a softmax function[1], is employed to generate the output for the binary classification task. So, the output is a set of propabilities scores normalized to lie between 0 and 1.

To perform our jet tagging, we utilize the output of the ParticleNet tagger [77], which

[1] The softmax function is defined as

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

for $i = 1, \ldots, K$, where $\mathbf{z} = (z_1, \ldots, z_K)$ is the input to the softmax function. In the context of the Boltzmann distribution, it represents the probability of a system being in a certain state i, where z_i corresponds to the energy of state i, and the denominator normalizes the probabilities over all possible states.

is a multiclass graph neural network trained to discriminate between different jet origins. We utilize the scores corresponding to the the classes described in Table 5.

Class	Jets, at generator level, ...	Definition
bb	with two or more b hadrons	nBHadrons > 1
b	with exactly one b hadron	nBHadrons = 1
cc	with two or more c, but no b hadrons	nBHadrons = 0 & nCHadrons > 1
c	with exactly one c, but no b hadrons	nBHadrons = 0 & nCHadrons = 1
uds	produced by u, d, or s quarks	hadronFlavour = 0 & |partonFlavour| $\in \{1, 2, 3\}$
g	produced by gluons	hadronFlavour = 0 & partonFlavour = 21

Table 5: Jet classes described in ParticleNet and their definitions.

Based on this, we define two scores for discrimination. One differentiates between heavy flavor (HF) jets (types bb, b, cc, or c) and light flavor (LF) jets (types uds or g):

$$\text{score}[\text{HF vs. LF}] = \frac{\text{score}[bb] + \text{score}[b] + \text{score}[cc] + \text{score}[c]}{\text{score}[bb] + \text{score}[b] + \text{score}[cc] + \text{score}[c] + \text{score}[uds] + \text{score}[g]} \tag{87}$$

The second differentiates between charm and bottom-induced jets:

$$\text{score}[b \text{ vs. } c] = \frac{\text{score}[bb] + \text{score}[b]}{\text{score}[cc] + \text{score}[c] + \text{score}[bb] + \text{score}[b]} \tag{88}$$

Instead of using the entire score output in our analysis, we divide the two-dimensional plane defined by these two scores into categories enriched in either LF, c, or b jets. The tagging efficiencies in each category are then calibrated. Each category is labeled with L, C, or B, indicating the dominating flavor (LF, c, or b jets) in that category. The following index increases with higher purity in the corresponding flavor. We further define loose, medium, and tight working points, starting from index 0, 1, and 2, respectively.The performance of the algorithm compared with the previous CMS tagger can be seen in Figure 28.

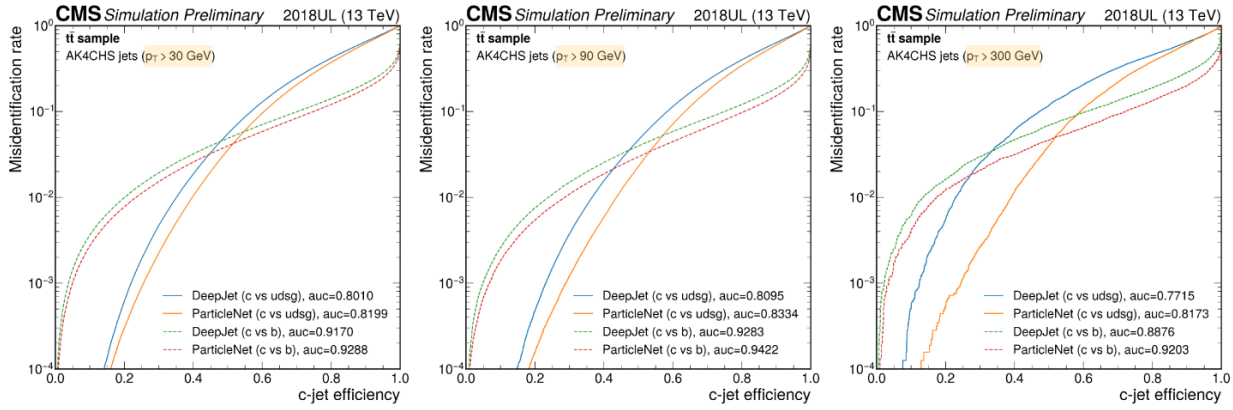


Figure 28: Performance for c-tagging.

To also improve the sensitivity of this analysis, we are using discrete working points (WPs) instead of full shapes of the flavor tagging discriminants where we defined a set of exclusive b-tagging and c-tagging WPs and each jet is uniquely classifed into one of the 11 categories (B0-B4, C0-C4, or L0) as Figure 29 indicates.

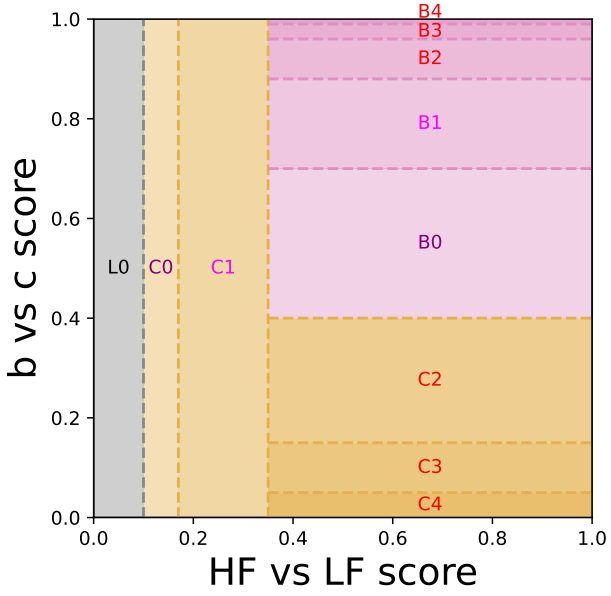


Figure 29: Jet categories categorization used in the analysis

3.2 Particle Transformer and Event Classifier

In terms of the event classification we are using a state-of-the-art graph neural network architecture inspired by the "jet" <=> "particle cloud" representation proposed in ParticleNet, treating an event as a cloud of objects and exploit permutation-invariant graph networks for event-level classification. We use for JETCLASS, a new large and compre-

hensive dataset to advance deep learning for jet tagging. The JETCLASS dataset consists of 100 M jets for training, about two orders of magnitude larger than existing public datasets. It also includes more types of jets, several of which have not been explored for tagging yet but are promising for future applications at the LHC. Based on this dataset, we use Particle Transformer(ParT) [80], a new Transformer-based architecture for jet tagging, which here is used for event classification.

The dataset includes a total of 10 types of jets. Representative jets of each type are visualized as particle clouds in Figure 30. The circles, triangles(upward- or downward-directed), and pentagons represent the particle types, which are hadrons, leptons (electrons or muons), and photons,respectively. The solid (hollow) markers stand for electrically charged (neutral) particles. The marker color reflects the displacement of the particle trajectory from the interaction point of the proton-proton collision, where a larger displacement results in more blue. The jets in this dataset generally fall into two categories. The background jets are initiated by light quarks or gluons (q/g) and are ubiquitously produced at the LHC. The signal jets are those arising either from the top quarks (t), or from the W, Z or Higgs (H) bosons. For top quarks and Higgs bosons, we further consider their different decay modes as separate types, because the resulting jets have rather distinct characteristics and are often tagged individually. The use of jet tagging typically involves selecting one (or a few) specific type of signal jets with high confidence, and rejecting background jets as much as possible, since the background jets usually appear orders of magnitude more frequently than the targeted signal jets. Note that for several types of signal jets in this dataset, such as $H \to 4q$, $H \to \ell\nu qq'$, and $t \to b\ell\nu$, no dedicated methods have been developed so far to tag them.

The dataset offers all constituent particles of each jet as inputs for jet tagging. It's worth noting that the number of particles fluctuates from jet to jet, usually ranging from 10 to 100, with an average of 30 to 50, contingent upon the jet type. For each particle of a jet, three categories of features are provided:

× **Kinematics**: This encompasses energy and momentum, represented by the 4-vector (E, px, py, pz) in GeV units, fundamental quantities detected by a particle detector. All other kinematic variables can be derived from these 4-vectors.

× **Particle Identification**: This includes the electric charge, denoted as ±1 for positively/negatively charged particles and 0 for neutral particles, as well as the particle identity determined by the detector systems. A 5-class encoding is utilized to maintain consistency with current LHC experiments: charged hadron ($\pi^{\pm}, K^{\pm}, \rho/r\bar{h}o$), neutral hadron (0), electron (±11), muon (±13), and photon (22). This information

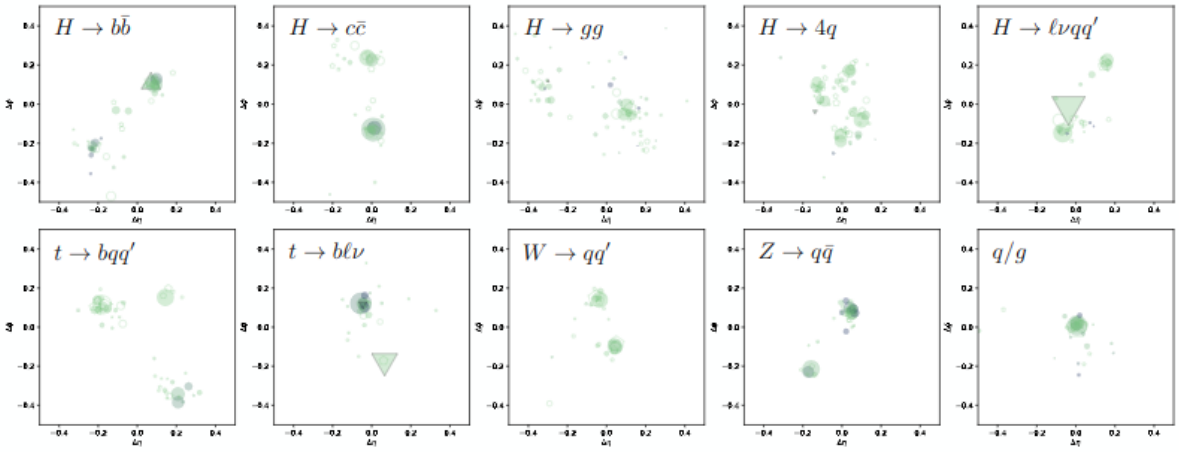


Figure 30: Examples of the 10 types of jets in the JETCLASS dataset, viewed as particle clouds. Each particle is displayed as a marker,with its coordinates corresponding to the flying direction of the particle, and its size proportional to the energy. [80]

is particularly vital for tagging jets involving a charged lepton, such as $H \to \ell\nu qq'$ and $t \to b\ell\nu$, where leptons can be reliably identified at the LHC.

× **Trajectory Displacement**: This comprises the measured values and errors of the transverse and longitudinal impact parameters of the particle trajectories in mm units, totaling 4 variables. These measurements are solely accessible for electrically charged particles, with a value of 0 assigned to neutral particles.

An overview of the Particle Transformer (ParT) architecture is presented in Figure 31. For a jet with N particles, ParT makes use of two sets of inputs: the particle input includes a list of C features for every particle and forms an array of a shape (N, C); the interaction input is a matrix of C' features for every pair of particles, in a shape (N, N, C'). The particle and interaction inputs are each followed by an MLP to project them to d- and d'-dimensional embedding, $x_0 \in \mathbb{R}^{N\times d}$ and $U \in \mathbb{R}^{N\times N\times d'}$, respectively. Unlike Transformers for Natural Language Processing (NLP) and vision, we do not add any ad-hoc positional encodings, as the particles in a jet are permutation invariant. The spatial information (i.e., the flying direction of each particle) is directly included in the particle inputs. We feed the particle embedding x_0 into a stack of L particle attention blocks to produce new embeddings, $x_1, ..., x_L$ via multi-head self-attention. The interaction matrix U is used to augment the scaled dot-product attention by adding it as a bias to the pre-softmax attention weights. The same U is used for all the particle attention blocks. After that, the last particle embedding x_L is fed into two class attention blocks, and a global class token x_{class} is used to extract information for jet classification via attention to all the particles, follow-

ing the CaiT approach [81]. The class token is passed to a single-layer MLP, followed by softmax, to produce the final classification scores.

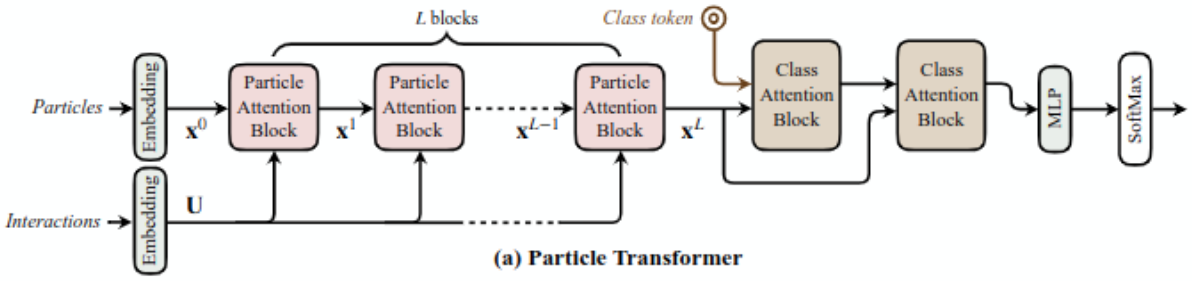


Figure 31: The architecture of Particle Transformer. [80]

For the particle interaction features, in a pair of particles a, b with 4-vectors p_a, p_b, we calculate the following 4 features:

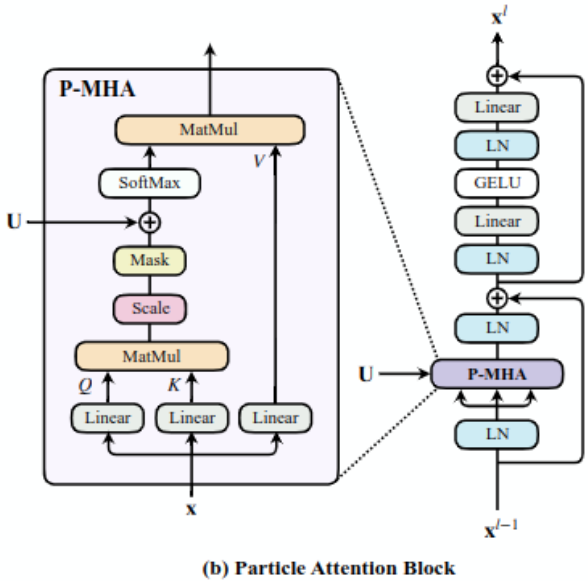
$$\Delta = \sqrt{(y_a - y_b)^2 + (\phi_a - \phi_b)^2}, \quad k_T = \min(p_{T,a}, p_{T,b})\Delta, \quad z = \frac{\min(p_{T,a}, p_{T,b})}{p_{T,a} + p_{T,b}}, \quad m^2 = (E_a + E_b)^2 - \|p_a + p_b\|^2,$$

where y_i is the rapidity, ϕ_i is the azimuthal angle, $p_{T,i} = \sqrt{p_{x,i}^2 + p_{y,i}^2}$ is the transverse momentum, $p_i = (p_{x,i}, p_{y,i}, p_{z,i})$ is the momentum 3-vector and $\|\cdot\|$ is the norm, for $i = a, b$. Since these variables typically have a long-tail distribution, we take the logarithm and use $(\ln\Delta, \ln k_T, \ln z, \ln m^2)$ as the interaction features for each particle pair.

A key component of ParT is the particle attention block. As illustrated in Figure 32 , the particle attention block consists of two stages. The first stage includes a multi-head attention (MHA) module with a LayerNorm (LN) layer both before and afterwards. The second stage is a 2-layer MLP, with an LN before each linear layer and Gaussian Error Linear Unit (GELU) nonlinearity in between. Residual connections are added after each stage. The overall block structure is based on NormFormer , however, we replace the standard MHA with P-MHA, an augmented version that can also exploit the pairwise particle interaction directly. The P-MHA is computed as

$$\text{P-MHA}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}} + U\right)V,$$

where Q, K and V are linear projections of the particle embedding x_l. Essentially, we add the interaction matrix U to the pre-softmax attention weights. This allows P-MHA to incorporate particle interaction features designed from physics principles and modify the dot-product attention weights, thus increasing the expressiveness of the attention mechanism.



(b) Particle Attention Block

Figure 32: The architecture of Particle Attention Block. [80]

As illustrated in Figure 33, the class attention block has a similar structure as the particle attention block. However, unlike in the particle attention block where we compute the self attention between particles, here we compute the attention between a global class token x_{class} and all the particles using the standard MHA. Specifically, the inputs to the MHA are $Q = W_q x_{\text{class}} + b_q$, $K = W_k z + b_k$, $V = W_v z + b_v$, where $z = [x_{\text{class}}, x_L]$ is the concatenation of the class token and the particle embedding after the last particle attention block, x_L.

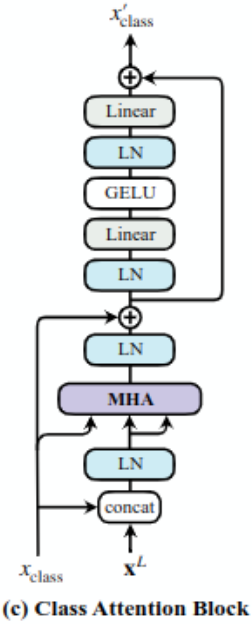
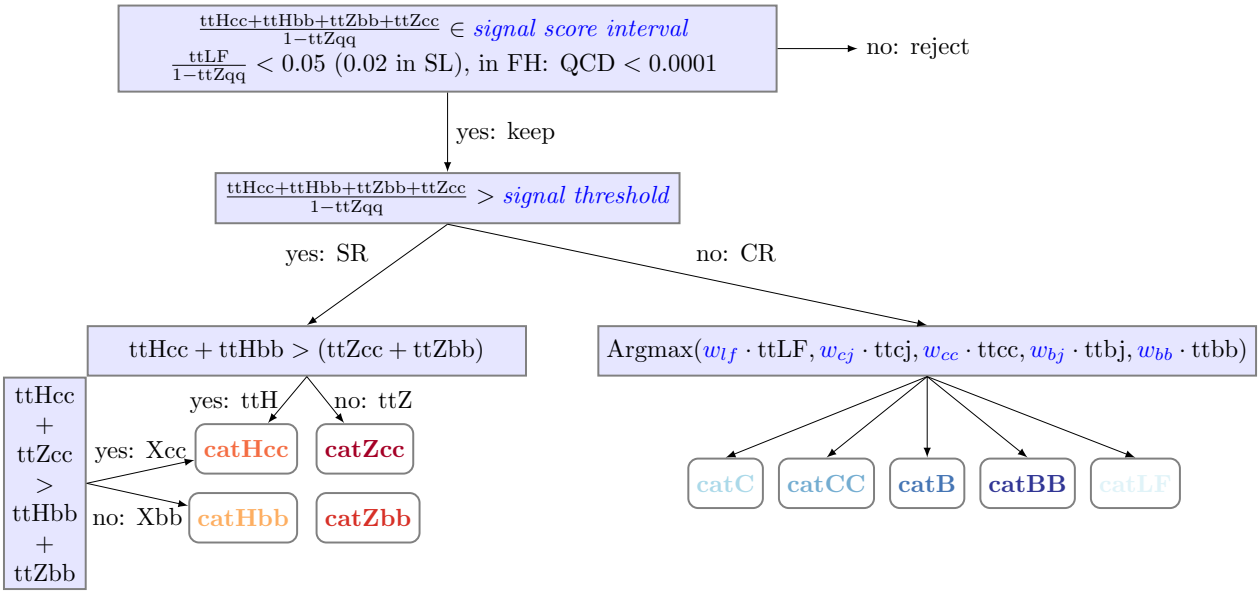


Figure 33: The architecture of Class Attention Block.[80]

Especially for the analysis' perspective, the Particle Transformer model takes $\ln p_T$, $\ln E$, η, and ϕ for each lepton and jet, as well as $\ln$MET. Additionally, leptons are assigned an extra "isEl/isMu" flag to give Particle Transfomer the ability to distinguish between the different streams more clearly. For each event used during training, all highest $p_T$jets are used up to a channel dependent maximum: 10 for full-hadronic, 8 for single leptonic, and finally 6 for the dilepton channel. Each jet is paired with eight flags indicating which ParticleNet b- and c-tagging score thresholds the jet passes. Events in the training sample are weighted by cross-section, then renormalized such that the sum of the weights in each category is equal. In the single-lepton and dilepton channels, a total of ten event categories are used during the training, one for each of the following processes: The five $t\bar{t}$ backgrounds described previously; the three Z backgrounds $t\bar{t}Z(Z\rightarrow c\bar{c})$, $t\bar{t}Z(Z\rightarrow b\bar{b})$, and $t\bar{t}Z(Z\rightarrow q\bar{q})$; the $t\bar{t}H(H\rightarrow b\bar{b})$ background, and finally the $t\bar{t}H(H\rightarrow c\bar{c})$ signal process. The reason fot the use of the similar topologically $t\bar{t}Zc\bar{c}$ process is a control measure applied in this analysis. The trained model assigns ten scores that reflect the probability of an event falling into each category, but since the scores must sum to 1, the result is nine independent Particle Transformer scores. The fully hadronic channel uses the same set-up, but adds an additional category for QCD events, bringing the total event category number up to eleven and the number of independent scores to 10.

To validate the modeling of the backgrounds and the strategy of the background estimation in the analysis region, a dedicated validation region (VR) is designed, as close

as possible to the signal region. Within the VR, similar control regions (CRs) and signal regions (SRs) are designed as in the analysis region, all signal-depleted. Similarly to the analysis region, a rescaling is performed to achieve a similar purity and event yield in each SR and CR in the VR, as in the analysis region. To account for differences, the cuts on the tt + lf score in the VR are adapted to be < 0.08 in the DL and FH channels, and < 0.03 in the DL channel.The full categorization scheme is summarized schematically in Figure 34,while a couple of scores for DL and SL channel are presented in Figure 35 for the midscore validaton region.



	analysis region (AR)	validation region (VR)
signal score interval	[0.6, 1]	[0.4, 0.6]
signal threshold	0.85	0.58
$w_{lf}, w_{cj}, w_{cc}, w_{bj}, w_{bb}$	100, 12, 4, 2, 1	100, 12, 4, 2, 1

Figure 34: Diagram of the categorization cutflow. What is indicated in the boxes in italic dark blue differs between search and validation regions as specified in the table below it.

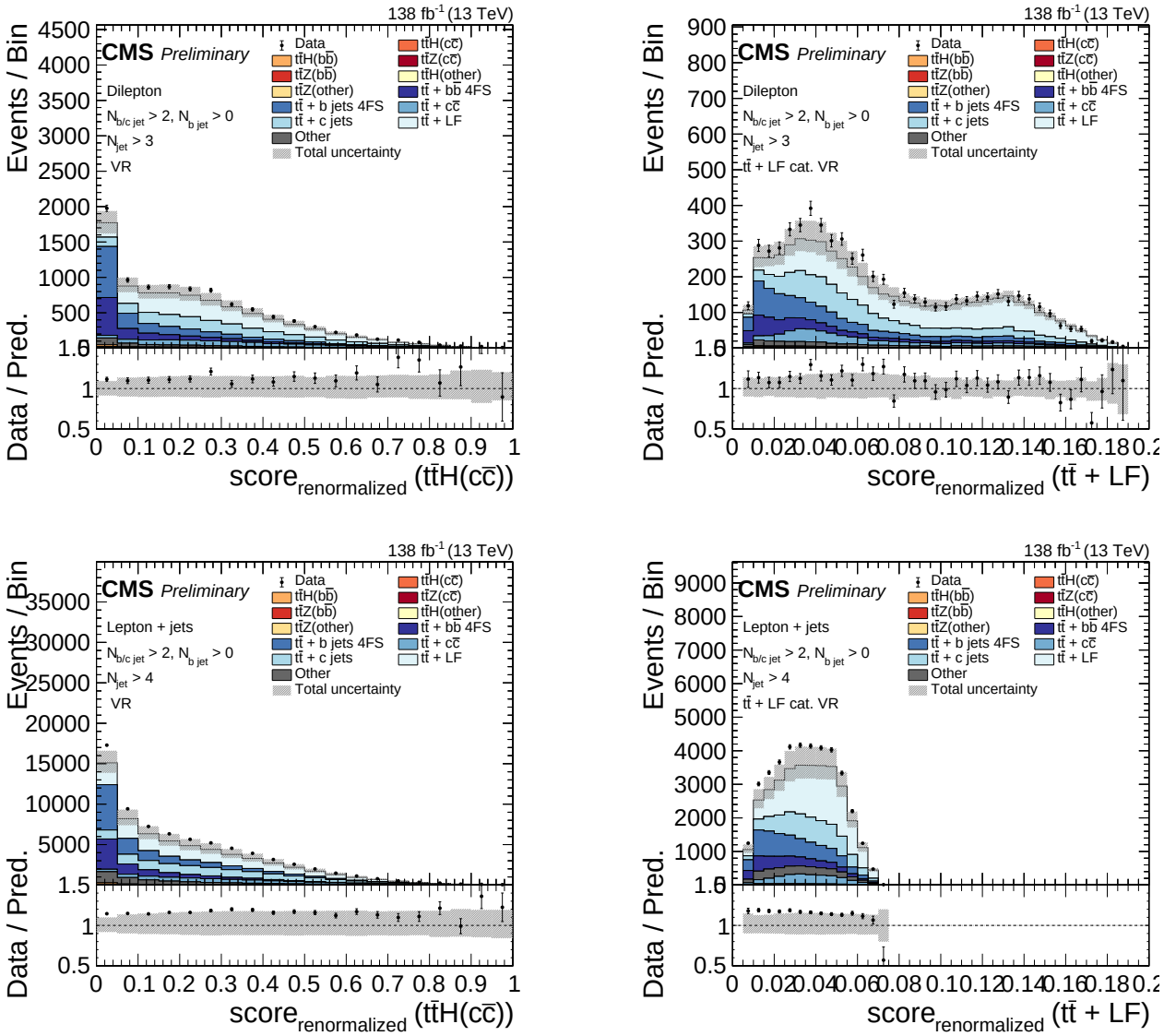


Figure 35: Renormalized Particle Transformer scores in all categories in the VR in the DL channel (up right and left plot) and SL channel (down right and left plot). The scores are plotted for one signal process ($t\bar{t}H(c\bar{c})$) and one background process($t\bar{t}+LF$) per channel.

3.3 Lepton Identification

After we explained the analysis aim and the state-of-start tools for its implementation, it is of crucial importance to identify techniques to select leptons arising from top quark decays via a subsequent W boson leptonic decay. In the associated production of top quarks with vector bosons, leptons can also originate from the vector boson decay themselves, with similar reconstructed properties of "prompt" leptons. The "nonprompt" leptons arise from hadron decays, can be hadrons misidentified as leptons, as well as to be electrons coming from external photon conversions in the detector. Such leptons ap-

pear to be less isolated than those of the prompt origin and can be additionally associated with displaced secondary vertices. The nonprompt leptons represent one of the major backgrounds in the study of multilepton final states. In the $t\bar{t}H(c\bar{c})$ analysis the 2 channels that have final states with multiple leptons are the dilepton Fig.36a and single-lepton Fig.36b channels.

(a) DL channel (b) SL channel

Figure 36: Feynman diagrams for the leptonic channels.

For this analysis, a set of identification criteria for prompt leptons was set up and is summarised in the next subsections.

3.3.1 Muon identification

Three types of muons are defined, as summarized in Table 6. One type is designated for the single-lepton channel, while the other two define the leading and sub-leading muons in the dilepton channel. The latter definition is also applied in the fully-hadronic and single-lepton channels in order to veto additional muons. In each case, particle-flow reconstructed muons are used. They are required to fulfill the following quality criteria, designed to select high-purity prompt muons originating from weak boson decays [82]: the muons must fulfill the "CutBasedIdTight" quality criteria as well as different isolation criteria, depending on the type of the muon. The isolation criteria are based on the relative PF-isolation with an isolation cone of 0.4. The muon p_T scale and resolution are corrected for biases due to detector misalignment by applying the "Rochester corrections" [83].Finally, the muons must fulfill kinematic requirements on p_T and η, depending on the type of the muon. The p_T cut in the single-muon channel has been raised from 26 GeV

in 2016 and 2018 to 29 GeV in the 2017 dataset following an increased trigger threshold in 2017. The corresponding loss in signal efficiency amounts to approximately 5% and approximately 6% in $t\bar{t}$ + jets background events. The selection criteria are summarized in Table 6.

Channel	Particle	Muon ID	Max. rel-iso	Min. p_T [GeV]	Max. $\|\eta\|$
SL	Muon	CutBasedIdTight	0.15 (PFIsoTight)	26/29/26	2.4
DL	Leading muon	CutBasedIdTight	0.25 (PFIsoLoose)	25	2.4
DL	Sub-leading muon	CutBasedIdTight	0.25 (PFIsoLoose)	15	2.4

Table 6: Muon identification criteria for different channels.

3.3.2 Electron identification

Similar to the muon definition, three types of electrons are defined as summarized in Table 7. One type is designated for the single-lepton channel, while the other two define the leading and sub-leading electrons in the dilepton channel. The latter definition is also applied in the fully-hadronic and single-lepton channels in order to veto additional electrons.

In each case, particle-flow reconstructed electrons are utilized, selected based on stringent criteria to identify high-purity prompt electrons originating from weak boson decays [84]. Specifically, electrons must satisfy the tight working point criteria of the cut-based electron ID ("cutBasedElectronID-Fall17-94X-V2-tight"), which includes recommended cuts on impact parameters. Moreover, they must meet isolation requirements based on effective-area-corrected isolation, calculated within an isolation cone of 0.3. The effective area method is employed to correct for the influence of pileup, estimating and subtracting the contribution from nearby particles not associated with the electron of interest. Electrons located outside the barrel ($|\eta_{\text{Supercluster}}| < 1.4442$) and endcap ($|\eta_{\text{Supercluster}}| > 1.5560$) regions are excluded from the analysis. Additionally, stringent kinematic requirements on transverse momentum (p_T) and pseudorapidity (η) are imposed depending on the electron type. Notably, in the single-electron channel, the p_T threshold was increased from 29 GeV in 2016 to 30 GeV in the 2017 and 2018 datasets due to heightened trigger thresholds.

Channel	Particle	Electron ID	Min. p_T [GeV]	Max. $\|\eta\|$
SL	Electron	mvaFall17V2Iso-WP80	29/30/30	2.4
DL	Leading electron	mvaFall17V2Iso-WP90	25	2.4
DL	Sub-leading electron	mvaFall17V2Iso-WP90	15	2.4

Table 7: Electron identification criteria for different channels.

4 Boosted Decision trees

4.1 Decision trees

Decision trees are a machine learning technique initially developed within data mining and pattern recognition contexts, later finding applications in various domains such as medical diagnosis, insurance, loan screening, and optical character recognition of handwritten text. Breiman et al. [85] formalized this technique, introducing the CART algorithm (Classification And Regression Trees) with a comprehensive implementation of decision trees. The fundamental principle involves extending a simple cut-based analysis into a multivariate technique by continuing to analyze events that fail a particular criterion. Many events lack all characteristics of either signal or background in a two-class problem, prompting decision trees to not immediately reject events failing a criterion, but instead checking whether other criteria may assist in proper classification. While decision trees can handle multiple output classes, this chapter primarily focuses on binary trees with two possible classes: signal and background, although the concepts generalize to non-binary trees with multiple outputs.

4.1.1 Algorithm

Mathematically, decision trees are represented as rooted binary trees, focusing solely on trees with two classes: signal and background. Beginning at the initial node, or root node, the tree recursively splits into two branches until a stopping condition is met. The various stages of this process, interchangeably termed growing, training, building, or learning, are delineated in subsequent sections.

Consider a dataset comprising signal (si) and background (bj) events, each characterized by weights wsi and wbj, respectively, and described by a set of variables x. This dataset serves as the root node for a new decision tree. The algorithm proceeds as follows from this root node:

- × If the node satisfies any termination criterion, it is declared terminal, or a leaf, and the algorithm halts.
- × Events are sorted based on each variable in x.
- × For each variable, the algorithm identifies the splitting value that maximizes the separation between two branches—one dominated by signal events and the other by background events. If no further improvement in separation is achievable through splitting, the node is converted into a leaf, and the algorithm terminates.

× The variable and splitting value that yield the optimal separation are selected, and the node is split into two new branches: one containing events failing the criterion and the other containing events satisfying it.

× Recursively, the algorithm returns to step 1 for each newly created node.

This algorithm is greedy and does not guarantee the discovery of the optimal solution. At each node, all variables are considered, enabling the identification of intervals of interest within a particular variable rather than restricting each variable to a single use.

It's worth noting that decision trees are interpretable by humans, allowing for the tracing of the criteria satisfied by an event to reach a specific leaf. This interpretability facilitates understanding the tree in terms of, for instance, physics, enabling the definition of selection rules beyond mere mathematical abstraction.

To illustrate the entire process, consider the tree depicted in Fig. 37.

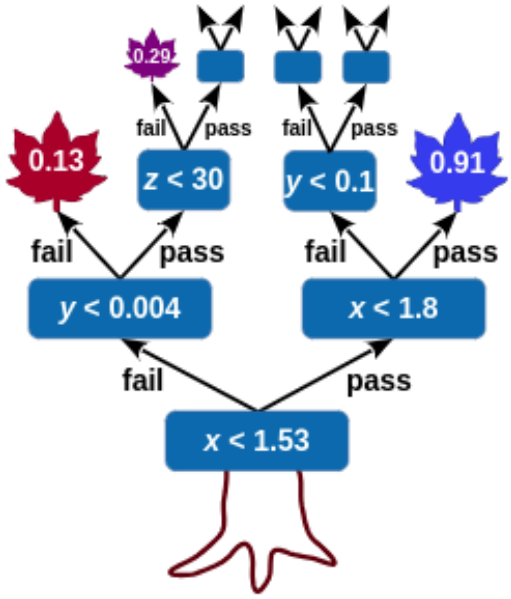


Figure 37: Graphical representation of a decision tree. Blue rectangles are internal nodes with their associated splitting criterion; leaves are terminal nodes with their purity. [86]

4.1.2 Hyperparameters

The number of hyperparameters for a decision tree is relatively limited. The first concerns the normalization of signal and background before training, typically achieved by equalizing the sums of event weights for signal and background (creating balanced

classes), resulting in a root node purity of 0.5. Decision trees are not highly sensitive to this initial normalization due to subsequent splits typically leading to more balanced nodes, causing only minor inefficiency in training that marginally affects final discriminatory power.

Other hyperparameters pertain to split selection. This involves determining discriminating variables and evaluating the best separation between signal and background events.

The splitting process must eventually terminate, declaring certain nodes as terminal leaves. Conditions for termination may include:

- × A minimum leaf size, often requiring a minimum number N_{min} of training events in each node post-split to ensure statistically significant purity measurements, with a statistical uncertainty of $\sqrt{N_{\text{min}}}$. Handling weighted events, common in high-energy physics, may involve using the effective number of events $N_{\text{eff}} = \left(\sum_{i=1}^{N} w_i\right)^2 / \sum_{i=1}^{N} w_i^2$ (where N is the number of events and w_i are their weights).
- × Achieving perfect separation (all events in the node belong to one class).
- × Insufficient improvement with further splitting.
- × A maximum tree depth to limit the number of layers, either for computational efficiency or to ensure uniform tree sizes.

Finally, a terminal leaf must be assigned a class. Conventionally, a leaf is labeled as signal if its purity $p > 0.5$ and background otherwise.

The core of a decision tree algorithm lies in splitting a node into two. An impurity measure $i(t)$ for node t quantifies the extent to which the node comprises a mix of signal and background events. This measure ideally:

- × Is maximal for an equal mix of signal and background (no separation).
- × Is minimal for nodes containing only signal or only background events (perfect separation).
- × Is symmetric in signal and background purities.
- × Is strictly concave to favor purer nodes.

A figure of merit can be constructed using this impurity measure, measuring the decrease in impurity for a split S of node t into two children t_{P} (passed) and t_{F} (failed). The goal is to find the split S^* that maximizes the decrease in impurity, resulting in the smallest residual impurity and minimizing overall tree impurity.

A stopping condition can be defined based on the decrease in impurity, avoiding node splitting if the decrease falls below a predefined value. Care is required in setting this early-stopping criterion, as sometimes seemingly weak splits enable powerful subsequent splits in child nodes.

Common impurity functions include:

× The misclassification error: $1 - \max(p, 1-p)$

× The (cross) entropy: $-\sum_{i=s,b} p_i \log p_i$, with $p_b = 1 - p_s$ and $p_s = p$

× The Gini index of diversity $= p_s p_b + p_b p_s$ [87]

The Gini index is the most commonly used impurity function in decision tree implementations, typically yielding similar performance to entropy.

4.2 Boosting algorithm

The boosting algorithm has turned into a very successful way of improving the performance of any type of classifier, not only decision trees. Creating a highly effective discriminant is challenging, but it's relatively straightforward to produce simple ones, albeit more error-prone (high bias), that still perform marginally better than random guessing. These less effective discriminants are termed weak classifiers. Boosting aims to amalgamate such weak classifiers into a new, more robust one with a lower error rate (reduced bias compared to individual classifiers) and enhanced performance.

Let T_k represent a training sample containing N_k events. Each event i is associated with a weight w_{ki}, a vector of discriminating variables $\mathbf{x}_i$, and a class label $y_i = +1$ for signal or -1 for background. The pseudocode for a generic boosting algorithm is as follows:

× Initialize T_1 for k in $1..N_{\text{tree}}$

× Train classifier T_k on T_k

× Assign weight α_k to T_k

× Modify T_k into T_{k+1}

The boosted output is a function $F(T_1, \ldots, T_{N_{\text{tree}}})$, typically a weighted average:

$$F(i) = \sum_{k=1}^{N_{\text{tree}}} \alpha_k T_k(\mathbf{x}_i)$$

This averaging renders the output quasi-continuous, alleviating one of the limitations of single decision trees. It's important to note that once a particular tree is trained, it remains unmodified and is simply added to the ensemble. This differs from approaches like neural networks, where weights are repeatedly updated over epochs to converge towards the final classifier.

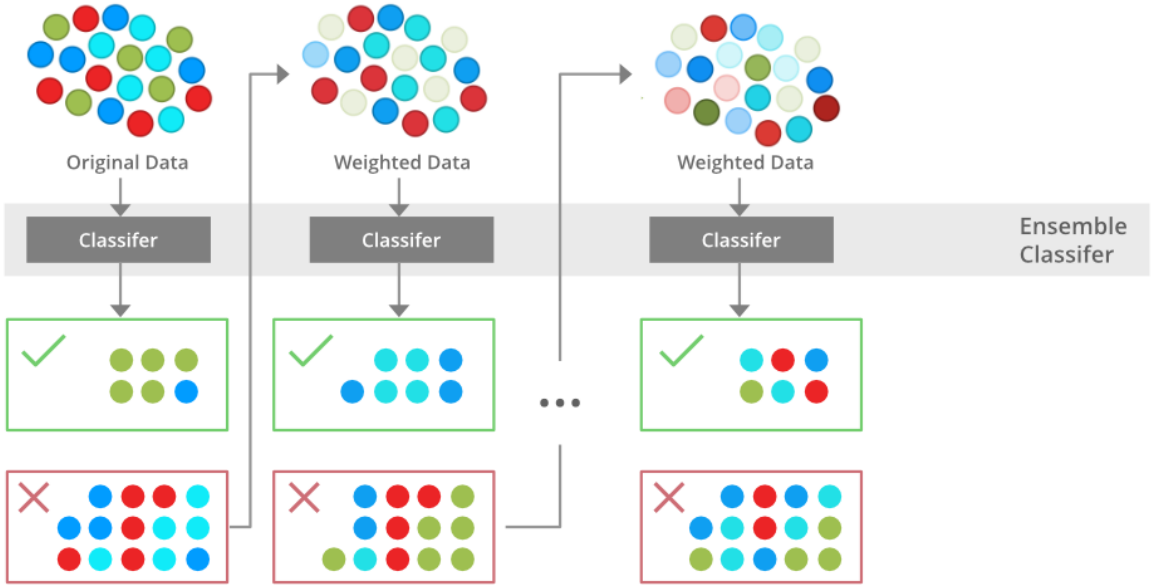


Figure 38: Example of the boosting algorithm [88]

4.3 Gradient boosting

Other boosting algorithms were initially conceptualized within the statistical framework of arcing algorithms (adaptive reweighting and combining) [89],[90]. At each step, a weighted minimization is conducted, followed by classifier re-computation and weighted input adjustment. This framework evolved into gradient boosting [91] . Boosting is viewed as a numerical optimization problem, aiming to minimize the loss function by iteratively adding trees using a gradient descent approach instead of assigning higher weights to misclassified events.

Formally, consider a model F constructed iteratively, with its imperfect instance at step k denoted as F_k. F_k approximates the best possible model (in some cases $F_k(x) \neq y$), which is to be refined at the next iteration. This refinement involves adding a new component h_k such that:

$$F_{k+1}(x) = F_k(x) + h_k(x) = y$$

or equivalently,

$$h_k(x) = y - F_k(x)$$

Instead of training F_{k+1} anew, a new classifier can be trained to fit the residual $y - F_k(x)$, representing the part that the current model F_k cannot handle correctly. If $F_{k+1}(x)$ remains unsatisfactory, further iterations can be fitted.

The connection with gradient descent becomes evident when considering the specific case of the mean squared error (MSE) loss function (commonly used for regression problems):

$$L_{\text{MSE}}(x, y) = \frac{1}{2}(y - F_k(x))^2$$

Minimizing the loss $J = \sum_i L_{\text{MSE}}(x_i, y_i)$ by adjusting all $F_k(x_i)$ yields:

$$\frac{\partial J}{\partial F_k(x_i)} = \frac{\partial L_{\text{MSE}}(x_i, y_i)}{\partial F_k(x_i)} = F_k(x_i) - y_i$$

Thus, residuals can be interpreted as negative gradients: $h_k(x_i) = y_i - F_k(x_i) = -\frac{\partial J}{\partial F_k(x_i)}$. This concept extends to any differentiable loss function, not just MSE.

5 Lepton MVA identification

For this thesis the aim was to develop an identification technique in order to enhance the performance of the existing lepton-ID. For this reason I used the TOPLeptonMVA algorithm developed by the Ghent University and used in the $t\bar{t}t\bar{t}$ analysis[92]. The training of the leptonMVA ID was initially done using the gradient boosted decision trees (GBDT) available within the TMVA package [93]. In the new version of the ID training which is used in this thesis, the extreme GBDT methods are used, as provided by XGBoost [94] . The latter generally comes with a better performance and is much faster in training and optimizing the hyperparameters, which is done with a cross-validation approach. The optimization of the considered hyperparameter space includes the number of boosting rounds (n estimators = 2000), the maximum depth of a tree (max depth = 4), the minimum sum of instance weight on a leaf node to perform a partition (min child weight = 500), and the minimum loss reduction required to make a further partition on a leaf node (gamma = 5). In order to correct for the possible imbalance of the data set, the scale pos weight parameter is set to the ratio of negative to positive instances. In order to avoid the overtraining effects, the early stopping is used if no further improvement is observed in the loss function after ten boosting rounds. Other parameters are set to their default values.

5.1 Input features

A distinction between prompt and nonprompt leptons can be done by using input features exhibiting significant differences between these types of leptons. A set of common features for electrons and muons includes variables that are listed in Table 8.

Nonprompt leptons are typically soft in terms of their momentum, while prompt leptons from top quark decays have a significantly harder p_T distribution. Therefore, the basic lepton kinematic variables, p_T and η, can be used to distinguish between the prompt and nonprompt leptons. As described hereafter, several of the input features, such as the lepton relative isolation, ratio, also depend on the lepton p_T.

The relative isolation of a lepton is the scalar p_T sum of PF objects in $\Delta R = 0.3$ around the lepton, divided by the lepton p_T:

$$I_{\text{rel}} = \frac{\sum p_T^{\text{ch. hadr.}} + \max(0, \sum p_T^{\text{neu. hadr.}} + \sum p_T^{\text{pho.}} - \rho A_{\text{eff}})}{{p_T}^l} \tag{89}$$

where $\sum p_T^{\text{ch. hadr.}}$ is the scalar sum of p_T of charged hadrons originating from the primary vertex, while $\sum p_T^{\text{neu. hadr.}}$ and $\sum p_T^{\text{pho.}}$ correspond to the contributions from neutral hadrons and photons. The "jet area" method is used to mitigate the contribution

from pileup, with ρ being the average transverse-momentum flow density and A_{eff} the effective area defined as the geometric area of the isolation cone multiplied by the η-dependent correction factor that accounts for the residual dependence of the isolation on pileup. In case of a muon, the I_{rel} definition uses the $(\pi(\Delta R)^2)$ pileup correction defined as $(\sum p_T^{\text{pu ch. hadr.}})/2$, where the sum is performed over the charged PF candidates not associated with the primary vertex, and the factor 0.5 corresponds to a naive average of neutral to charged particles.

The mini-isolation variable (I_{mini}) introduces the p_T-dependent cone size in the relative isolation definition that reduces the impact of a possible inclusion of hadronic activity present in the event that is not directly connected with the studied lepton. The cone size varies between 0.2 and 0.05, when moving from low to high p_T. Two variables are defined that correspond to the sum over the charged ($I_{\text{mini}}^{\text{ch}}$) and neutral ($I_{\text{mini}}^{\text{neu}}$) PF candidates.

An important feature of the nonprompt lepton is the possible presence of a nearby hadronic jet that can be associated with a hadron decay. A lepton is considered to be found within a jet if it shares a PF candidate that is part of the jet. The reconstructed lepton-jet variables use the jet four momentum with energy corrections that are only applied to its hadronic component after subtracting the leptonic contribution from the L1 pileup-corrected four momentum of the jet. The lepton is then added back after the L2 and L3 jet response corrections are applied to the lepton-aware jet. This procedure allows to properly include the p_T- and η-dependent jet energy corrections that are initially derived in a sample with inclusive jets.

The number of selected tracks in the nearby jet (N_{trk}) is used as one of the input features. The p_{ratio}^T variable is defined as the ratio between the lepton and jet p_T. If no nearby jet is found, the relative isolation is used instead as $1/(1+I_{\text{rel}})$. This variable carries similar information as the relative isolation of the lepton. The p_{rel}^T feature represents the projection of the lepton p_T on the transverse plane relative to the jet axis. The contribution from the lepton is subtracted from the jet's four momentum. This variable is particularly sensitive to discriminating leptons originating from B hadron decays, due to the large b-quark mass with respect to charm and other lighter quarks. Finally, the b tagging DeepJet discriminant of the nearby jet (B_{tag}) provides an important handle on distinguishing between different hadron flavour origins associated with nonprompt leptons.

The track impact parameters provide invaluable information on the compatibility of the reconstructed lepton to be originated from the primary vertex. The impact parameter is defined as the distance from the point of closest approach of the track to the primary vertex. This distance is computed in the transverse ($|d_{xy}|$) and longitudinal ($|d_z|$) planes with respect to the beam line. Nonprompt leptons associated with displaced decays of

heavy hadrons tend to have large impact parameter values compared to prompt leptons, which have these parameter values consistent with the typical beam spread. The impact parameter significance of a lepton (SIP_{3D}) is defined as the ratio between the impact parameter of the associated track and its uncertainty, calculated in 3D space.

The electron MVA ID is a Boosted Decision Tree (BDT) discriminant (ID MVA) that is centrally provided by the EGamma Physics Object Group (POG) . This discriminant combines track information, supercluster observables, and track-cluster matching variables, allowing to decide whether a GSF track is matched to an ECAL cluster. The Fall17 version of the BDT training that does not include the PF isolation as input variable is used. The training sample contains Drell-Yan (DY) events.

A segment-based compatibility of a muon (P_{seg}) is computed from propagation of a tracker track to the muon system. This variable defines the probability that a given track corresponds to a muon.

Feature	**Description**
Kinematics	
p_T	Transverse momentum of a lepton
$\lvert\eta\rvert$	Pseudorapidity of a lepton
Isolation	
I^{ch}_{rel}	Relative isolation using the cone size of 0.4
I^{ch}_{mini}	Relative mini-isolation with p_T-dependent cone size including charged PF objects
I^{neu}_{mini}	Relative mini-isolation with p_T-dependent cone size including neutral PF objects
Properties of the closest jet	
N_{trk}	Number of charged particles associated with the jet
p_{T}^{rel}	Fraction of the lepton momentum in the transverse direction to the jet axis
p_{T}^{ratio}	Ratio between the lepton and jet transverse momenta
B_{tag}	The DeepJet b-tagging discriminator
Impact parameter	
SIP_{3D}	3D impact parameter significance
$\log(\lvert d_{xy}\rvert)$	Transverse impact parameter with respect to the primary vertex
$\log(\lvert d_z\rvert)$	Longitudinal impact parameter with respect to the primary vertex
Other	
ID MVA^e	POG electron MVA ID discriminant (Fall17v2noIso)
P^{μ}_{seg}	Compatibility of track segments in the muon system with the expected pattern of a minimum ionizing particle

Table 8: Description of the input features used in the lepton MVA.

The shape comparison between different lepton origins in the distributions of the described variables is presented in the Figures 40 and 42 for electrons and muons respectively. The sample used for the input features is an inclusive $t\bar{t}$ sample after a small preselection. Electrons must be particle-flow (PF) electrons associated with a track reconstructed by the Gaussian Sum Filter (GSF) algorithm [95] , which uses a weighted sum of Gaussian distributions in the track fitting, in contrast to the Kalman Filter [67]. An electron is selected if it has $p_T > 10$ GeV and $|\eta| < 2.5$. Additionally, an electron must

satisfy $I_{\text{mini}} < 0.4$, not more than one missing hit in the inner tracking detector (N_{miss}), $\text{SIP}_{3D} < 8$, $|d_{xy}| < 0.05$, and $|d_z| < 0.1$. Muons are reconstructed with the PF algorithm and have an associated track identified in both tracking and muon systems, with additional track-quality and muon-quality requirements. Muons must have $p_T > 10$ GeV and $|\eta| < 2.4$, and satisfy the same selection requirement that is applied to electrons, except for the N_{miss} requirement.

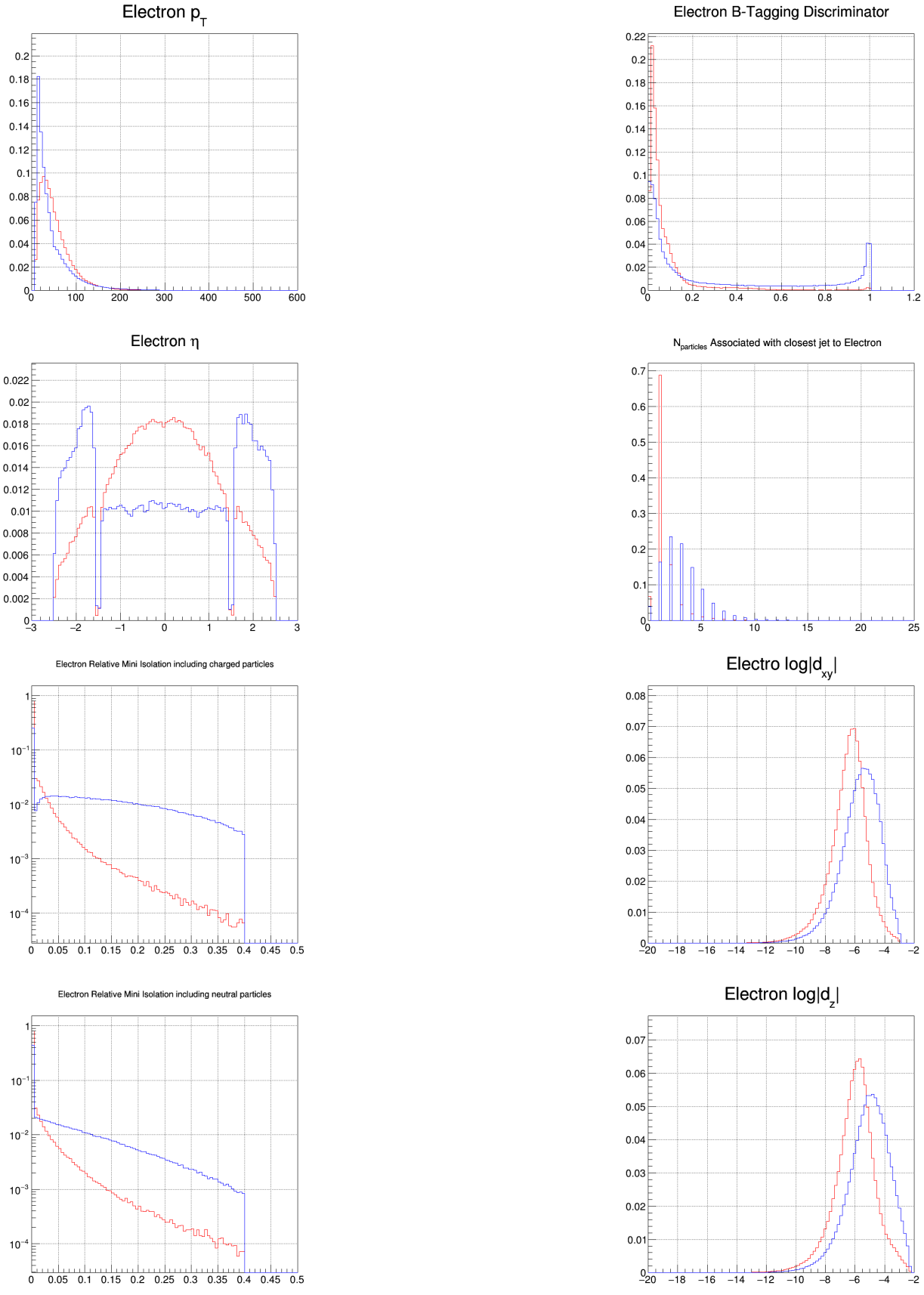


Figure 39: Shape comparison between prompt and nonprompt electron input features. Red corresponds to the prompt and blue to the non-prompt leptons.

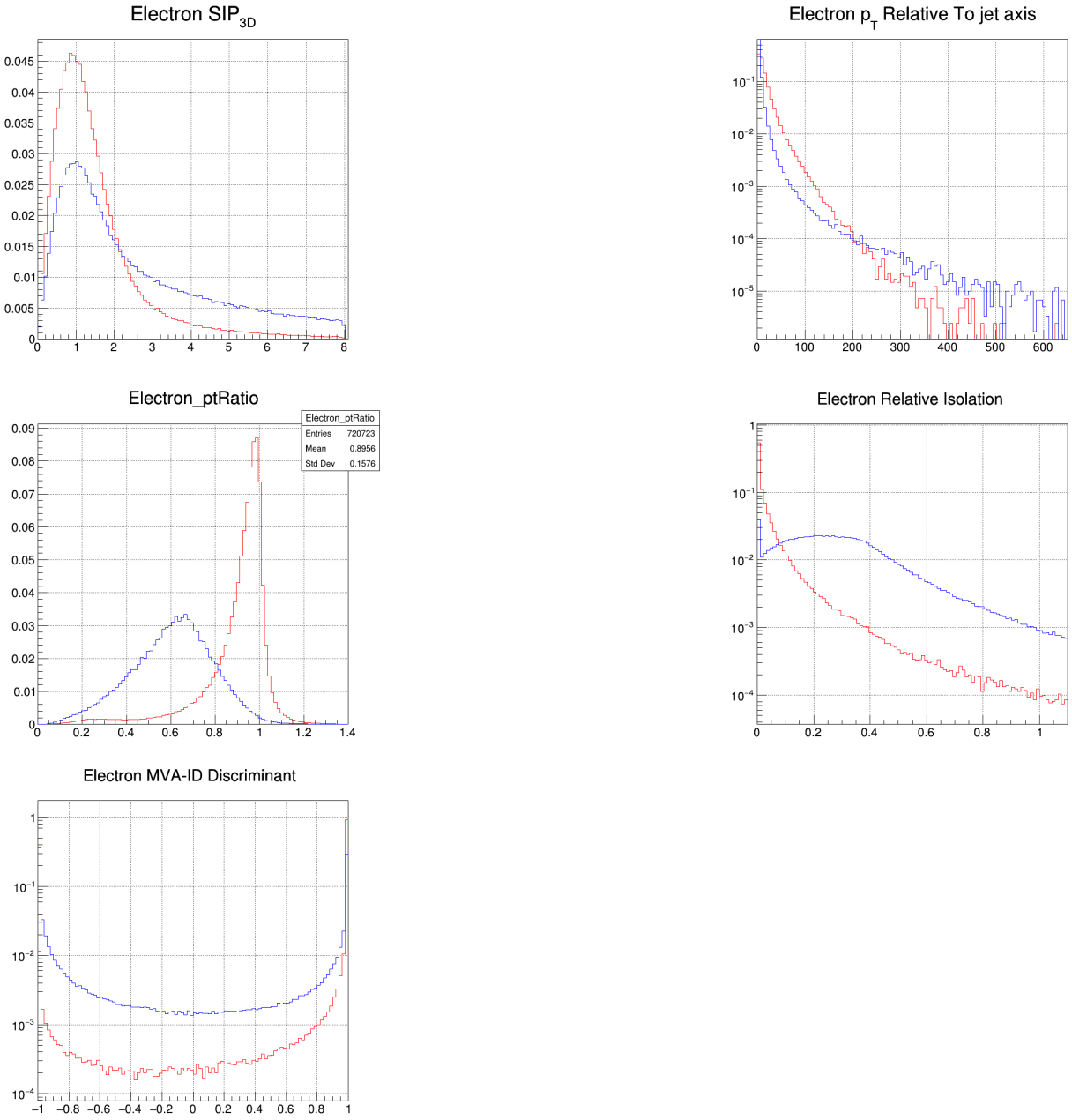


Figure 40: Shape comparison between prompt and nonprompt electron input features. Red corresponds to the prompt and blue to the non-prompt leptons.

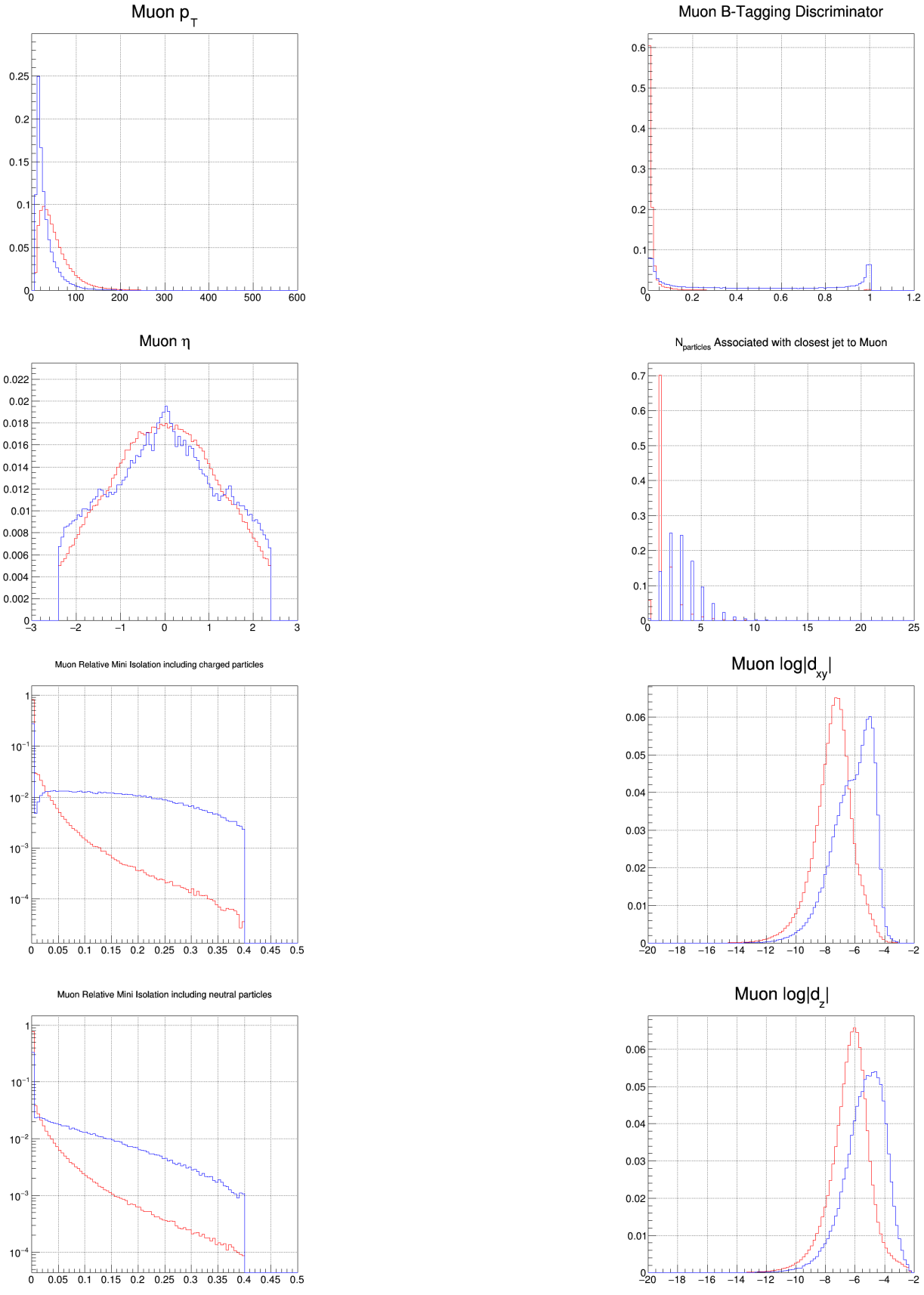


Figure 41: Shape comparison between prompt and nonprompt muon input features. Red corresponds to the prompt and blue to the non-prompt leptons.

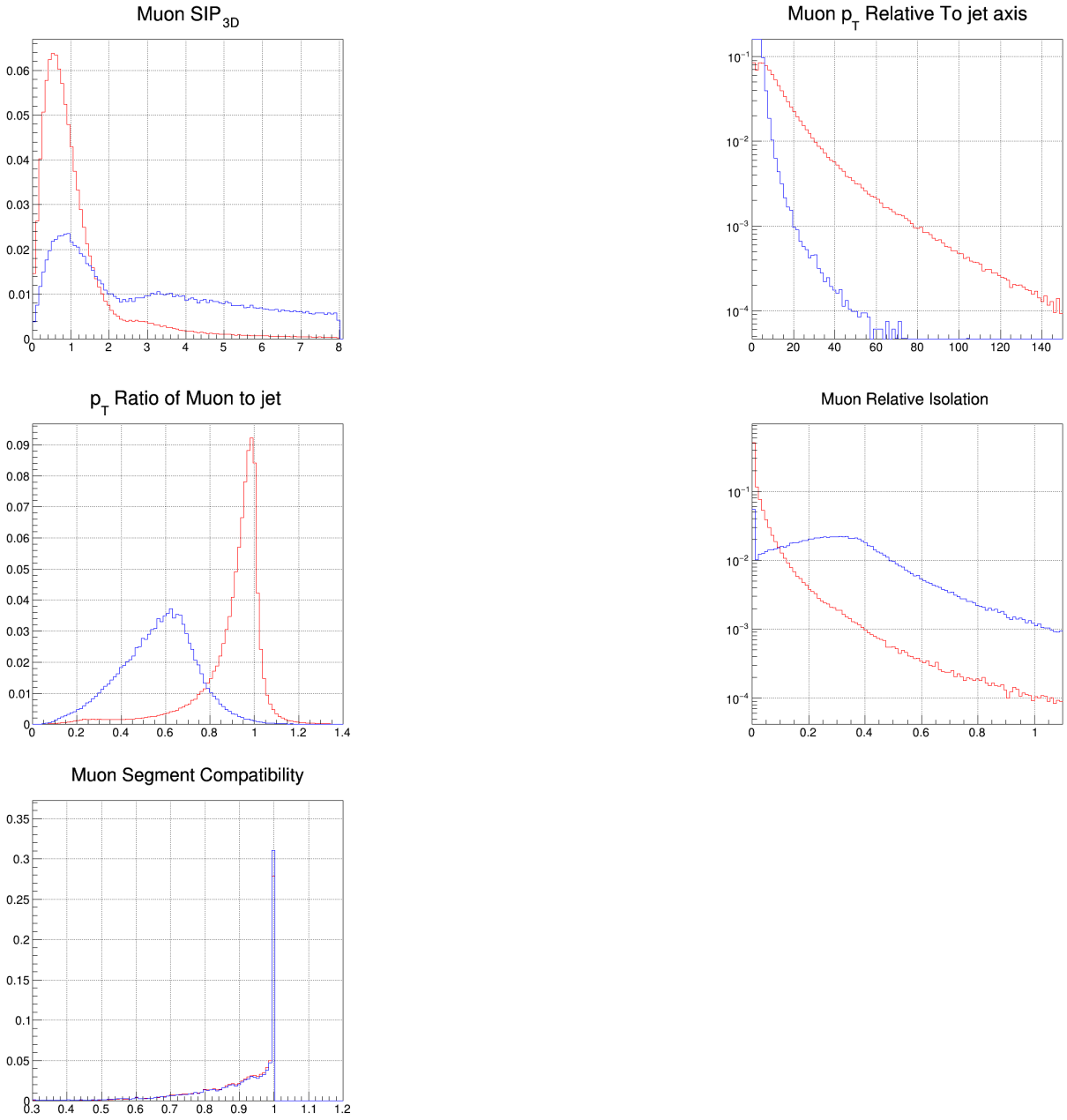


Figure 42: Shape comparison between prompt and nonprompt muon input features. Red corresponds to the prompt and blue to the non-prompt leptons.

5.2 ROC Curves

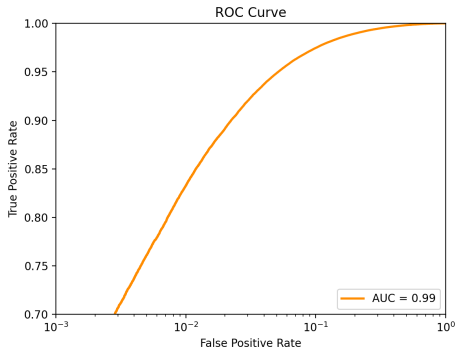
Receiver Output Curves (ROC) curves are commonly employed to assess the performance of a classifier. These curves enable a straightforward comparison between different classifiers by plotting the true positive rate against the false positive rate. The true positive rate (TPR) is defined as the ratio of correctly identified positive samples to the total number of positive samples, expressed as:

$$TPR = \frac{N_{\text{true positive}}}{N_{\text{true positive}} + N_{\text{false negative}}} \tag{90}$$

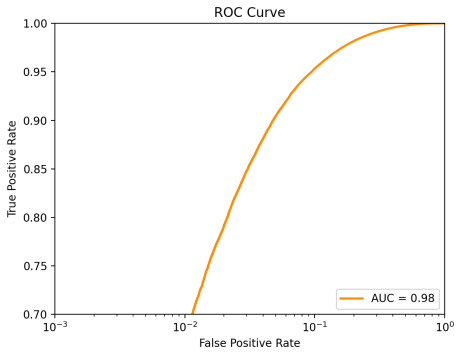
Similarly, the false positive rate (FPR) is defined as the ratio of incorrectly identified negative samples to the total number of negative samples, given by:

$$FPR = \frac{N_{\text{false positive}}}{N_{\text{false positive}} + N_{\text{true negative}}} \tag{91}$$

The ROC curves were made for the same $t\bar{t}$ sample as before and the results can be seen in the figures 43a and 43b.



(a) ROC curve for electrons



(b) ROC curve for muons

Figure 43: ROC Curves for LeptonMVA

These ROC curves also allow for the selection of the working point. This means, that it is possible to choose a desired false positive rate and receive the corresponding true positive rate. The true positive rate is then referred to as signal efficiency, the amount of signal leptons that are correctly classified as signal. Whereas the false positive rate is referred to as background efficiency, the amount of background leptons that are falsely classified as signal. Ideally, the signal efficiency is close to one, whereas the background efficiency stays close to zero.

5.3 Comparison with the existing ID

Before the implementation of the LeptonMVA algorithm in the analysis, it is imperative to conduct a comparative assessment between the existing identification (ID) method and the proposed ID based on the Multivariate Analysis (MVA) approach. To this end, a comprehensive evaluation was performed using a $t\bar{t}$ sample (Table 9) for both dilepton

(DL) and single-lepton (SL) channels. The evaluation involved the calculation of efficiencies and purities for leptons selected using the conventional ID methods described in sections 3.3.1 and 3.3.2, as well as those selected based on the LeptonMVA classifier.

Specifically, the LeptonMVA-based ID is delineated in Table 10 same for both channels. Notably, the inclusion of the Particle Flow (PF) candidate variable in the ID construction was necessitated by a discernible decrease in efficiency observed particularly in high-p_T muons when relying solely on the classifier score.

Channel	Dataset
DL	TTTo2L2Nu-TuneCP5-13TeV-powheg-pythia8/106X-upgrade2018-realistic
SL	TTToSemiLeptonic-TuneCP5-13TeV-powheg-pythia8/RunIISummer20UL18NanoAODv9

Table 9: Channels and Datasets for efficiencies and purities

Electron ID	**Electron Score** > 0.9
Muon ID	**Muon Score** > 0.9 **and Muon is PF candidate**

Table 10: Table with Electron and Muon IDs

Also it is important to give the definition of the efficiency and the purity as metrics of how accurate an identification method is.

The efficiency (ε) is calculated using the following formula:

$$\varepsilon = \frac{\text{Number of generated prompt leptons matched to offline leptons}}{\text{Total number of generated prompt leptons}} \tag{92}$$

where:

× **Number of generated prompt leptons matched to offline leptons** represents the count of prompt leptons generated by the simulation that have a corresponding match in the reconstructed leptons, satisfying the following criteria:

  ✓ For muons:

    * Same pdgId
    * Respective ID requirements
    * $\Delta R < 0.02$

  ✓ For electrons:

    * Same pdgId

    - Respective ID requirements
    - $\Delta R < 0.02$

× **Total number of generated prompt leptons** denotes the overall count of prompt leptons generated by the simulation within $|\eta| < 2.4$.

ΔR is calculated between the η and ϕ of the recontructed and generated leptons.

The purity is calculated using the following formula:

$$P = \frac{\text{Number of offline leptons matched to gen leptons}}{\text{Total number of offline leptons}} \tag{93}$$

where:

× **Number of offline leptons matched to gen leptons** represents the count of offline leptons that have a corresponding match in the **prompt** generated leptons, meeting the following criteria:

✓ For muons:

- Same pdgId
- Respective ID requirements
- $\Delta R < 0.02$

✓ For electrons:

- Same pdgId
- Respective ID requirements
- $\Delta R < 0.02$

× **Total number of offline leptons** refers to the overall count of the offline leptons that successfully pass the ID.

5.3.1 Single-lepton channel

Figures 44-47 show the comparison of efficiencies in the SL channel.

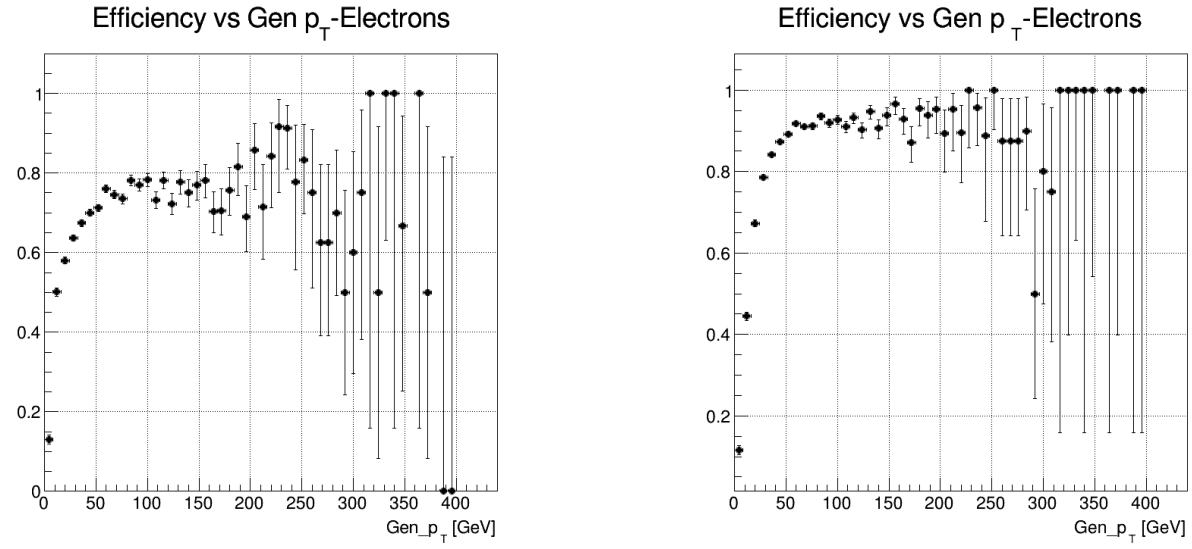


Figure 44: Efficiencies vs p_T for different IDs for electrons
Old ID:left New ID:right

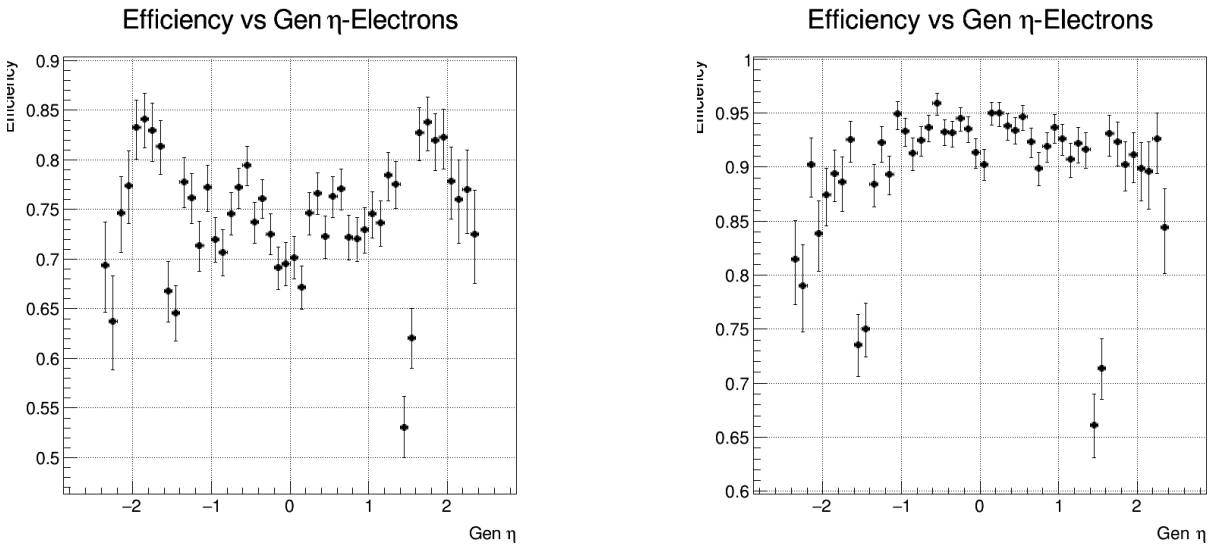


Figure 45: Efficiencies vs η for different IDs for electrons
Old ID:left New ID:right

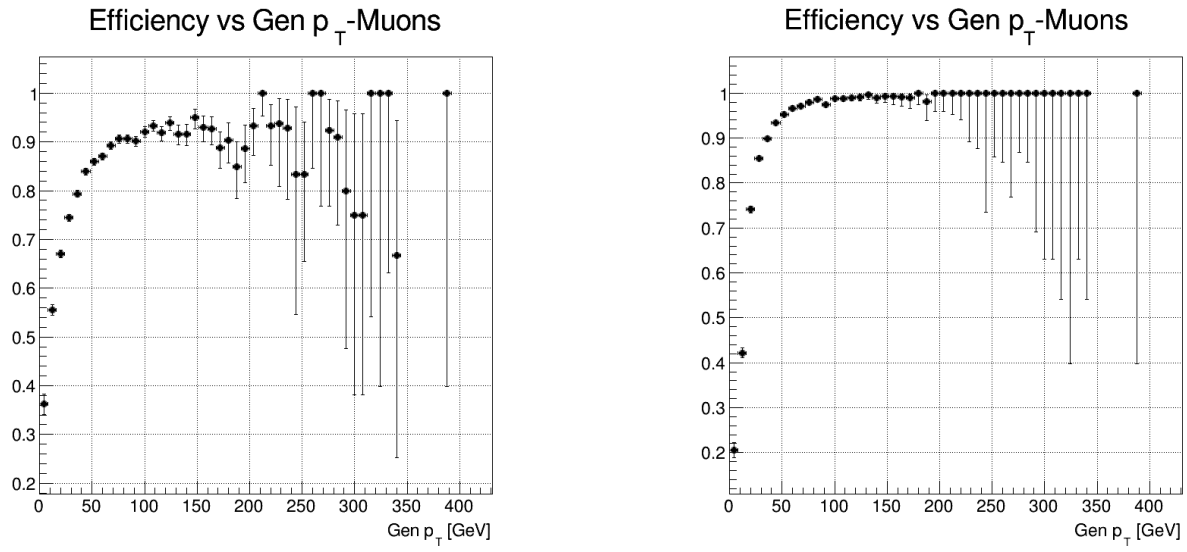


Figure 46: Efficiencies vs p_T for different IDs for muons
Old ID:left New ID:right

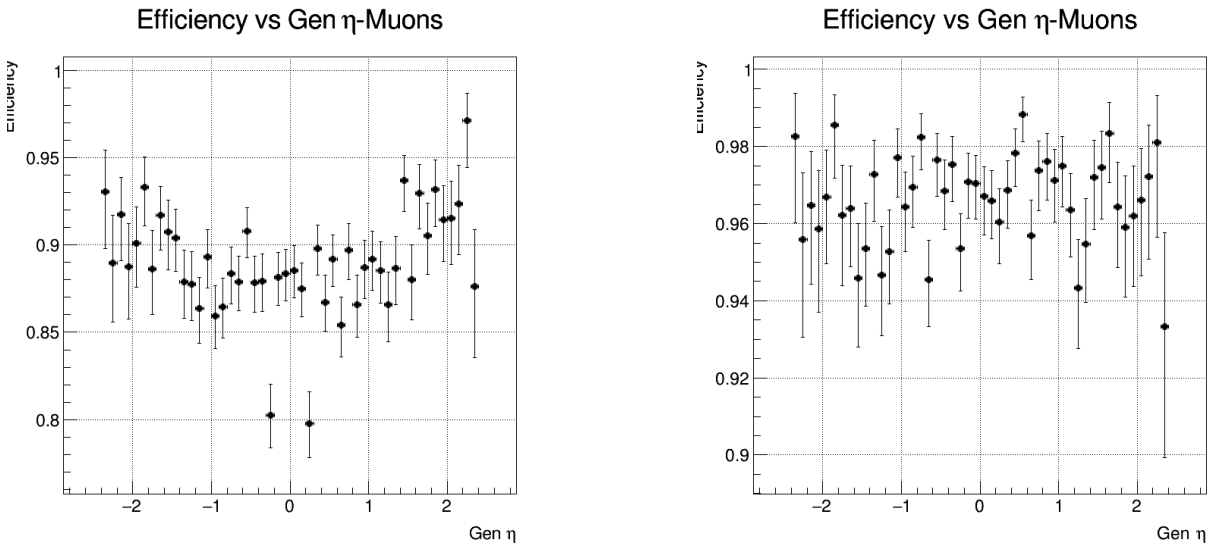


Figure 47: Efficiencies vs η for different IDs for muons
Old ID:left New ID:right

From the efficiency plots, it is clear that the new ID (the one using the LeptonMVA) is more efficient than the old one in selecting prompt leptons. The gain in electrons is bigger than the one in muons with aprroximately 20% increase in electrons and 10% in muons. Figures 48-51 show the comparison of puritires in the SL channel.

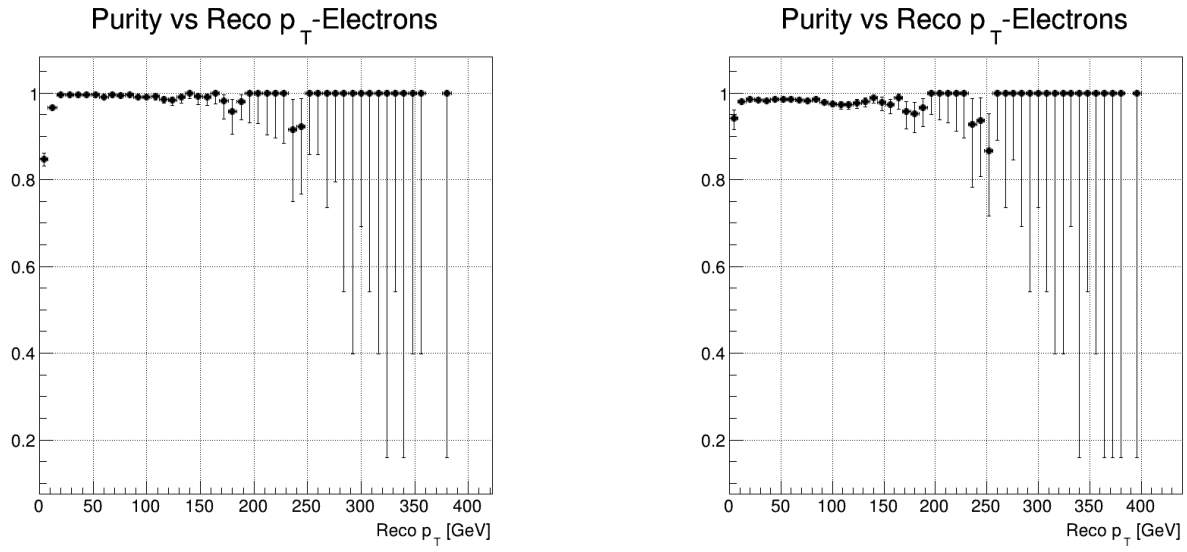


Figure 48: Purities vs p_T for different IDs for electrons
Old ID:left New ID:right

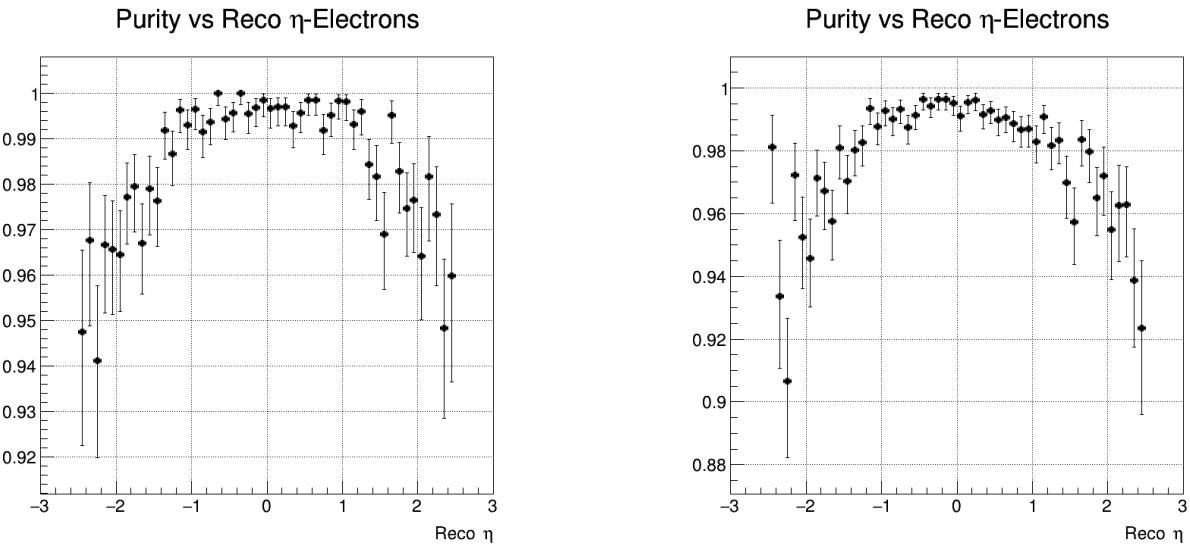


Figure 49: Purities vs η for different IDs for electrons
Old ID:left New ID:right

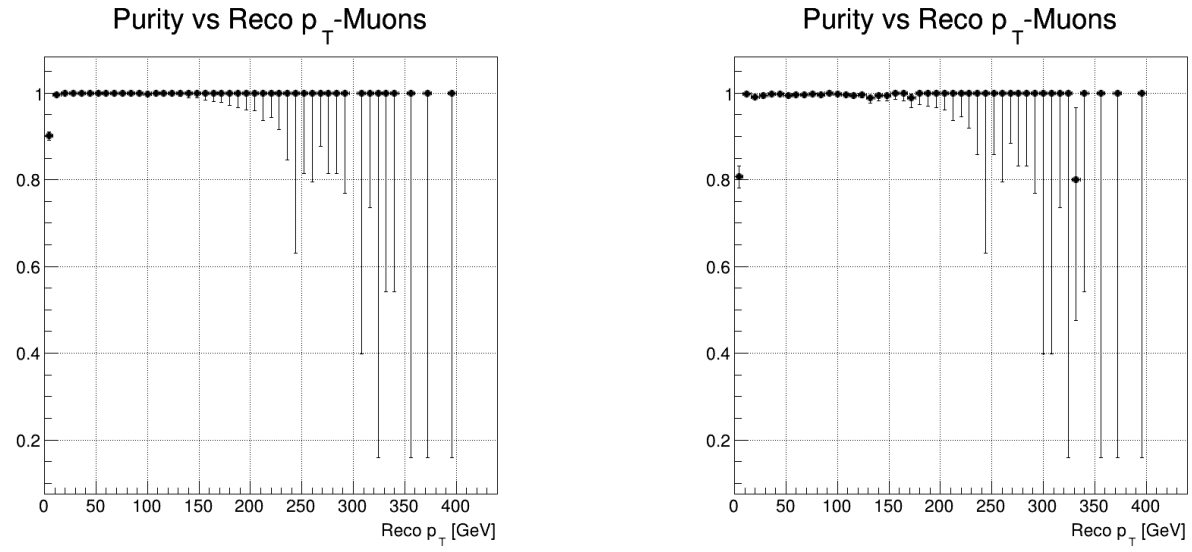


Figure 50: Purities vs p_T for different IDs for muons
Old ID:left New ID:right

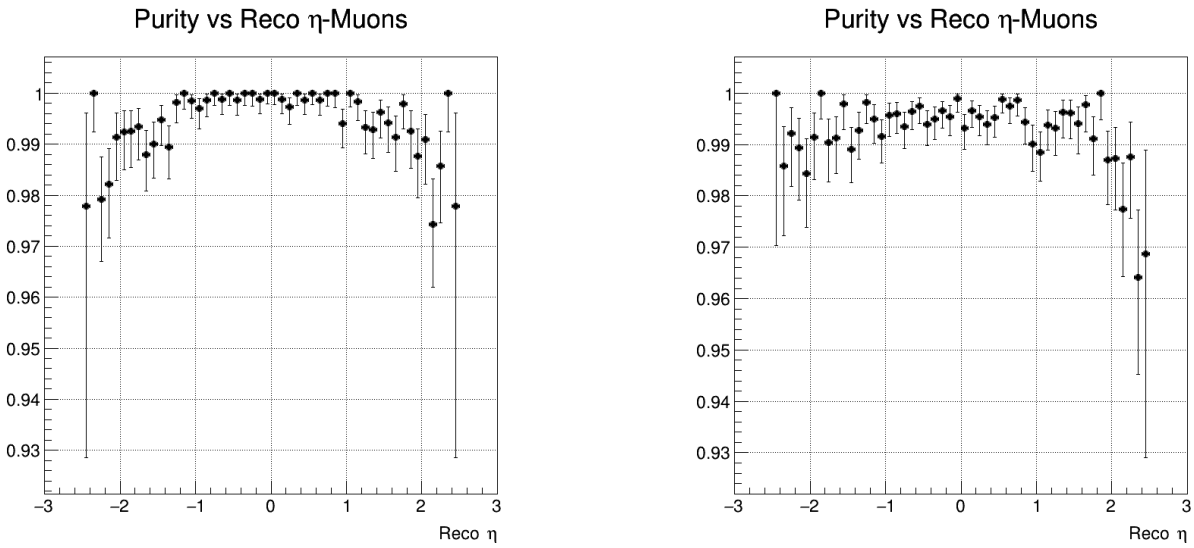


Figure 51: Purities vs *eta* for different IDs for muons
Old ID:left New ID:right

Despite the gain in efficiency, the purity still remains close to 1 in both IDs which is quite important and verifies that the new ID provides better results. This is prominent from the ratio plots (Fig. 52-55) that show the comparison between the 2 different IDs in every p_T and η range.

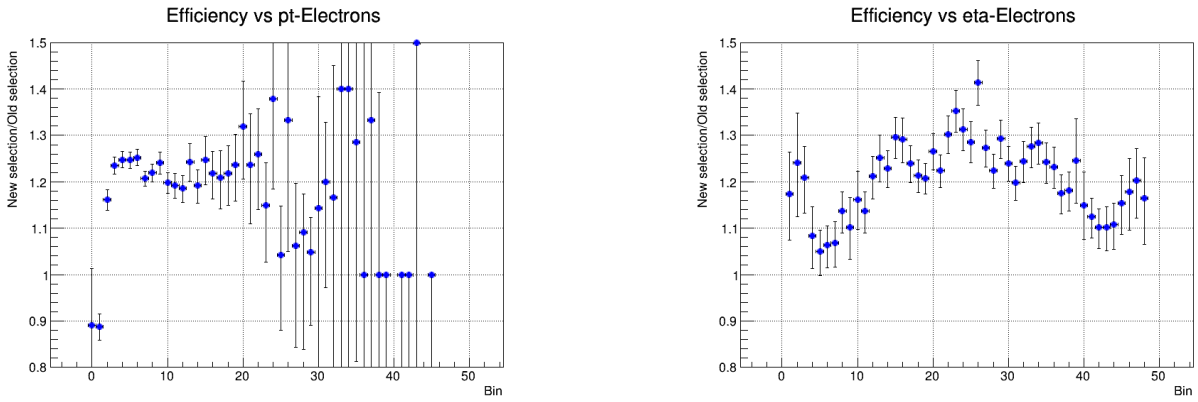


Figure 52: Ratio plots for efficiencies (Electrons)
p_T:left η:right

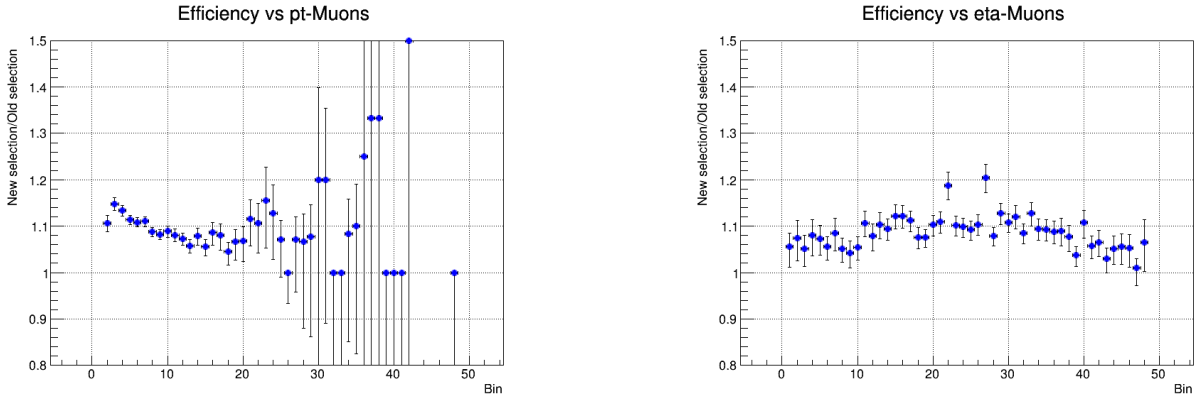


Figure 53: Ratio plots for efficiencies (Muons)
p_T:left η:right

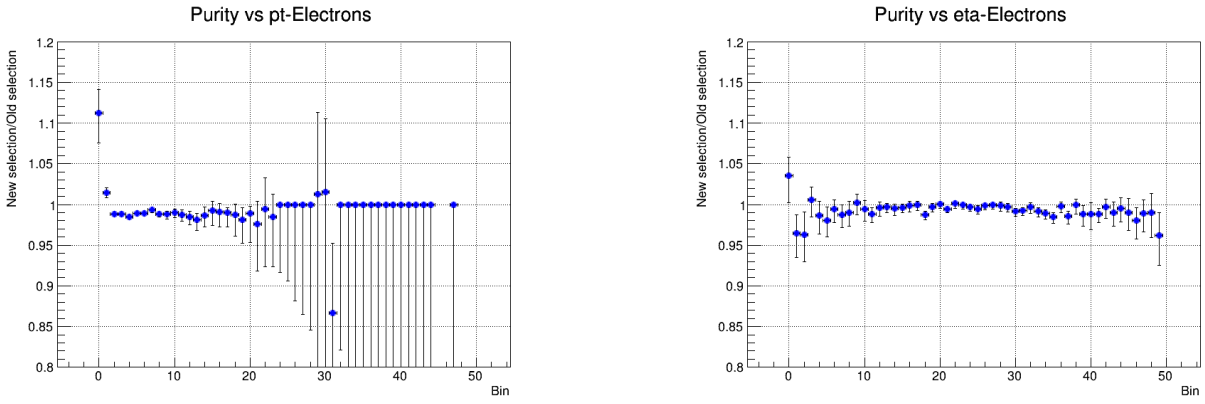


Figure 54: Ratio plots for purities (Electrons)
p_T:left η:right

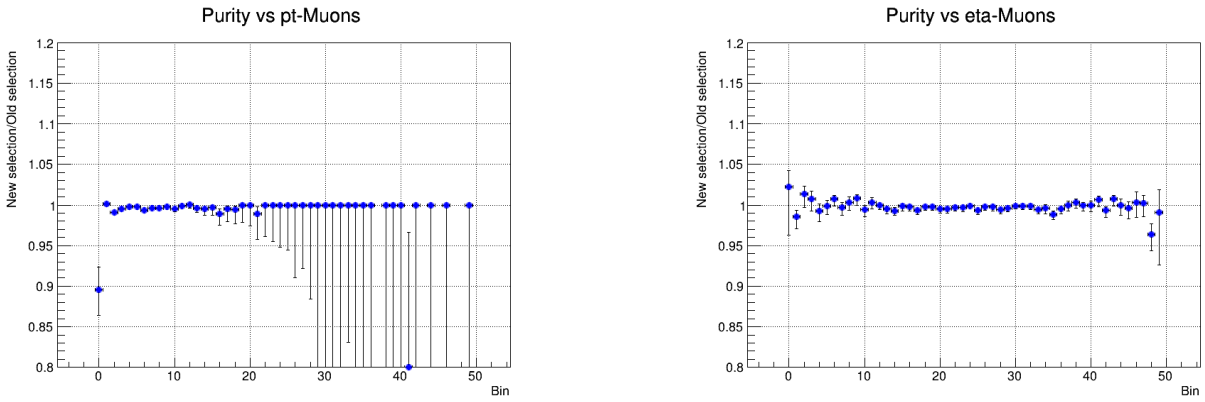


Figure 55: Ratio plots for purities (Muons)
p_T:left η:right

5.3.2 Dilepton channel

Figures 56-59 show the comparison of efficiencies in the DL channel. Just a clarification note here is that for the efficiencies and the purities both leptons are considered (leading and subleading) for the calculation of efficiencies and purities.

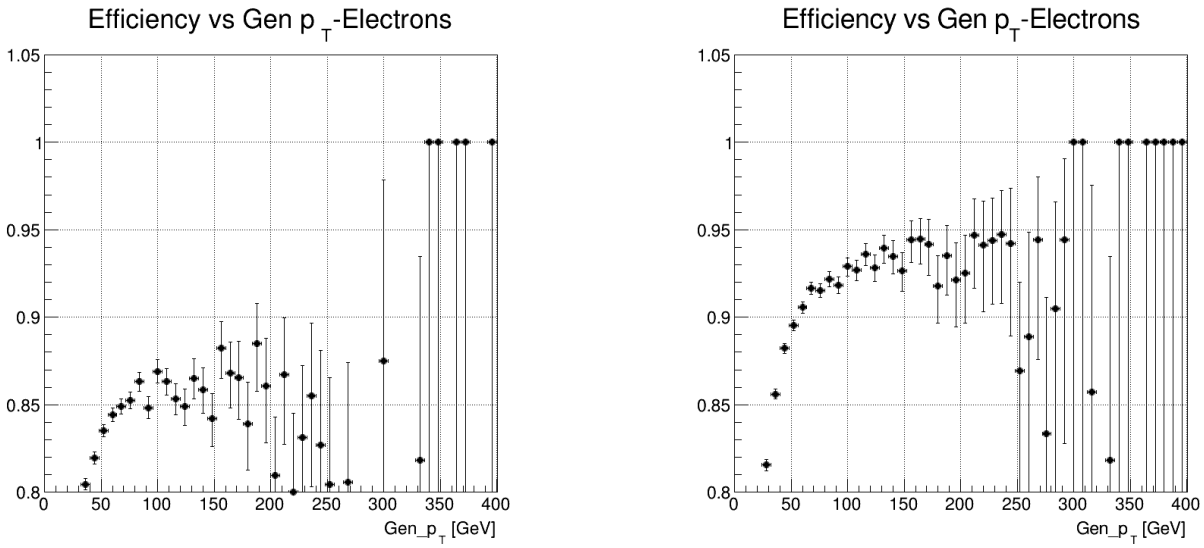


Figure 56: Efficiencies vs p_T for different IDs for electrons
Old ID:left New ID:right

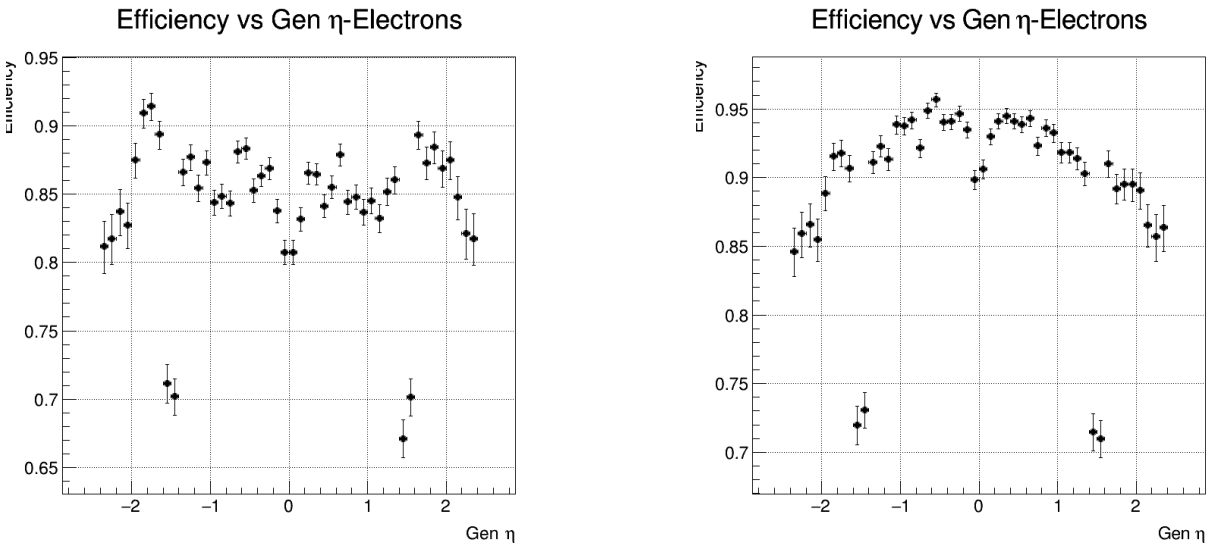


Figure 57: Efficiencies vs η for different IDs for electrons
Old ID:left New ID:right

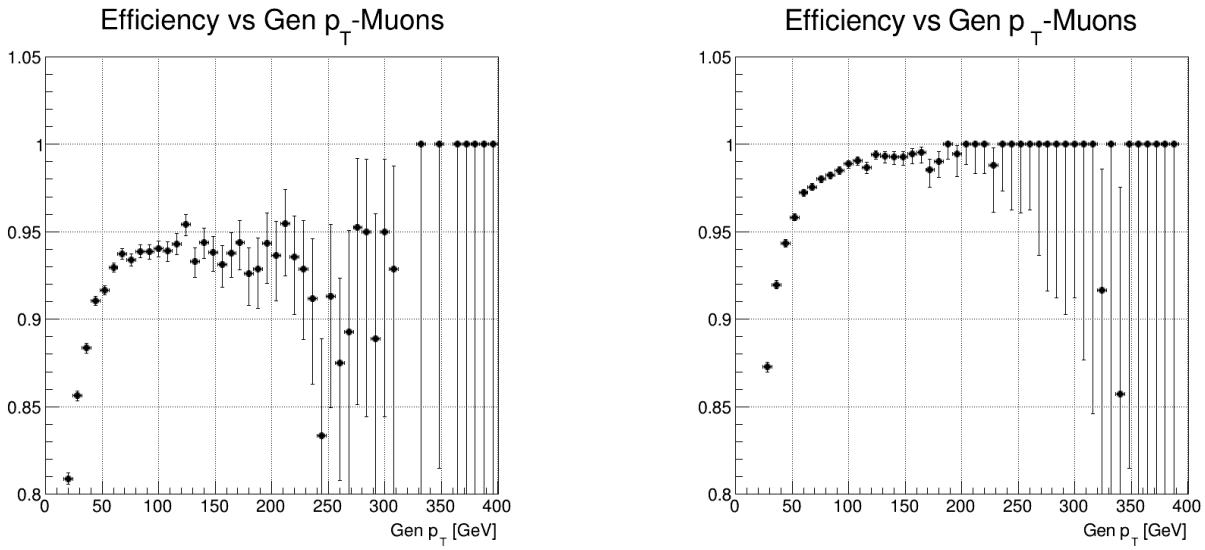


Figure 58: Efficiencies vs p_T for different IDs for muons
Old ID:left New ID:right

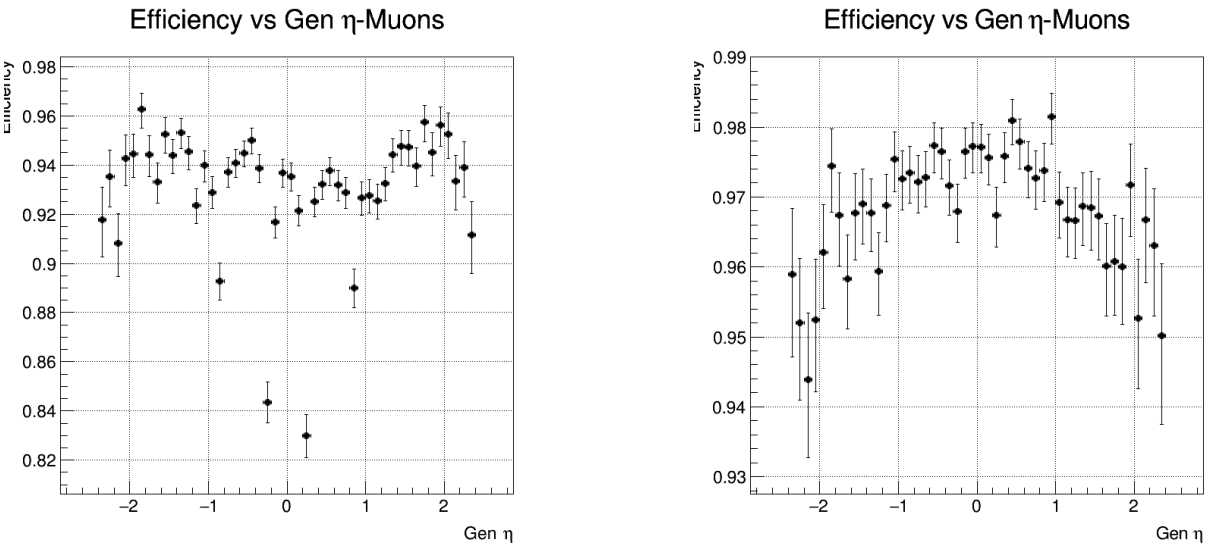


Figure 59: Efficiencies vs η for different IDs for muons
Old ID:left New ID:right

From the efficiency plots, it is seen that the new ID still provides better results for the DL channel but with a smaller gain than the SL channel because from Table 6 and Table 7, the old ID in DL is lenier than the SL. The goal here was to develop a global ID independent of the channel so it is of crucial importance to see if the purity in Figures 60-63 still remains close to 1.

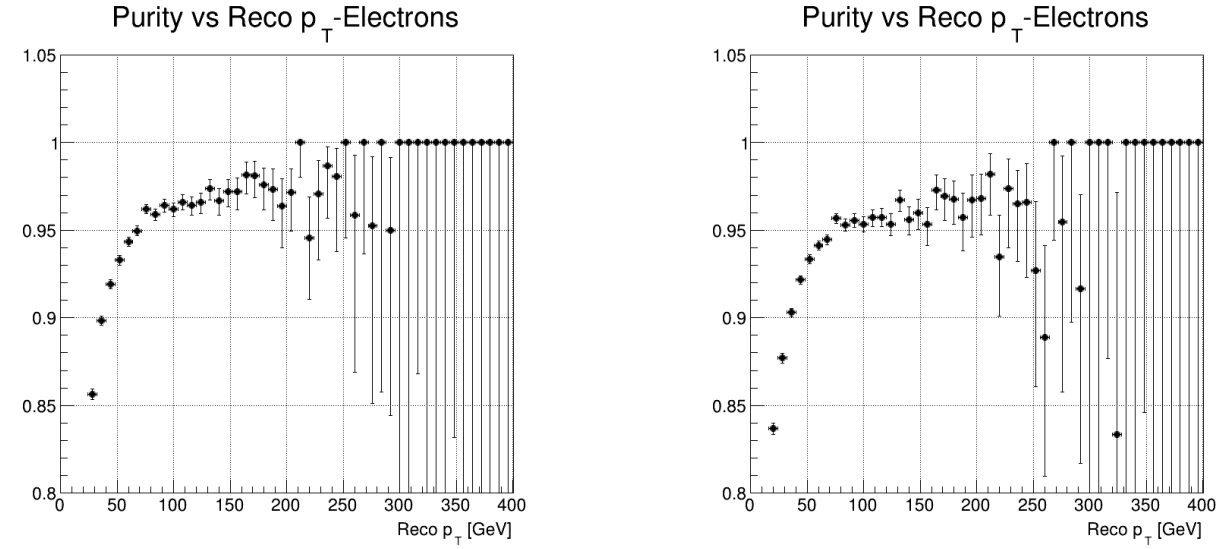


Figure 60: Purities vs p_T for different IDs for electrons
Old ID:left New ID:right

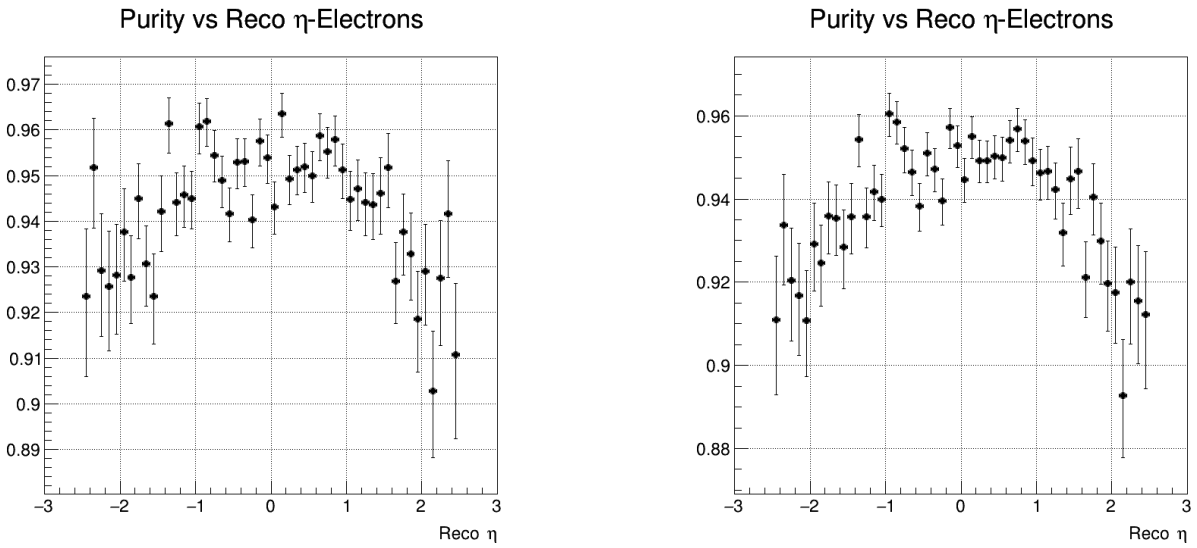


Figure 61: Purities vs *eta* for different IDs for electrons
Old ID:left New ID:right

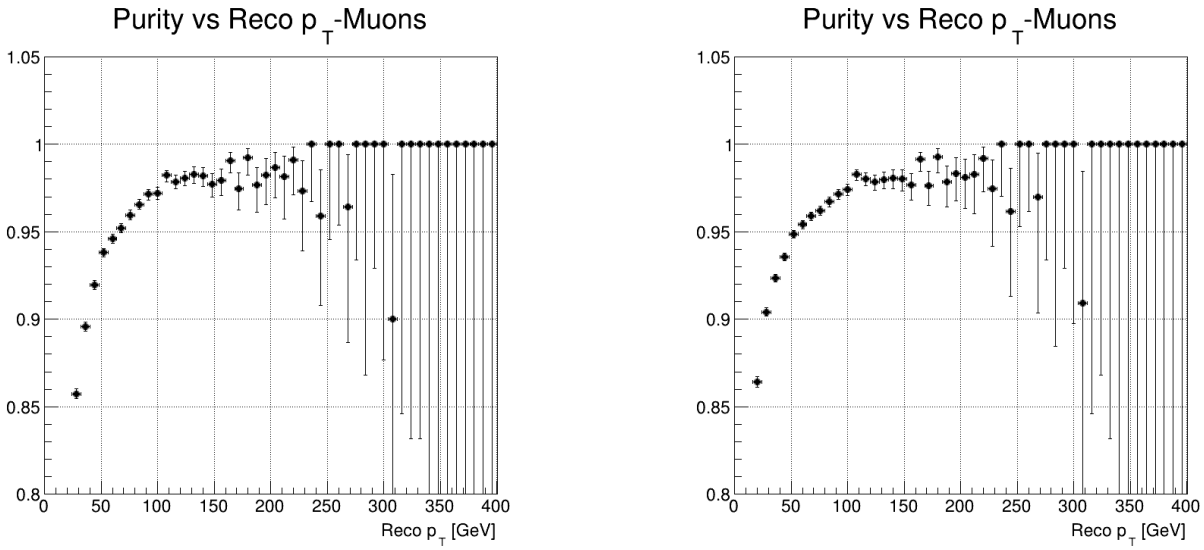


Figure 62: Purities vs p_T for different IDs for muons
Old ID:left New ID:right

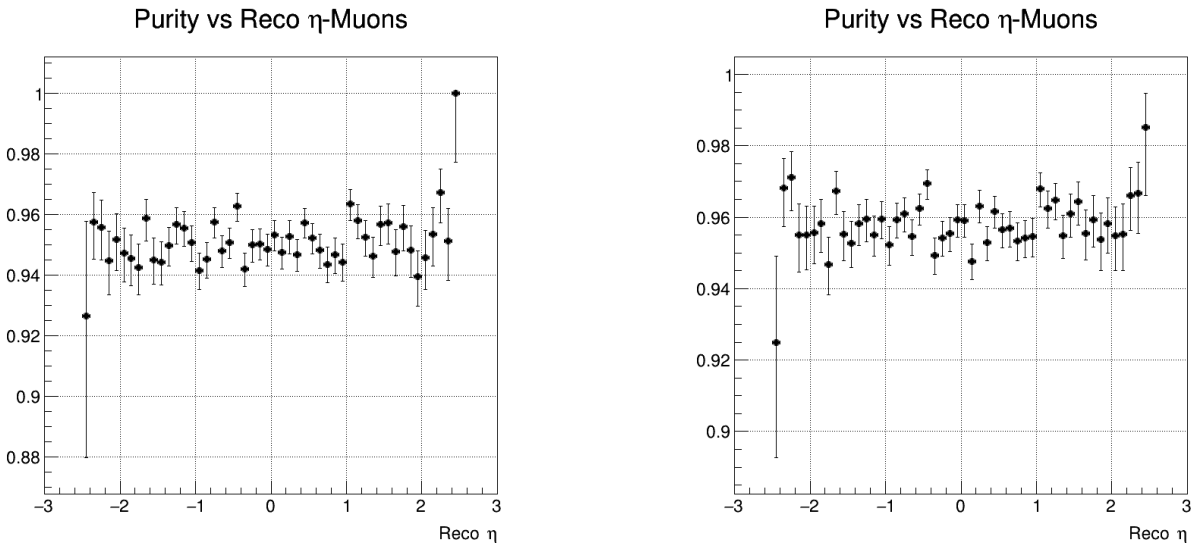


Figure 63: Purities vs *eta* for different IDs for muons
Old ID:left New ID:right

Finally, the ratio plots in Figures 64-67 illustrate the gain with the LeptonMVA ID applied in the DL channel.

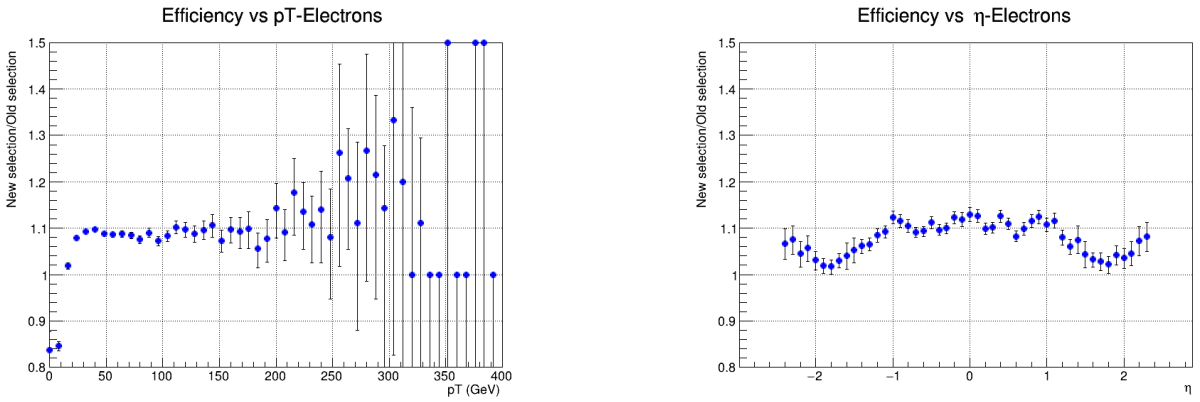


Figure 64: Ratio plots for efficiencies (Electrons)
p_T:left η:right

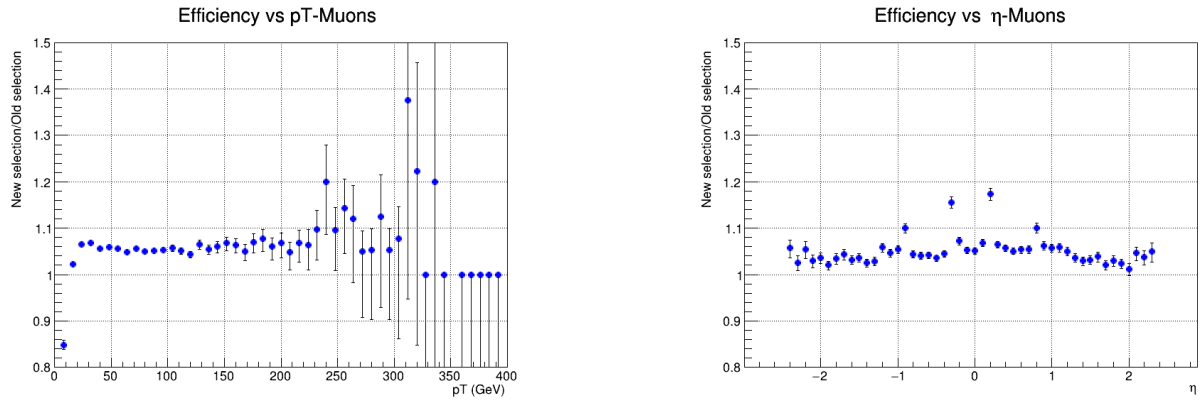


Figure 65: Ratio plots for efficiencies (Muons)
p_T:left η:right

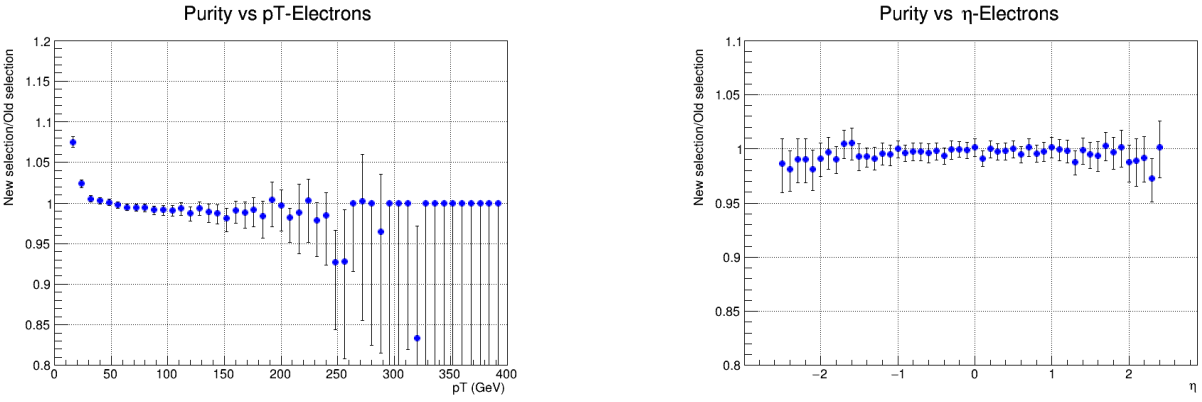


Figure 66: Ratio plots for purities (Electrons)
p_T:left η:right

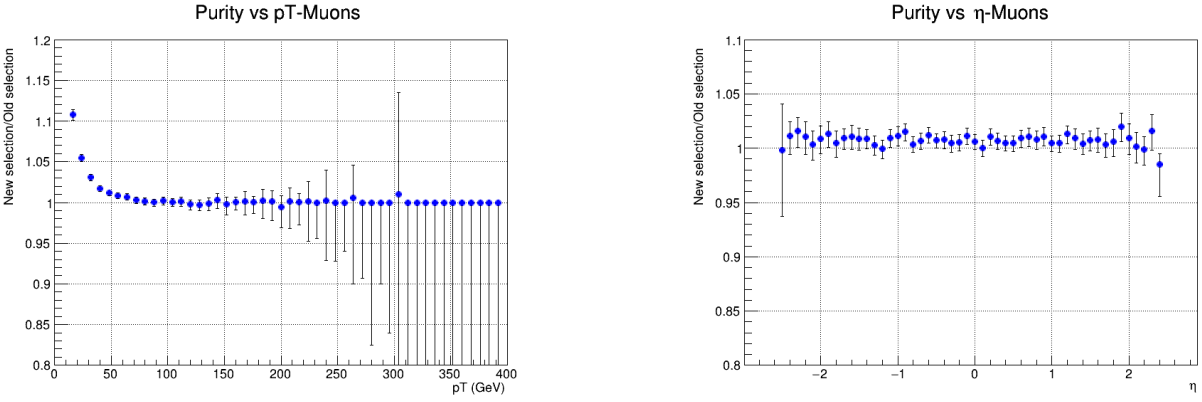


Figure 67: Ratio plots for purities (Muons)
p_T:left η:right

Based on the observations derived from the efficiency ratio plots and purity analyses,

it is evident that the proposed new identification (ID) algorithm offers enhanced efficiency across both channels. Consequently, the subsequent course of action involves examining the impact of this ID on the analysis outlined in Chapter 3, encompassing all signal and background samples. Additionally, it is imperative to assess how the adoption of this ID will influence the sensitivity of the analysis.

6 Results

6.1 Implementation of the new lepton ID

Having demonstrated that the new lepton identification method is more efficient while still selecting high-purity leptons, we can now assess its impact on the results of the existing analysis.

The new identification method will be integrated into the offline event selection, replacing the current lepton ID. This offline event selection focuses on events where a Higgs boson is produced in association with $t\bar{t}$ events, and the Higgs boson decays into $c\bar{c}$. All three $t\bar{t}$ decay channels are considered: fully-hadronic (FH), semi-leptonic (SL), and dilepton (DL) decays. These channels are characterized by the presence of zero, one, or two isolated charged loose leptons (e, μ), missing transverse momentum due to neutrinos from W boson decays, and jets with transverse momenta typically in the range of several tens of GeV or more, originating from final-state quarks, many of which are b or c quarks. This selection process is termed as baseline selection, and the selection criteria are detailed in Table 11. With these changes and for the same samples new trees were created while having as reference the existing trees. The samples used for the trees are listed in Table 13 and Table 14.

	FH Channel	SL Channel	DL Channel
Number of loose leptons	0	1	2
Sign and flavour of leptons	-	$e^{\pm}$, $\mu^{\pm}$	$e^{+}e^{-}$, $e^{\pm}\mu^{\mp}$, $\mu^{\pm}\mu^{\pm}$
Min. p_T of p_T-leading electron [GeV]	-	29/30/30	25
Electron MVA	-	`mvaIso_WP80`	`mvaIso_WP90`
Min. p_T of p_T-leading muon [GeV]	-	26/29/26	25
Max. muon relative isolation	-	0.15	0.25
Max. $\|\eta\|$ of leptons	2.4	2.4	2.4
Min. number of jets	7	5	4
Min. p_T of 6th jets [GeV]	40	-	-
Min. number of b or c-jets (medium)	3	3	3
Min. number of b-jets (medium tagged)	1	1	1
Min. HT [GeV]	500	-	-
Min. $m_{ee,\mu\mu}$ [GeV]	-	-	20
$m_{ee,\mu\mu}$ [GeV]	-	-	<76 or >106
Min. MET [GeV]	-	20	20

Table 11: Baseline selection criteria in the fully-hadronic (FH), single-lepton (SL), and dilepton (DL) channels. Where the criteria differ per year, they are quoted as 2016/2017/2018.

For implementing the new ID we keep the kinematic cuts the same and we only change the selections as states in Table 12.

	FH Channel	SL Channel	DL Channel
Number of loose leptons	0	1	2
Sign and flavour of leptons	-	$e^{\pm}$, $\mu^{\pm}$	$e^{+}e^{-}$, $e^{\pm}\mu^{\mp}$, $\mu^{\pm}\mu^{\pm}$
Min. p_T of p_T-leading electron [GeV]	-	29/30/30	25
Electron ID	-	MVAscore>0.9	MVAscore>0.9
Min. p_T of p_T-leading muon [GeV]	-	26/29/26	25
Muon ID	-	MVAscore>0.9 and PFcand	MVAscore>0.9 and PFcand
Max. $\|\eta\|$ of leptons	2.4	2.4	2.4
Min. number of jets	7	5	4
Min. p_T of 6th jets [GeV]	40	-	-
Min. number of b or c-jets (medium)	3	3	3
Min. number of b-jets (medium tagged)	1	1	1
Min. HT [GeV]	500	-	-
Min. $m_{ee,\mu\mu}$ [GeV]	-	-	20
$m_{ee,\mu\mu}$ [GeV]	-	-	<76 or >106
Min. MET [GeV]	-	20	20

Table 12: Baseline selection criteria in the fully-hadronic (FH), single-lepton (SL), and dilepton (DL) channels with the new lepton ID. Changes are with red color.

Channel	**Sample**	**XS**
$t\bar{t}H(c\bar{c})$	`ttHTocc_M125_TuneCP5_13TeV-powheg-pythia8`	0.015
$t\bar{t}H(b\bar{b})$	`ttHTobb_M125_TuneCP5_13TeV-powheg-pythia8`	0.295

Table 13: Monte Carlo Signal Samples. XS is the corresponding cross section.

Channel	Sample	XS
$t\bar{t}$	TTToHadronic_TuneCP5_13TeV-powheg-pythia8	379.265
	TTToSemiLeptonic_TuneCP5_13TeV-powheg-pythia8	366.226
	TTTo2L2Nu_TuneCP5_13TeV-powheg-pythia8	88.409
$t\bar{t}b\bar{b}$	TTbb_4f_TTToHadronic_TuneCP5-Powheg-Openloops-Pythia8	19.902
	TTbb_4f_TTToSemiLeptonic_TuneCP5-Powheg-Openloops-Pythia8	19.218
	TTbb_4f_TTTo2L2Nu_TuneCP5-Powheg-Openloops-Pythia8	4.639
single top	ST_s-channel_4f_hadronicDecays_TuneCP5_13TeV-amcatnlo-pythia8	3.110
	ST_t-channel_antitop_4f_InclusiveDecays_TuneCP5_13TeV-powheg-madspin-pythia8	80.0
	ST_t-channel_top_4f_InclusiveDecays_TuneCP5_13TeV-powheg-madspin-pythia8	134.2
	ST_tW_antitop_5f_inclusiveDecays_TuneCP5_13TeV-powheg-pythia8	39.65
	ST_tW_top_5f_inclusiveDecays_TuneCP5_13TeV-powheg-pythia8	39.65
	ST_s-channel_4f_leptonDecays_TuneCP5_13TeV-amcatnlo-pythia8	3.729
	ST_tW_antitop_5f_NoFullyHadronicDecays_TuneCP5_13TeV-powheg-pythia8	21.617
	ST_tW_top_5f_NoFullyHadronicDecays_TuneCP5_13TeV-powheg-pythia8	21.617
$t\bar{t}W$	TTWJetsToLNu_TuneCP5_13TeV-amcatnloFXFX-madspin-pythia8	0.196
	TTWJetsToQQ_TuneCP5_13TeV-amcatnloFXFX-madspin-pythia8	0.405
$t\bar{t}Z$	TTZToLLNuNu_M-10_TuneCP5_13TeV-amcatnlo-pythia8	0.253
	TTZToQQ_TuneCP5_13TeV-amcatnlo-pythia8	0.586
qcd	QCD_HT300to500_TuneCP5_13TeV-madgraphMLM-pythia8	322600
	QCD_HT500to700_TuneCP5_13TeV-madgraphMLM-pythia8	29980
	QCD_HT700to1000_TuneCP5_13TeV-madgraphMLM-pythia8	6334
	QCD_HT1000to1500_TuneCP5_13TeV-madgraphMLM-pythia8	1088
	QCD_HT1500to2000_TuneCP5_13TeV-madgraphMLM-pythia8	99.11
	QCD_HT2000toInf_TuneCP5_13TeV-madgraphMLM-pythia8	20.23
W+jets	WJetsToQQ_HT-200to400_TuneCP5_13TeV-madgraphMLM-pythia8	2549.0
	WJetsToQQ_HT-400to600_TuneCP5_13TeV-madgraphMLM-pythia8	276.5
	WJetsToQQ_HT-600to800_TuneCP5_13TeV-madgraphMLM-pythia8	59.25
	WJetsToQQ_HT-800toInf_TuneCP5_13TeV-madgraphMLM-pythia8	28.75
	WJetsToLNu_0J_TuneCP5_13TeV-amcatnloFXFX-pythia8	48716.955
	WJetsToLNu_1J_TuneCP5_13TeV-amcatnloFXFX-pythia8	8107.312
	WJetsToLNu_2J_TuneCP5_13TeV-amcatnloFXFX-pythia8	3049.263
Z+jets	ZJetsToQQ_HT-200to400_TuneCP5_13TeV-madgraphMLM-pythia8	1012.0
	ZJetsToQQ_HT-400to600_TuneCP5_13TeV-madgraphMLM-pythia8	114.2
	ZJetsToQQ_HT-600to800_TuneCP5_13TeV-madgraphMLM-pythia8	25.34
	ZJetsToQQ_HT-800toInf_TuneCP5_13TeV-madgraphMLM-pythia8	12.99
	DYJetsToLL_M-10to50_TuneCP5_13TeV-madgraphMLM-pythia8	22635
	DYJetsToLL_M-50_TuneCP5_13TeV-amcatnloFXFX-pythia8	6077.22
$t\bar{t}H(\tau\bar{\tau})$	ttHToTauTau_M125_TuneCP5_13TeV-powheg-pythia8	0.032
ttHNonbb	ttHToNonbb_M125_TuneCP5_13TeV-powheg-pythia8	0.212

Table 14: Montecarlo Background Samples. XS is the corresponding cross section.

6.2 Comparison variables

The first step involved comparing the yields of every tree generated with both the new and the old ID to ensure that the change in ID continues to benefit this specific analysis with additional requirements, and to determine if this benefit is uniform or if a trend exists. The focus and comparison histograms are for the DL and SL channels, as these are affected by the electron ID. We plotted the kinematic lepton variables (p_T, η, ϕ), MET, and

the Particle Transformer score outputs as explained in section 3.2. To observe the impact on one signal and one background sample, the results will be examined for $t\bar{t}$ and $t\bar{t}H(c\bar{c})$.

6.2.1 Dilepton Channel-$t\bar{t}$

Figure 68 shows the comparison of various distributions of variables for dileptonic $t\bar{t}$ selected with different IDs.

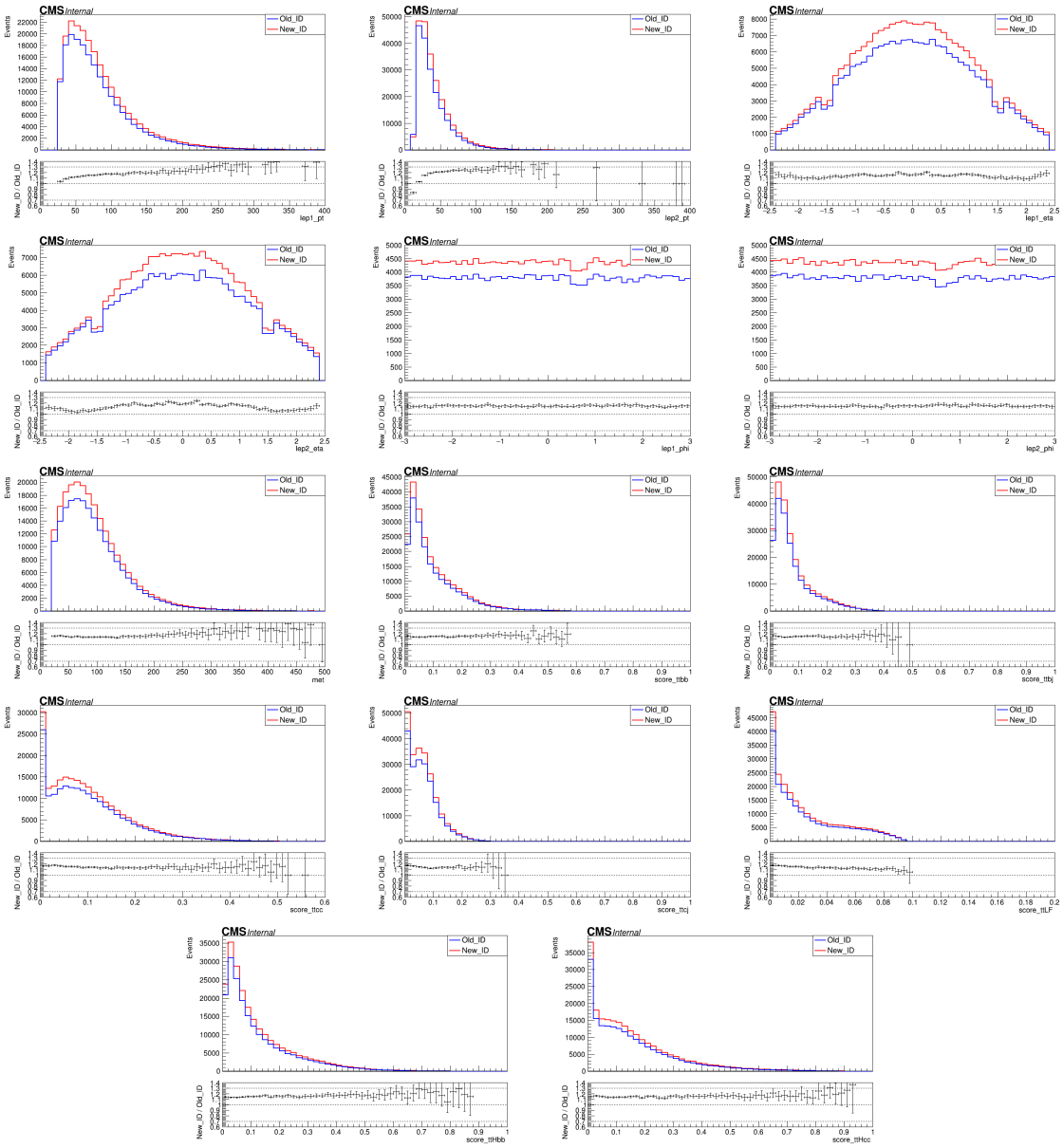


Figure 68: Comparison of di-leptonic $t\bar{t}$ event yield distributions for different leptons IDs

6.2.2 Dilepton Channel-$t\bar{t}H(c\bar{c})$

Figure 69 shows the comparison of various distributions of variables for dileptonic $t\bar{t}H(c\bar{c})$ selected with different IDs.

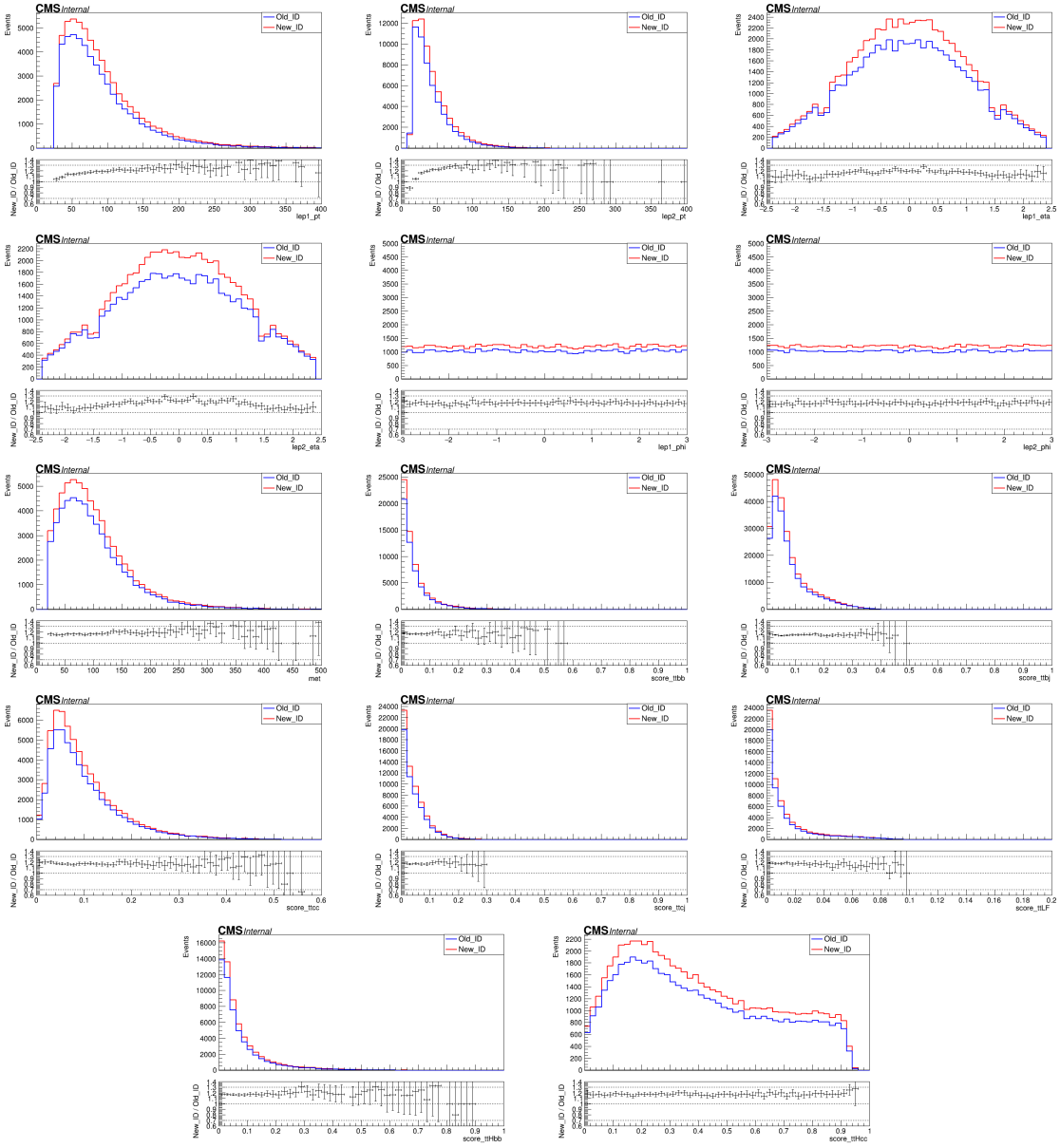


Figure 69: Comparison of di-leptonic $t\bar{t}H(c\bar{c})$ event yield distributions for different leptons IDs.

6.2.3 Single-lepton Channel-$t\bar{t}$

Figure 70 shows the comparison of various distributions of variables for semi-leptonic $t\bar{t}$ selected with different IDs.

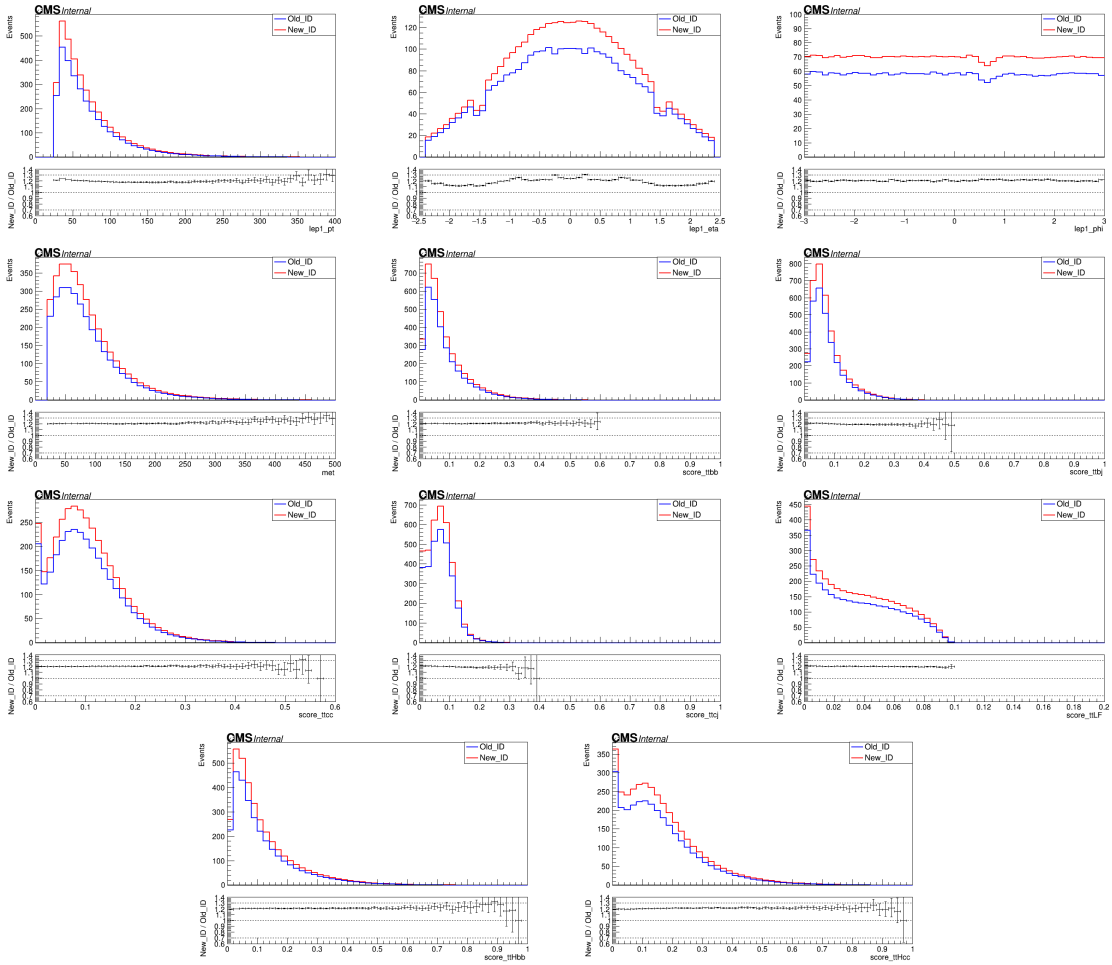


Figure 70: Comparison of semi-leptonic $t\bar{t}$ event yield distributions for different leptons IDs.

6.2.4 Single-lepton Channel-$t\bar{t}H(c\bar{c})$

Figure 71 shows the comparison of various distributions of variables for semi-leptonic $t\bar{t}H(c\bar{c})$ selected with different IDs.

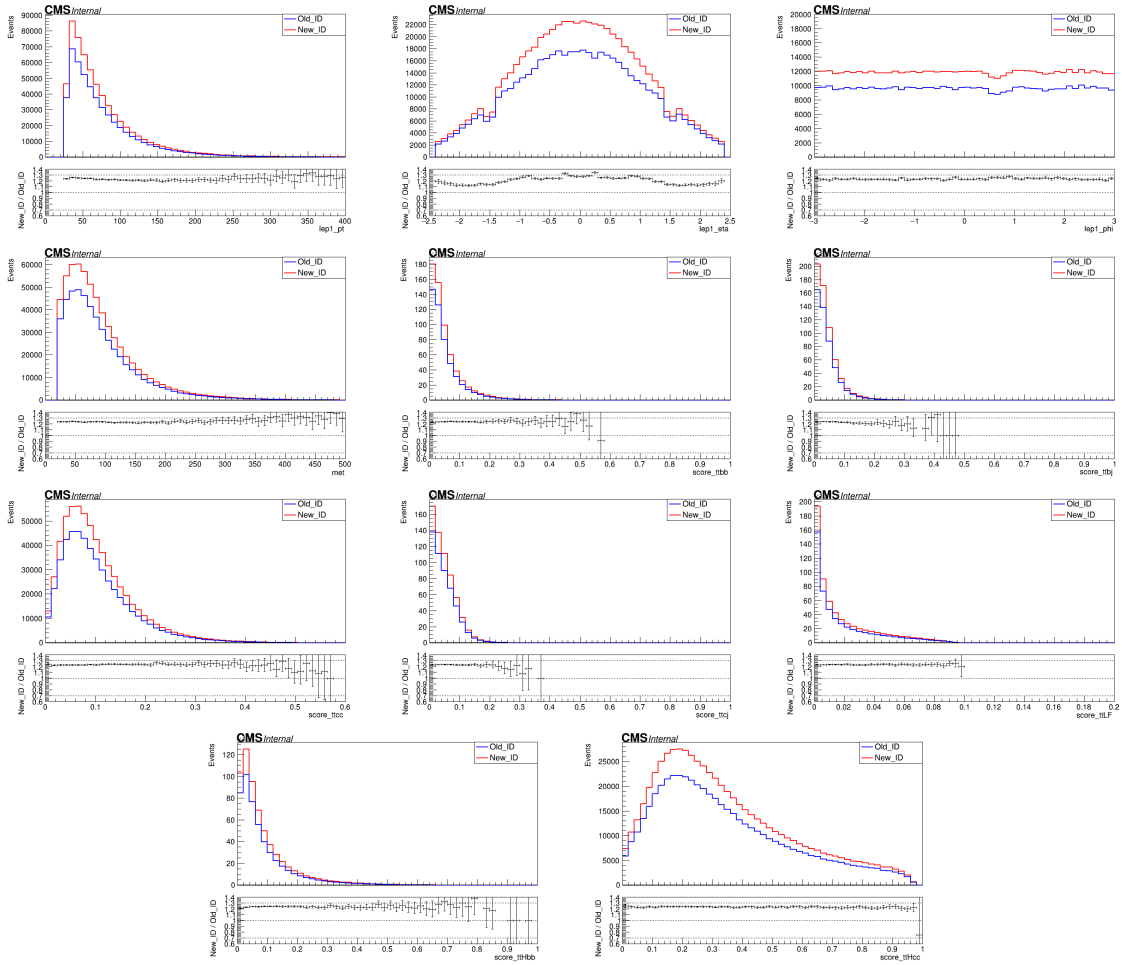


Figure 71: Comparison of semi-leptonic $t\bar{t}H(c\bar{c})$ event yield distributions for different leptons IDs.

The comparison reveals a p_T-dependent trend in the gains, particularly in the DL channel and the subleading lepton's p_T. In the low momentum region, the new ID selects fewer leptons compared to the existing selection ID. This is not a warning because otherwise there is a threshold for low p_T. Also, especially for the DL channel we observe a unique behaviour for the met variable where there is a minor drop in the met before the rise. The possible reason behind it is the refuting behaviour for the leading and subleading momentum which is linked with MET. Overall, the new ID shows a clear improvement in the number of events selected, while the distributions of scores and variables remain largely unchanged. To address the p_T trend, a unique set of scale factors (SFs) was developed for the new ID to achieve a specified agreement between the Data/MC plots, which will be discussed in the next section.

6.3 Data/MC plots

The next step to determine the suitability of the new ID for the analysis is to generate Data/MC plots to compare the results between the two IDs including the SFs. The selection criteria remain based on the baseline selection (Tables 11 and 12) and the midscore validation region as outlined in Figure 34. The plots are divided by lepton channel (ee, eμ, $\mu\mu$ for the DL channel and e, μ for the SL channel) to assess if the ID change impacts one channel more than the others.

6.3.1 Dilepton channel

Di-electron

The following figures present a comparison of Data/MC plots for different lepton identification methods in the electron-electron (EE) channel. The comparisons are made for various kinematic variables (lepton transverse momentum, pseudorapidity, azimuthal angle) and neural network scores. The left column shows the results using the new MVA-based lepton identification method, whereas the right column shows the results using the older cut-based lepton identification method.

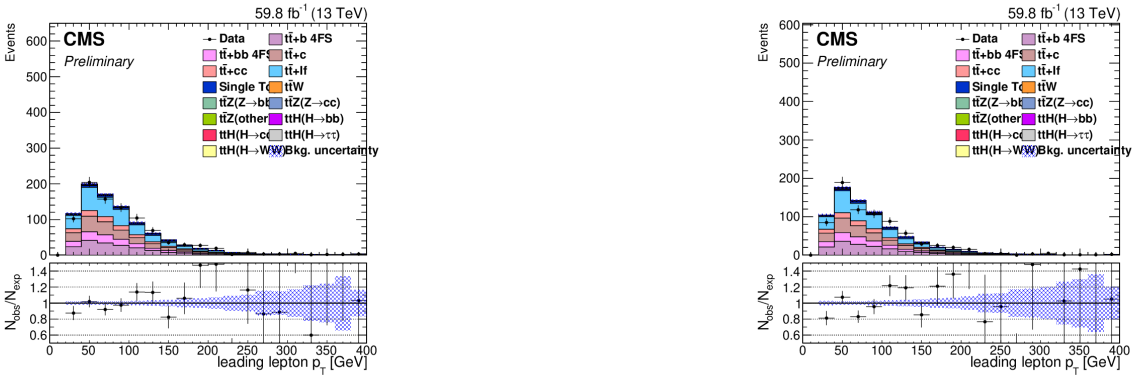


Figure 72: Comparison for Data/MC plots of lepton 1 transverse momentum in the electron-electron (EE) channel. **New: left. Old: right**

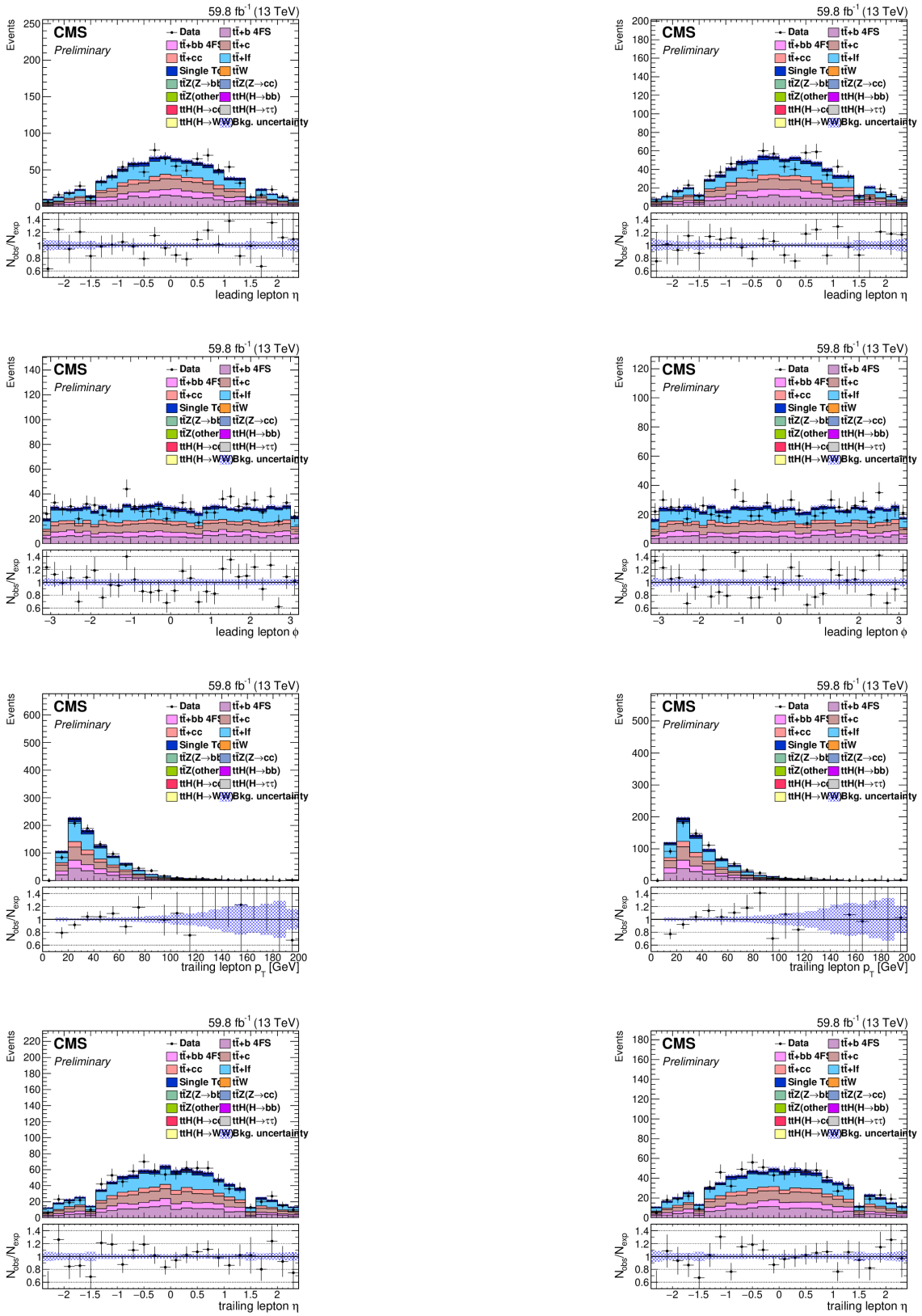


Figure 73: Comparison for Data/MC plots of lepton 1 pseudorapidity (top row), lepton 1 azimuthal angle (second row), lepton 2 transverse momentum (third row), and lepton 2 pseudorapidity (bottom row) in the electron-electron (EE) channel. **New: left. Old: right**

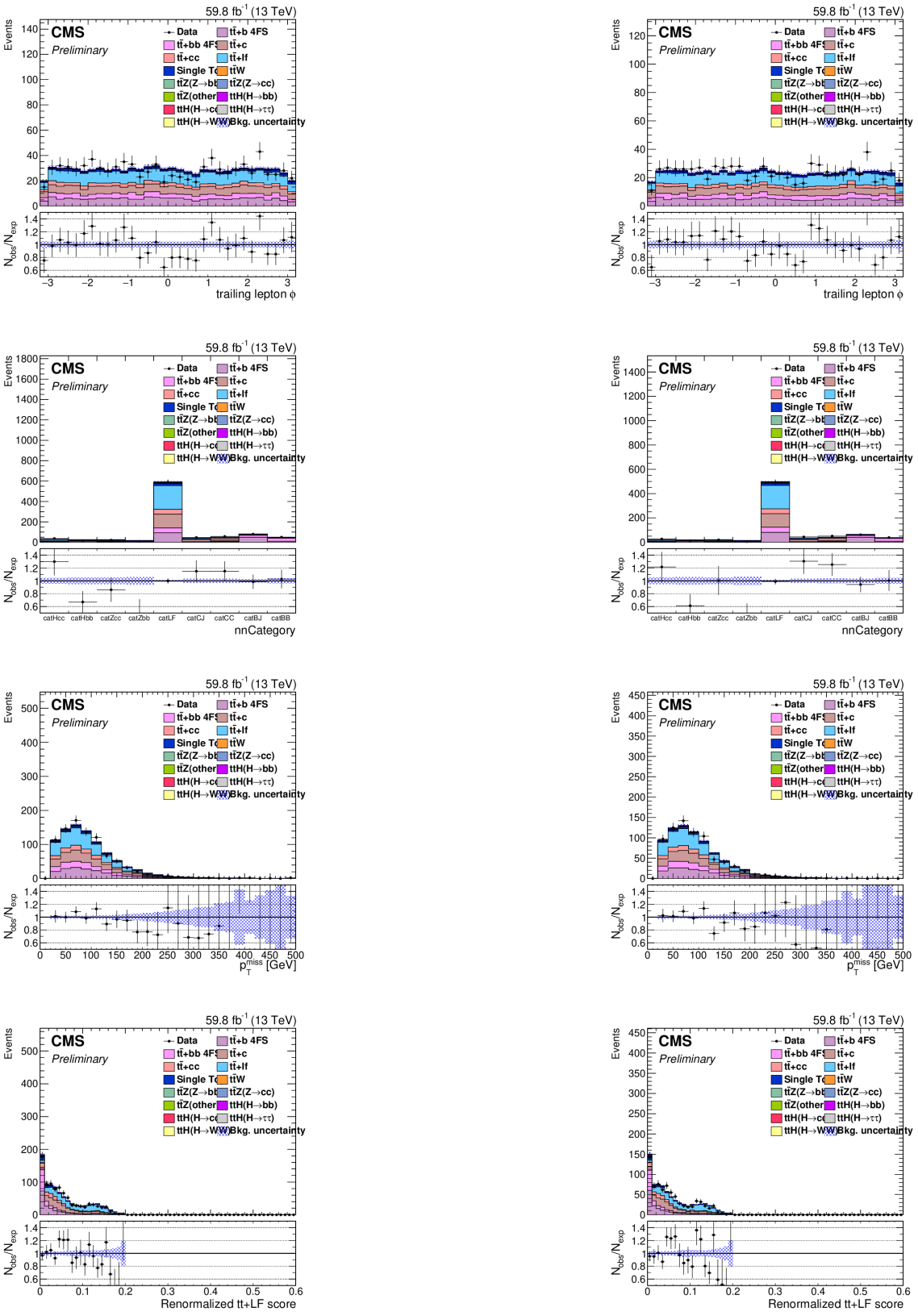


Figure 74: Comparison for Data/MC plots of lepton 2 azimuthal angle (top row), NN category (second row), MET (third row), and ParT score ratio for $t\bar{t}$+LF (bottom row) in the electron-electron (EE) channel. **New: left. Old: right**

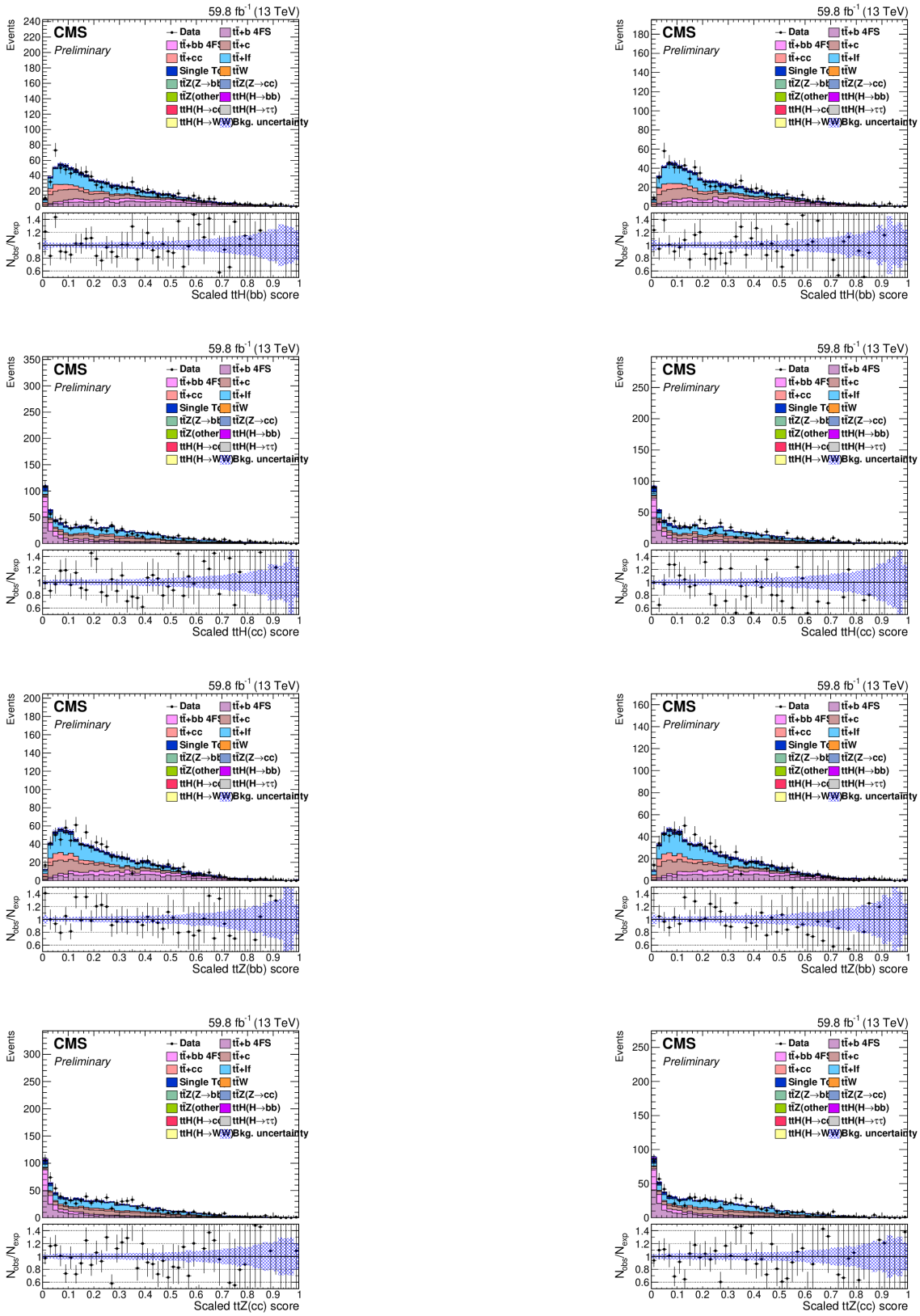


Figure 75: Comparison for Data/MC plots of ParT score ratios for $t\bar{t}H(b\bar{b})$ (top row), $t\bar{t}H(c\bar{c})$ (second row), $t\bar{t}Z(b\bar{b})$ (third row), and $t\bar{t}Z(c\bar{c})$ (bottom row) in the electron-electron (EE) channel. **New: left. Old: right**

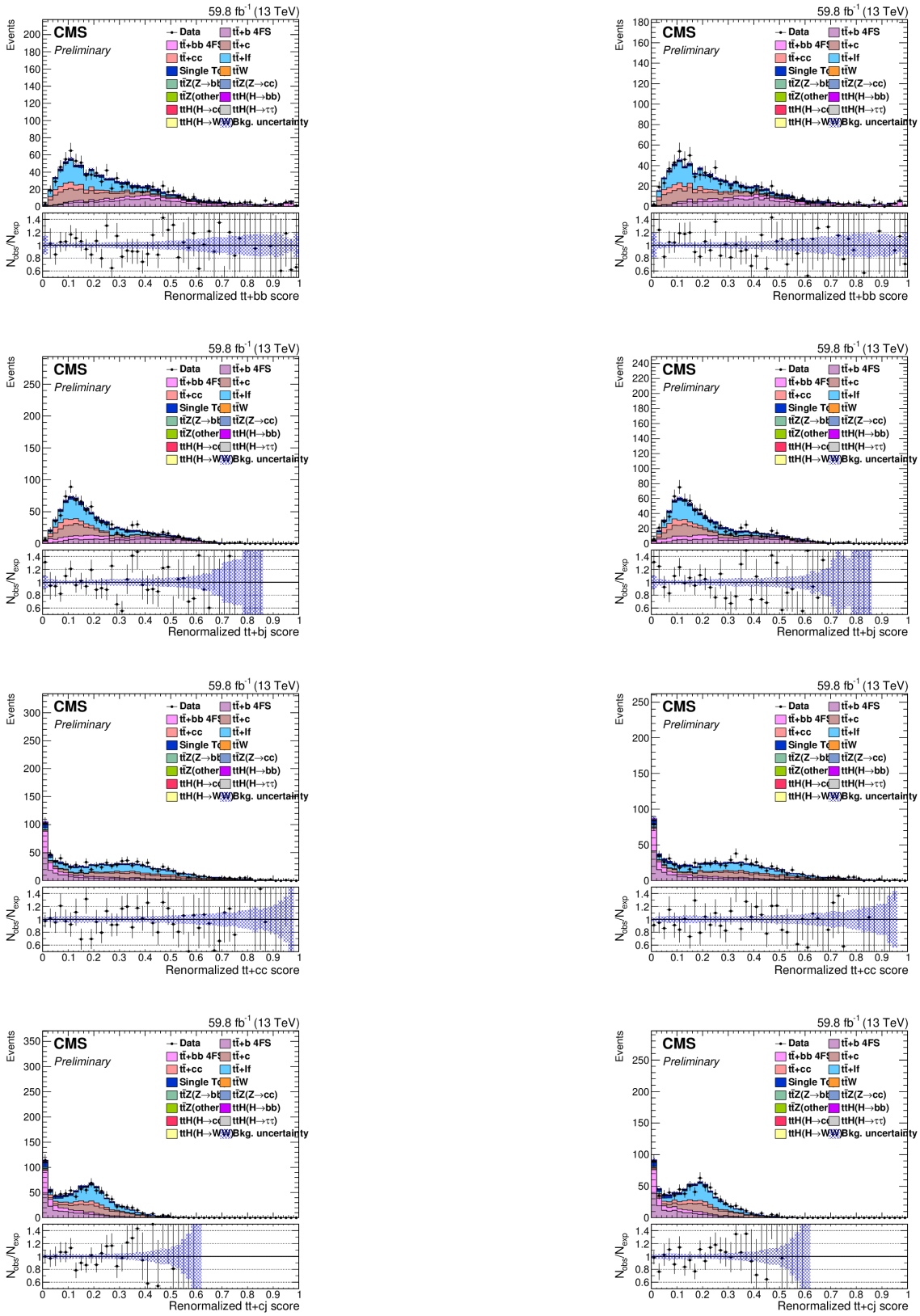


Figure 76: Comparison for Data/MC plots of ParT score ratios for $t\bar{t}+b\bar{b}$ (top row), $t\bar{t}$+bj (second row), $t\bar{t}+c\bar{c}$ (third row), and $t\bar{t}$+cj (bottom row) in the electron-electron (EE) channel. **New: left. Old: right**

Electron-Muon

The following figures present a comparison of Data/MC plots for different lepton identification methods in the electron-muon (EM) channel. The comparisons are made for various kinematic variables (lepton transverse momentum, pseudorapidity, azimuthal angle) and neural network scores. The left column shows the results using the new MVA-based lepton identification method, whereas the right column shows the results using the older cut-based lepton identification method.

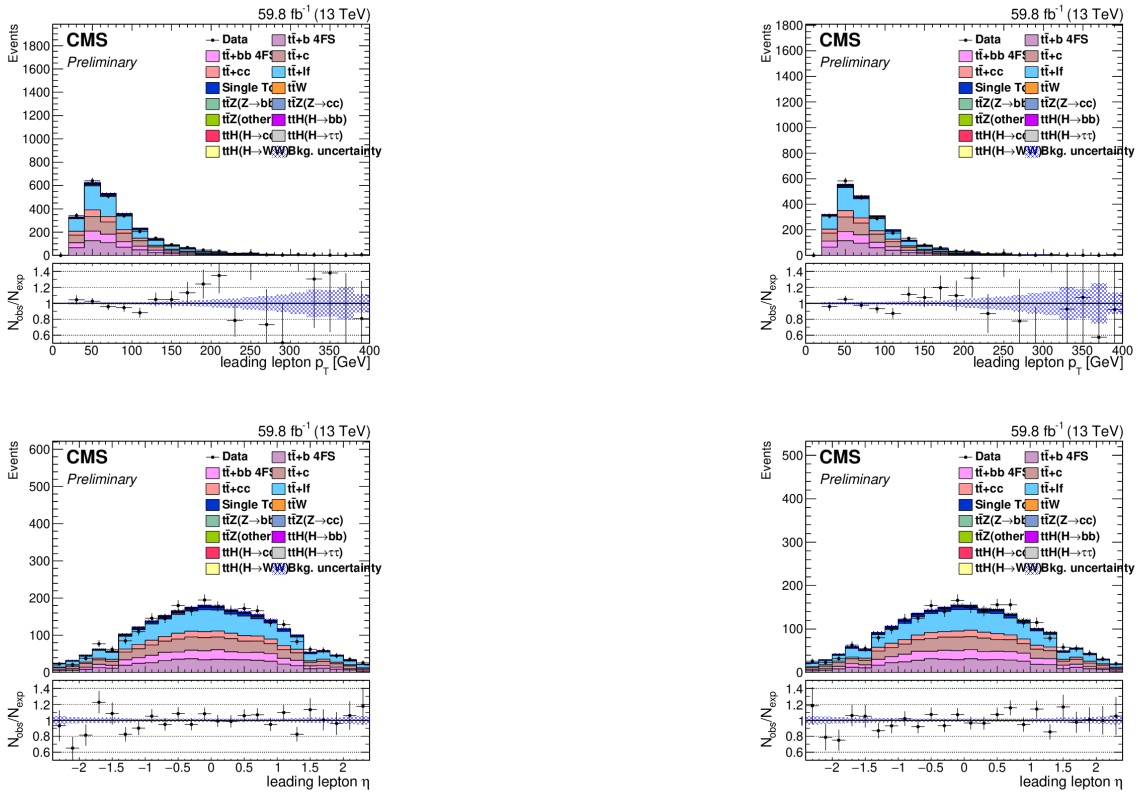


Figure 77: Comparison for Data/MC plots of lepton 1 transverse momentum (top row) and lepton 1 pseudorapidity (bottom row) in the electron-muon (EM) channel. **New: left. Old: right**

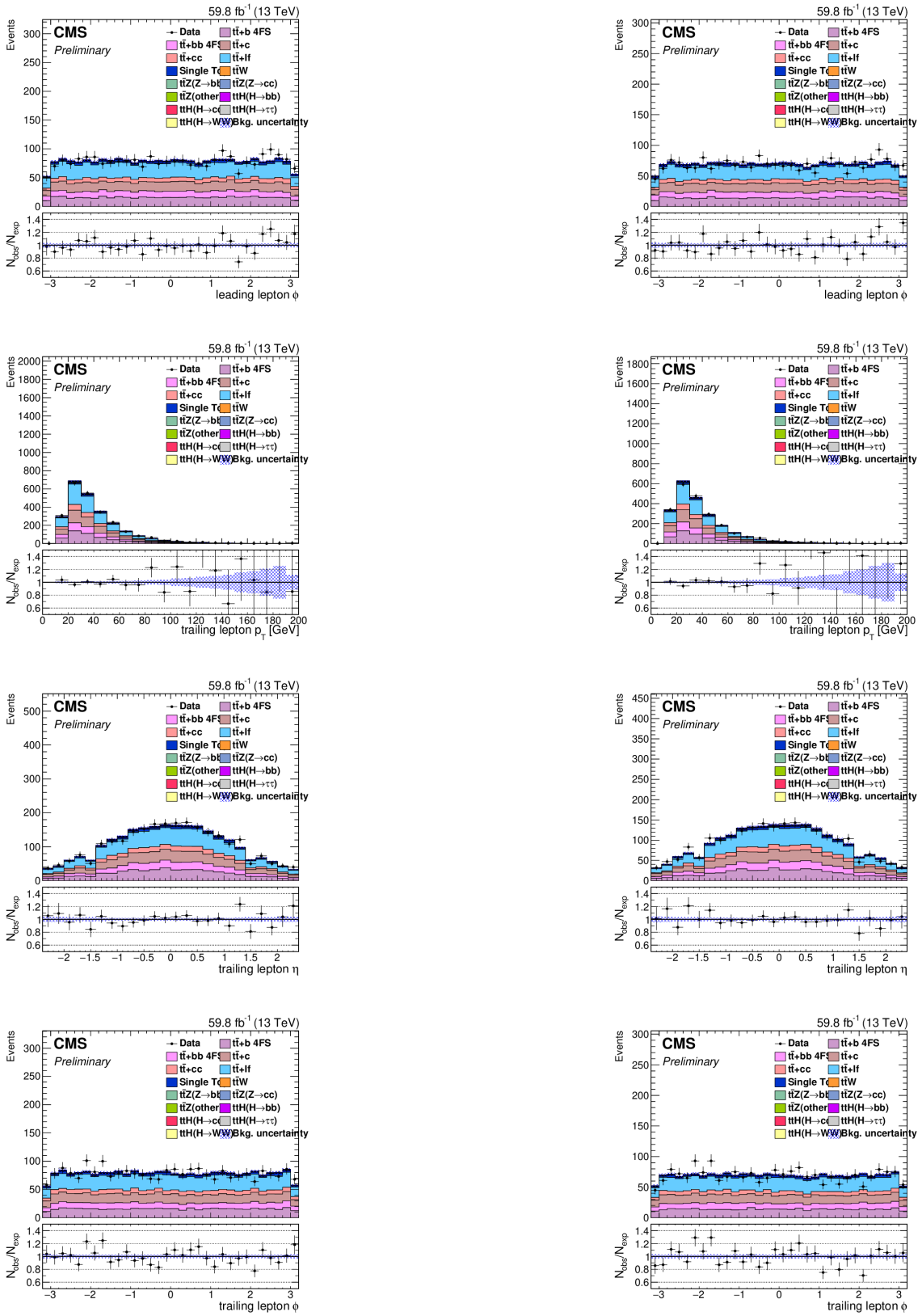


Figure 78: Comparison for Data/MC plots of lepton 1 azimuthal angle (top row), lepton 2 transverse momentum (second row), lepton 2 pseudorapidity (third row), and lepton 2 azimuthal angle (bottom row) in the electron-muon (EM) channel. **New: left. Old: right**

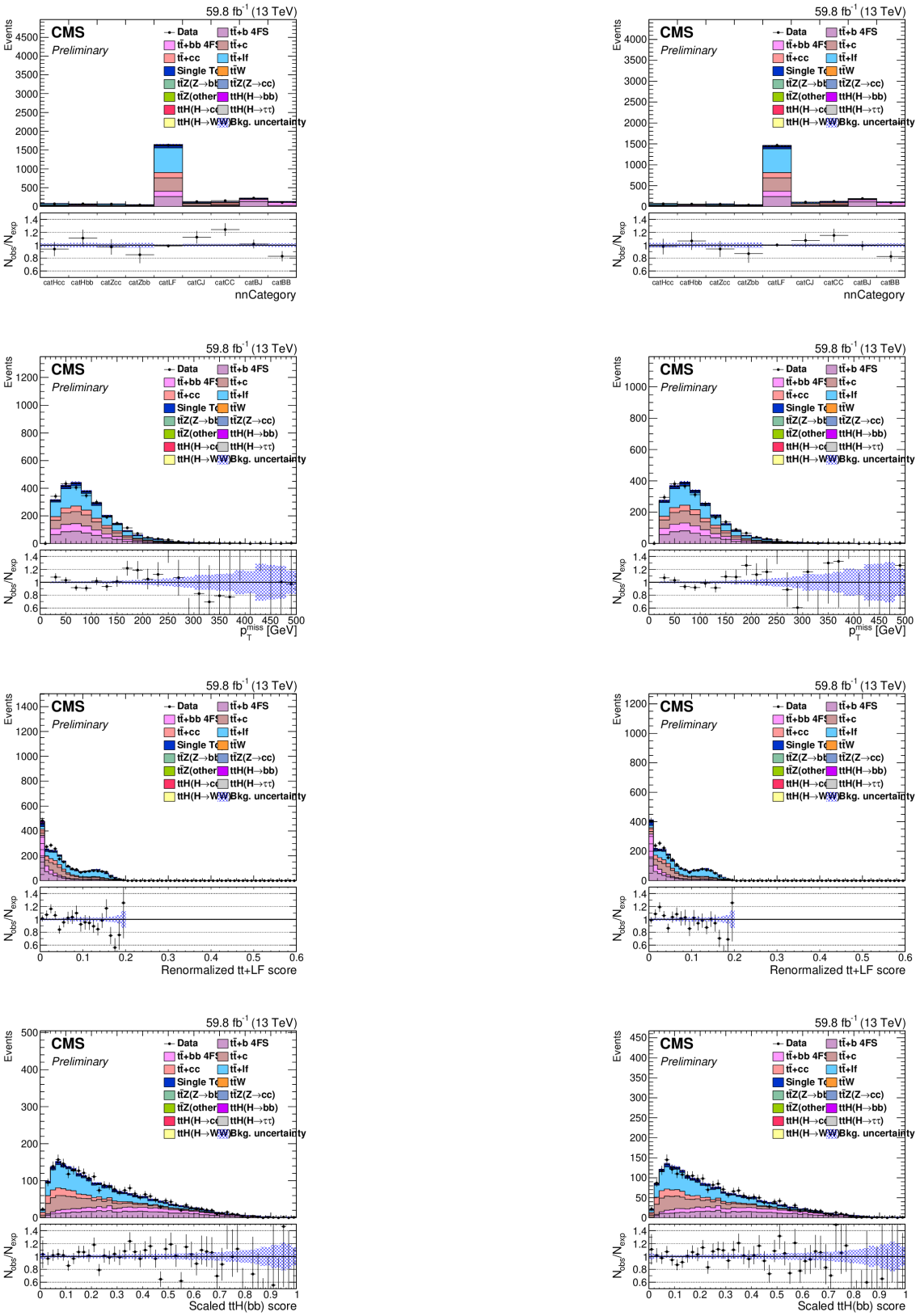


Figure 79: Comparison for Data/MC plots of NN category (top row), MET (second row), ParT score ratio for $t\bar{t}$+LF (third row), and ParT score ratio for $t\bar{t}H(b\bar{b})$ (bottom row) in the electron-muon (EM) channel. **New: left. Old: right**

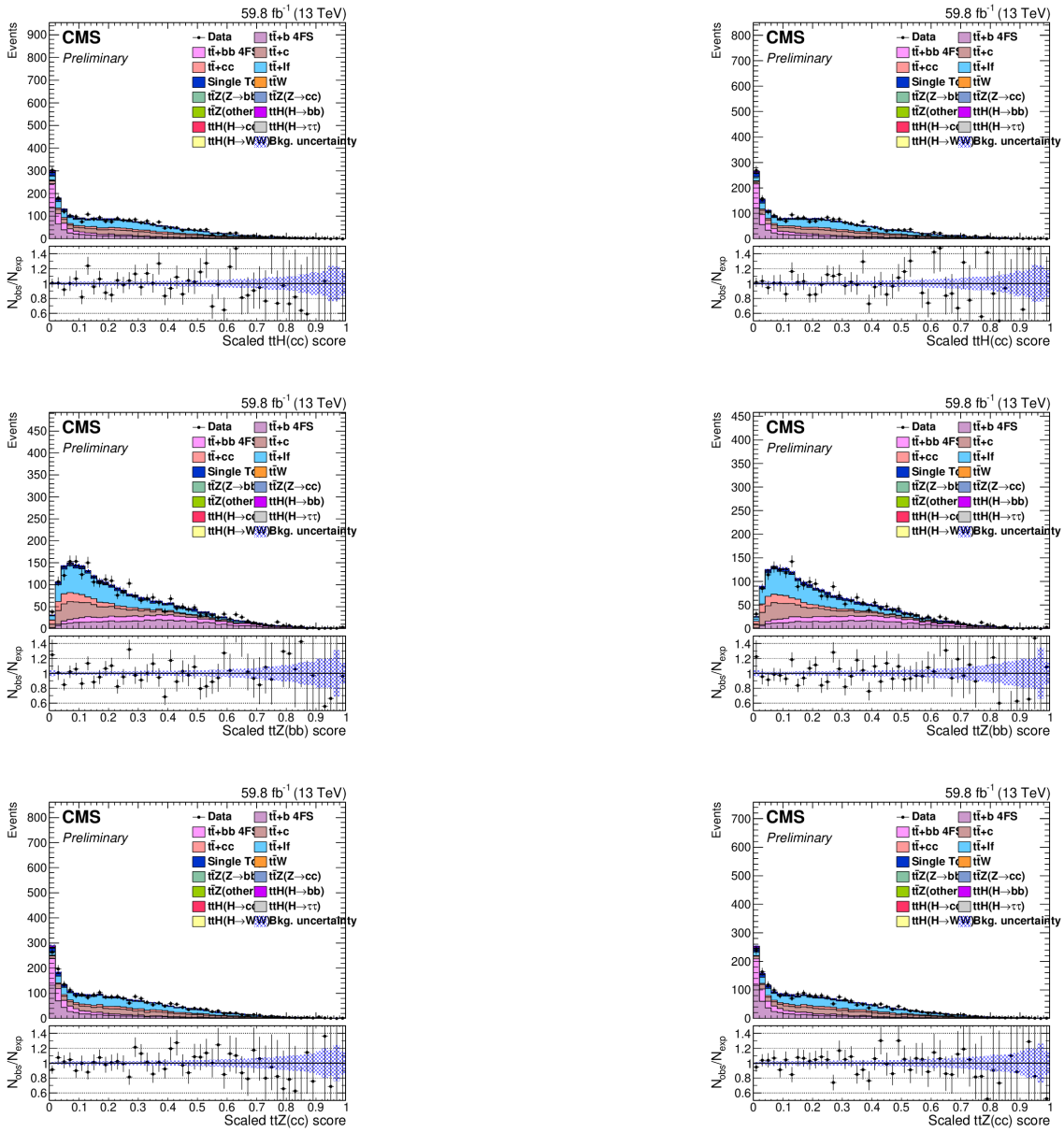


Figure 80: Comparison for Data/MC plots of ParT score ratios for $t\bar{t}H(c\bar{c})$ (top row), $t\bar{t}Z(b\bar{b})$ (second row), and $t\bar{t}Z(c\bar{c})$ (bottom row) in the electron-muon (EM) channel. **New: left. Old: right**

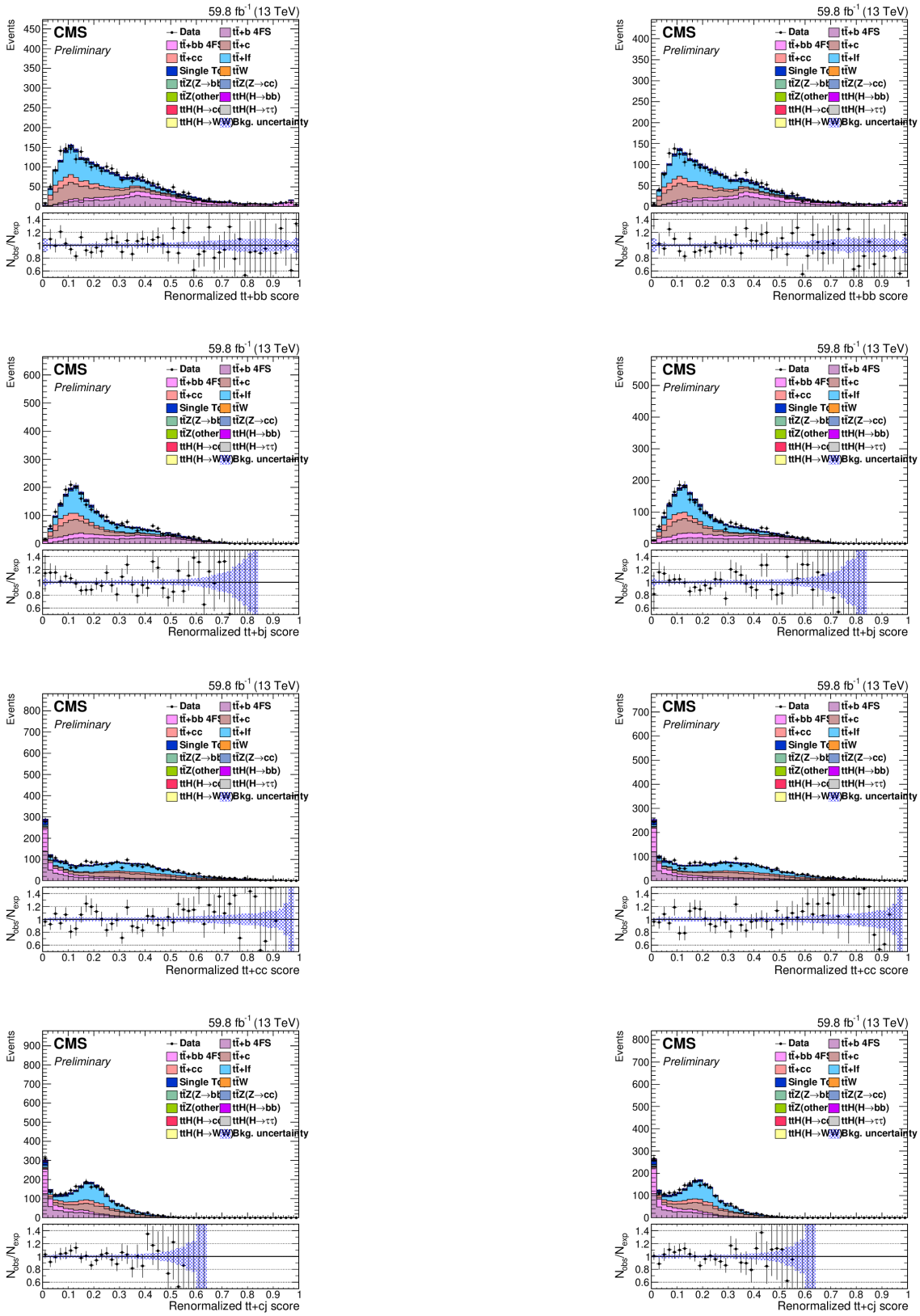


Figure 81: Comparison for Data/MC plots of ParT score ratios for $t\bar{t}+b\bar{b}$ (top row), $t\bar{t}$+bj (second row), $t\bar{t}+c\bar{c}$ (third row), and $t\bar{t}$+cj (bottom row) in the electron-muon (EM) channel. **New: left. Old: right**

Di-Muon

The following figures present a comparison of Data/MC plots for different lepton identification methods in the muon-muon (MM) channel. The comparisons are made for various kinematic variables (lepton transverse momentum, pseudorapidity, azimuthal angle) and neural network scores. The left column shows the results using the new MVA-based lepton identification method, whereas the right column shows the results using the older cut-based lepton identification method.

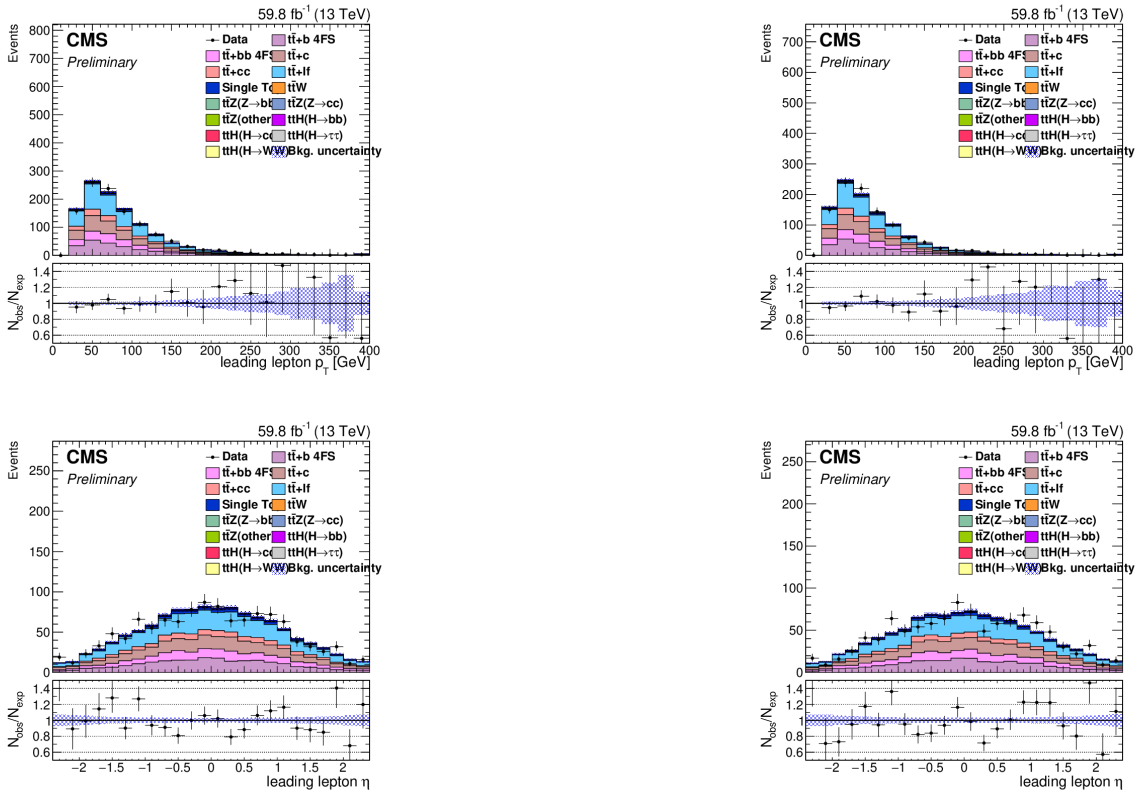


Figure 82: Comparison for Data/MC plots of lepton 1 transverse momentum (top row) and lepton 1 pseudorapidity (bottom row) in the muon-muon (MM) channel. **New: left. Old: right**

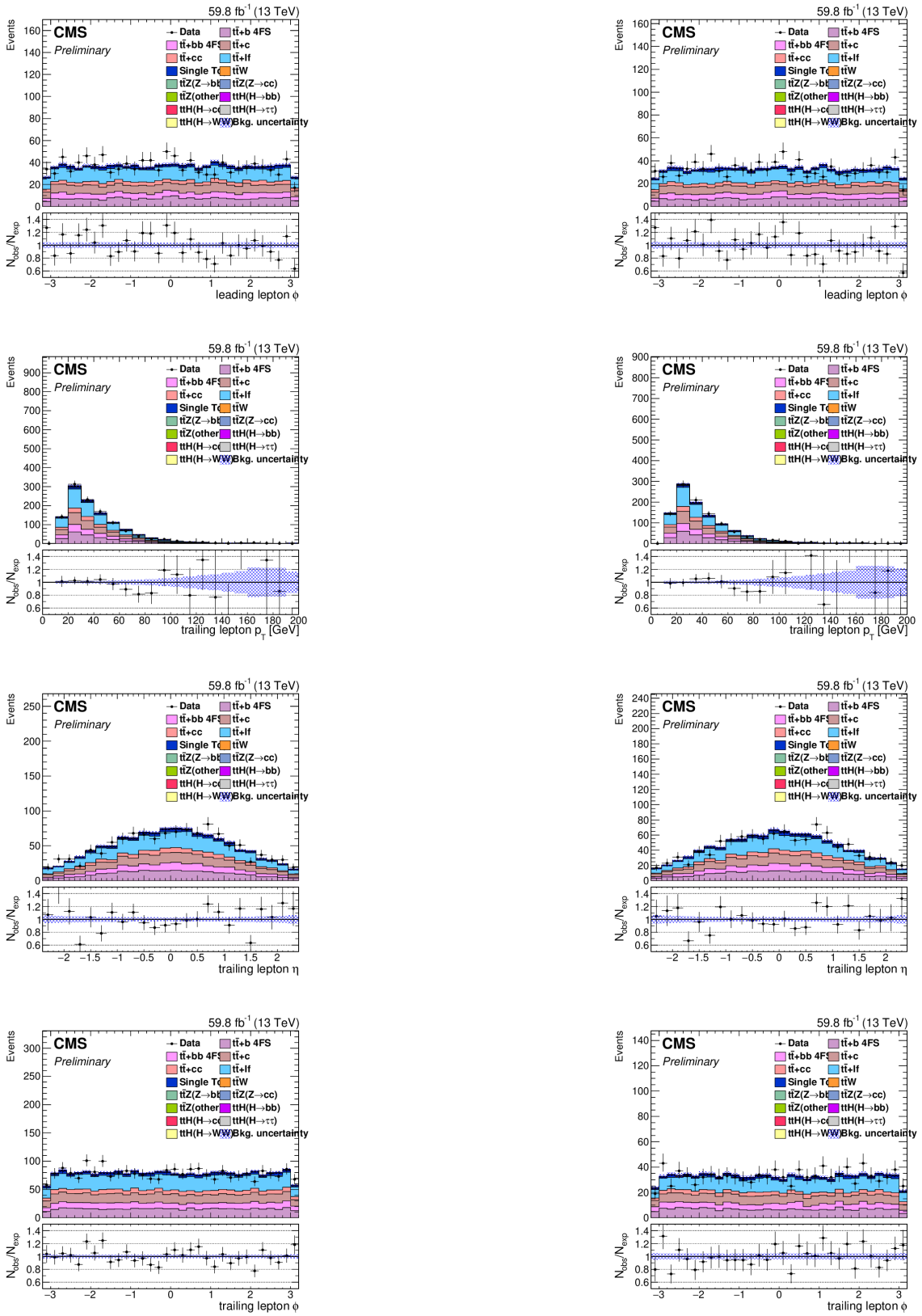


Figure 83: Comparison for Data/MC plots of lepton 1 azimuthal angle (top row), lepton 2 transverse momentum (second row), lepton 2 pseudorapidity (third row), and lepton 2 azimuthal angle (bottom row) in the muon-muon (MM) channel. **New: left. Old: right**

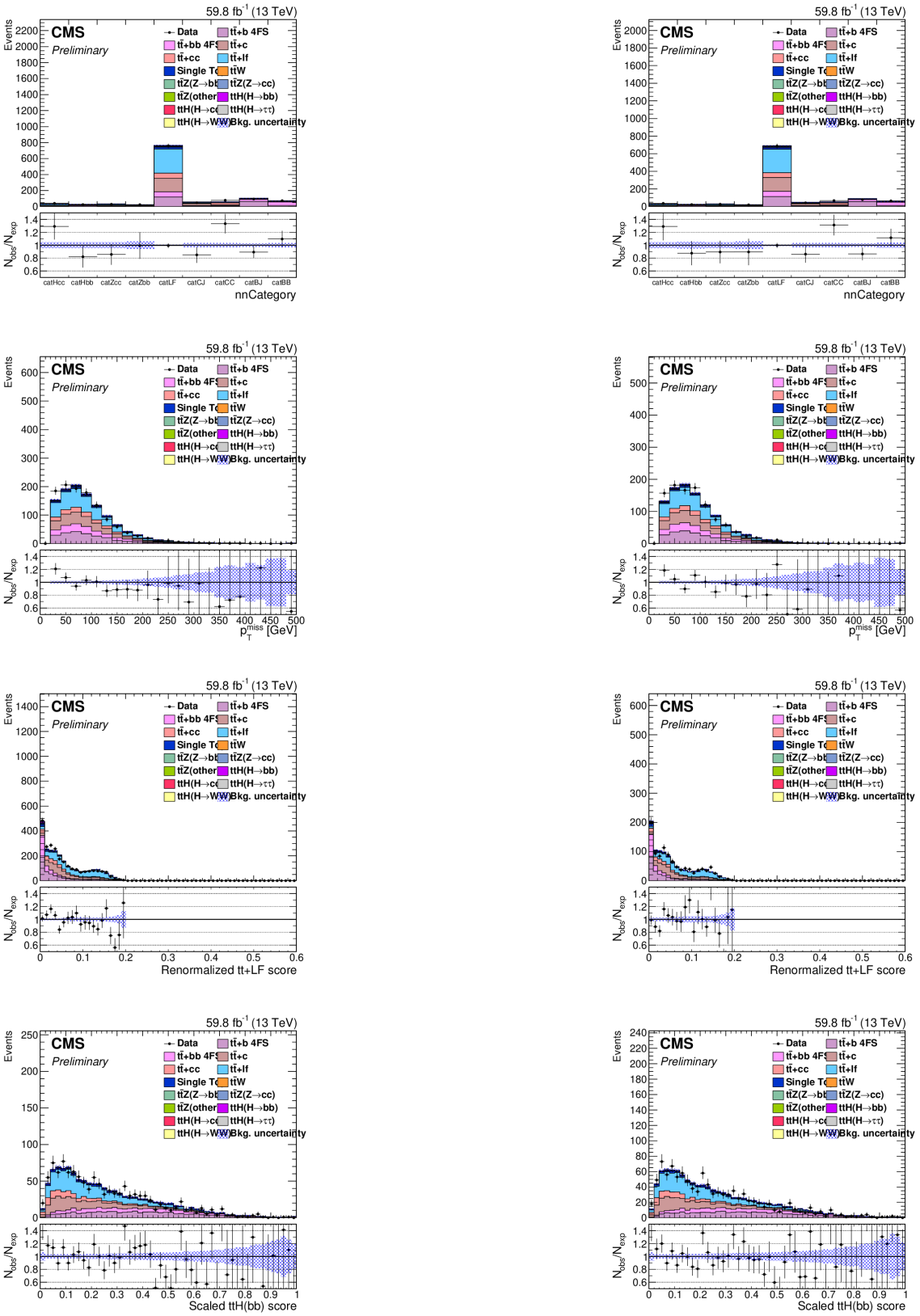


Figure 84: Comparison for Data/MC plots of NN category (top row), MET (second row), ParT score ratio for $t\bar{t}$+LF (third row), and ParT score ratio for $t\bar{t}H(b\bar{b})$ (bottom row) in the muon-muon (MM) channel. **New: left. Old: right**

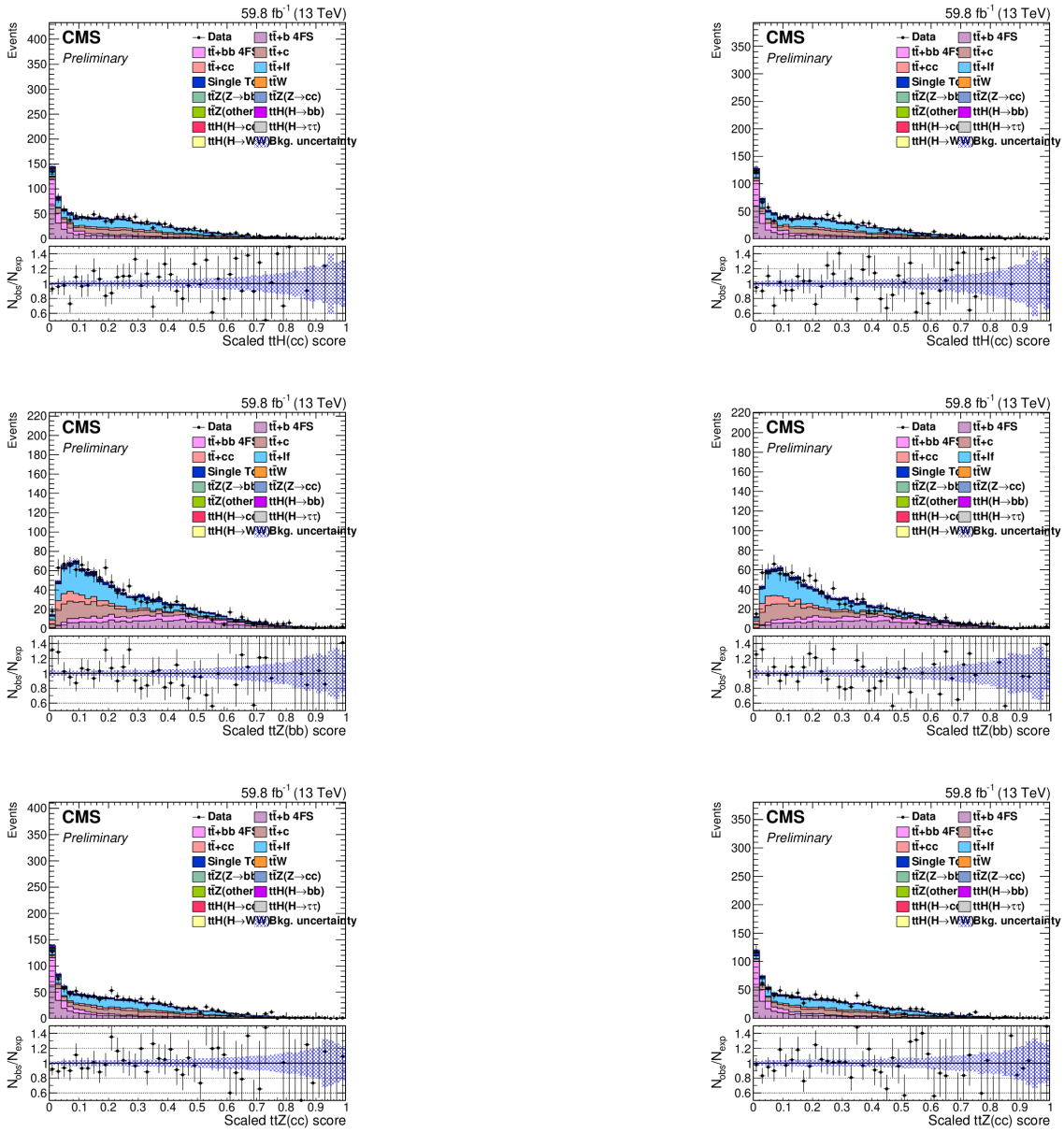


Figure 85: Comparison for Data/MC plots of ParT score ratios for $t\bar{t}H(c\bar{c})$ (top row), $t\bar{t}Z(b\bar{b})$ (second row), and $t\bar{t}Z(c\bar{c})$ (bottom row) in the muon-muon (MM) channel. **New: left. Old: right**

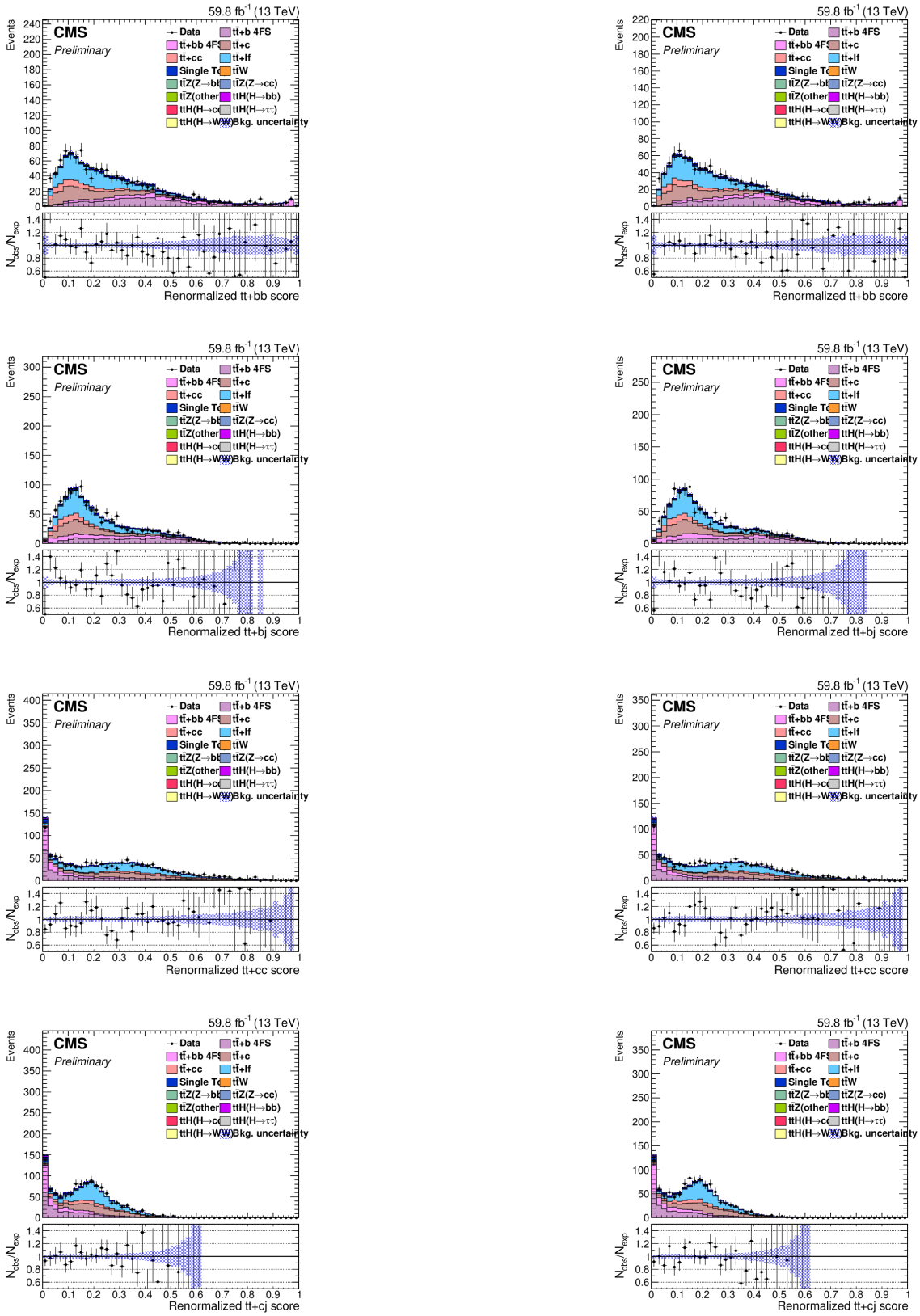


Figure 86: Comparison for Data/MC plots of ParT score ratios for $t\bar{t}+b\bar{b}$ (top row), $t\bar{t}$+bj (second row), $t\bar{t}+c\bar{c}$ (third row), and $t\bar{t}$+cj (bottom row) in the muon-muon (MM) channel. **New: left. Old: right**

For each of the three channels, the implementation of the new lepton ID results in a gain. The channel most significantly affected by this change is the di-electron channel, followed by the electron-muon (EM) channel, and lastly the dimuon channel. The Data/MC plots remain consistent across all cases after adjusting the different scale factors. All in all, the picture is very encouraging for the DL channel since there are more events spread uniformly which may result in better sensitivity for the analysis.

6.3.2 Single-lepton channel

Electron

The following figures present a comparison of Data/MC plots for different lepton identification methods in the case that the lepton in the event is an electron. The comparisons are made for various kinematic variables and neural network scores. The left column shows the results using the new MVA-based lepton identification method, whereas the right column shows the results using the older cut-based lepton identification method.

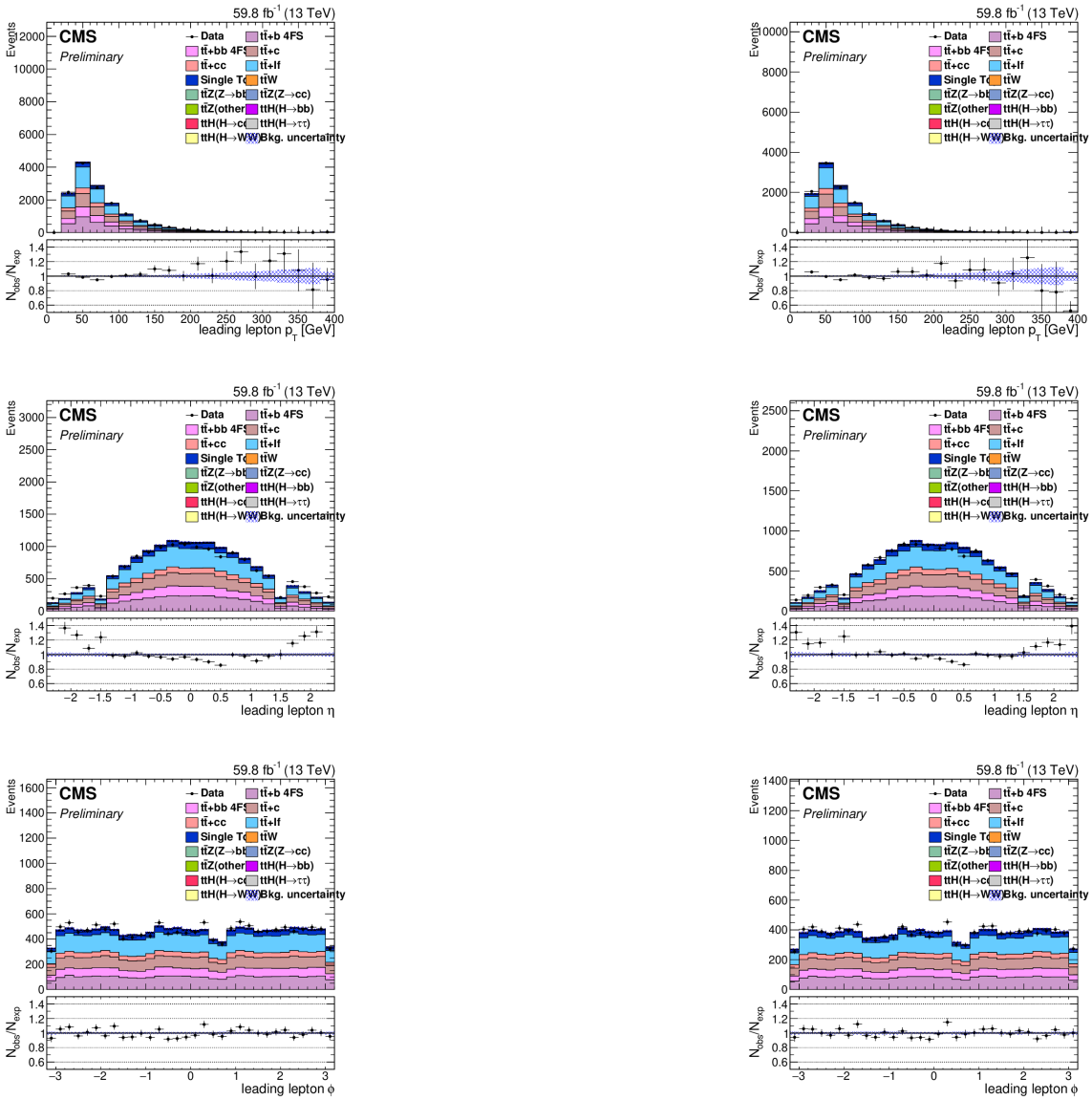


Figure 87: Comparison for Data/MC plots of electron transverse momentum (top row),electron pseudorapidity (middle row) and electron azimuthal angle (bottom row) in the electron channel. **New: left. Old: right**

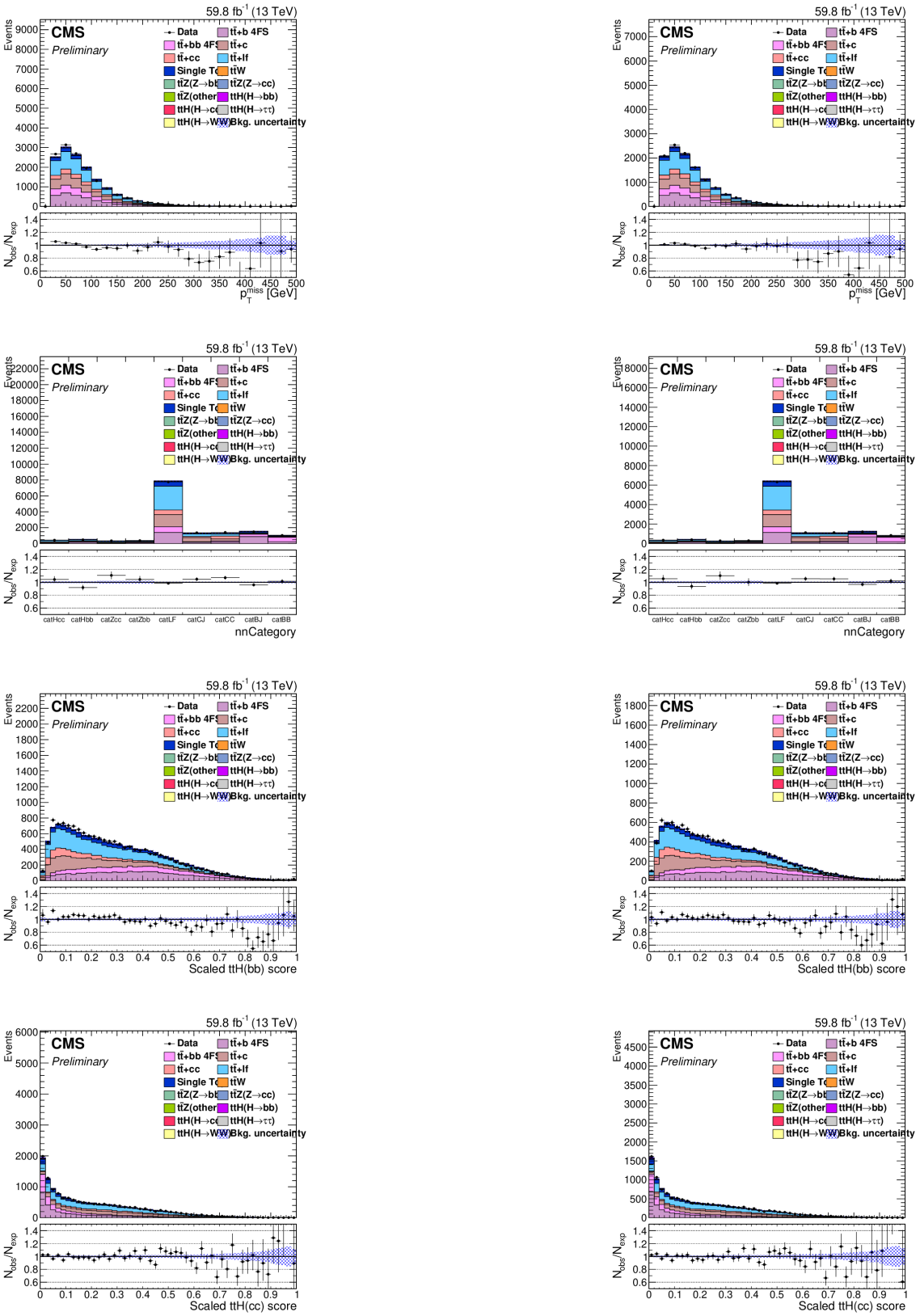


Figure 88: Comparison for Data/MC plots of MET (top row), NN category (second row), ParT score ratio for $t\bar{t}H(b\bar{b})$ (third row), and ParT score ratio for $t\bar{t}H(c\bar{c})$ (bottom row) in the electron channel. **New: left. Old: right**

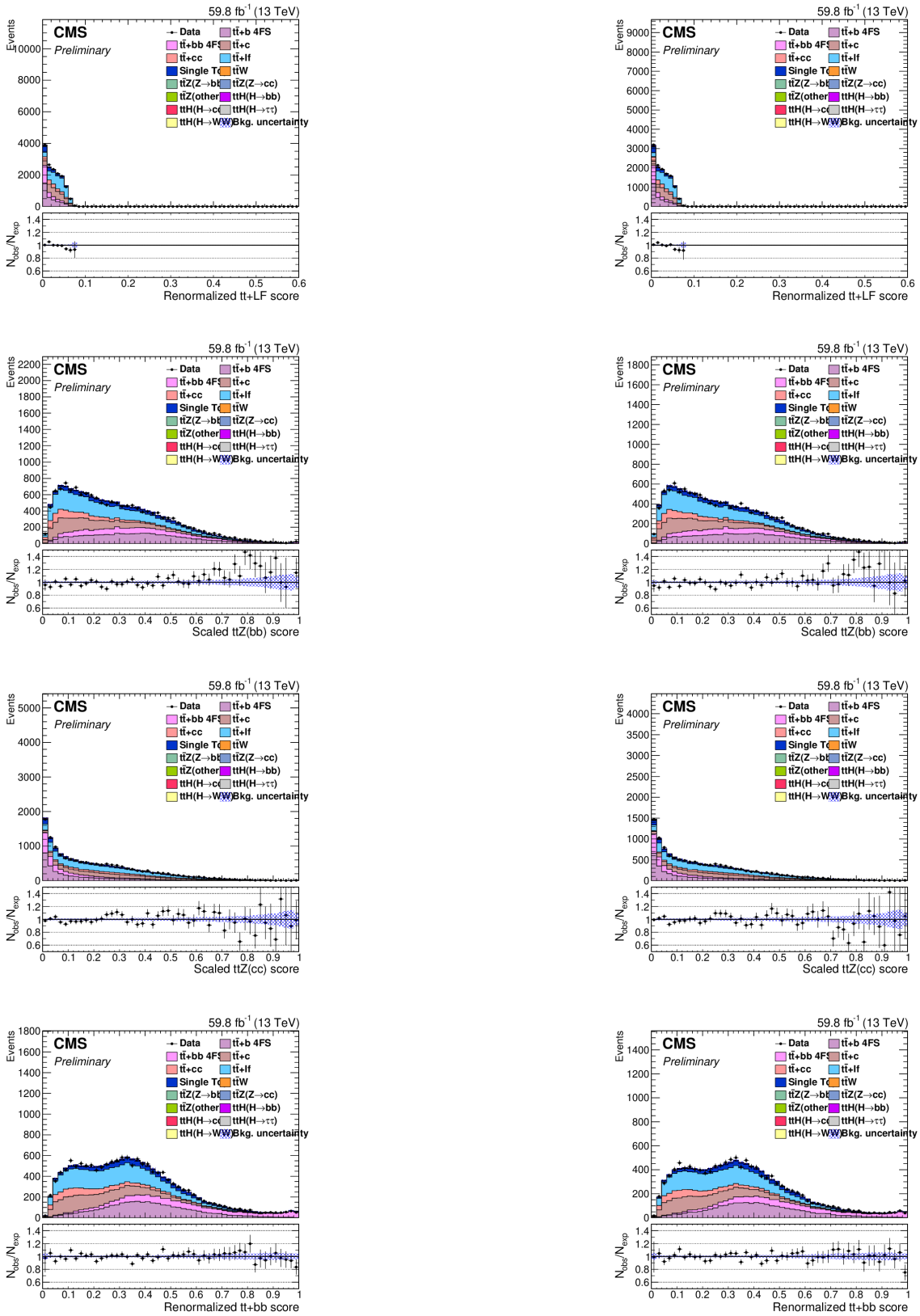


Figure 89: Comparison for Data/MC plots of ParT score ratios for $t\bar{t}$+LF (top row), $t\bar{t}Z(b\bar{b})$ (second row), $t\bar{t}Z(c\bar{c})$ (third row) and $t\bar{t}+b\bar{b}$ (bottom row) in the electron channel. **New: left. Old: right**

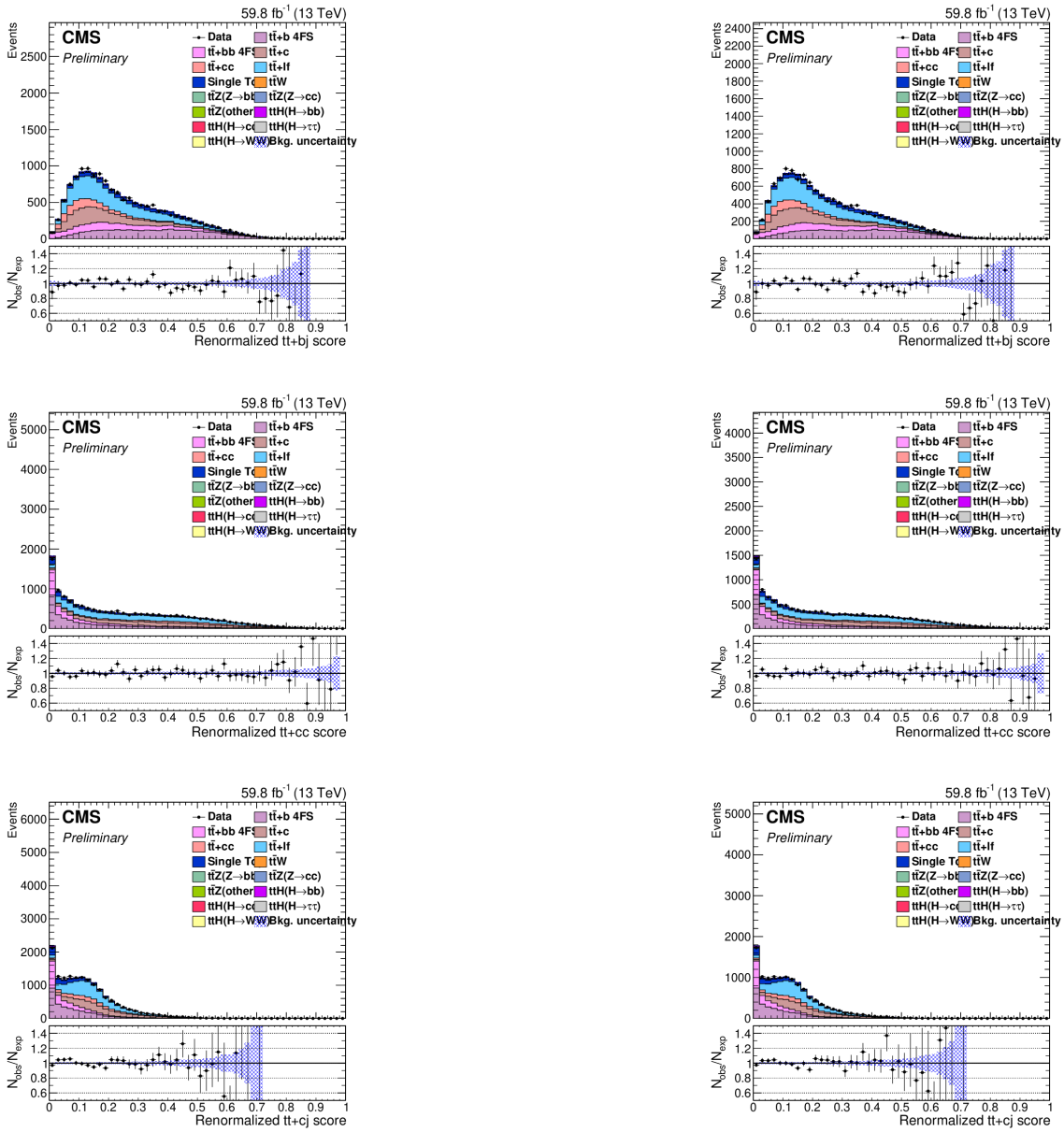


Figure 90: Comparison for Data/MC plots of ParT score ratios for $t\bar{t}+bj$ (top row), $t\bar{t}+c\bar{c}$ (second row), $t\bar{t}$+cj (bottom row) in the electron channel. **New: left. Old: right**

Muon

The following figures present a comparison of Data/MC plots for different lepton identification methods in the case that the lepton in the event is a muon. The comparisons are made for various kinematic variables and neural network scores. The left column shows the results using the new MVA-based lepton identification method, whereas the right column shows the results using the older cut-based lepton identification method.

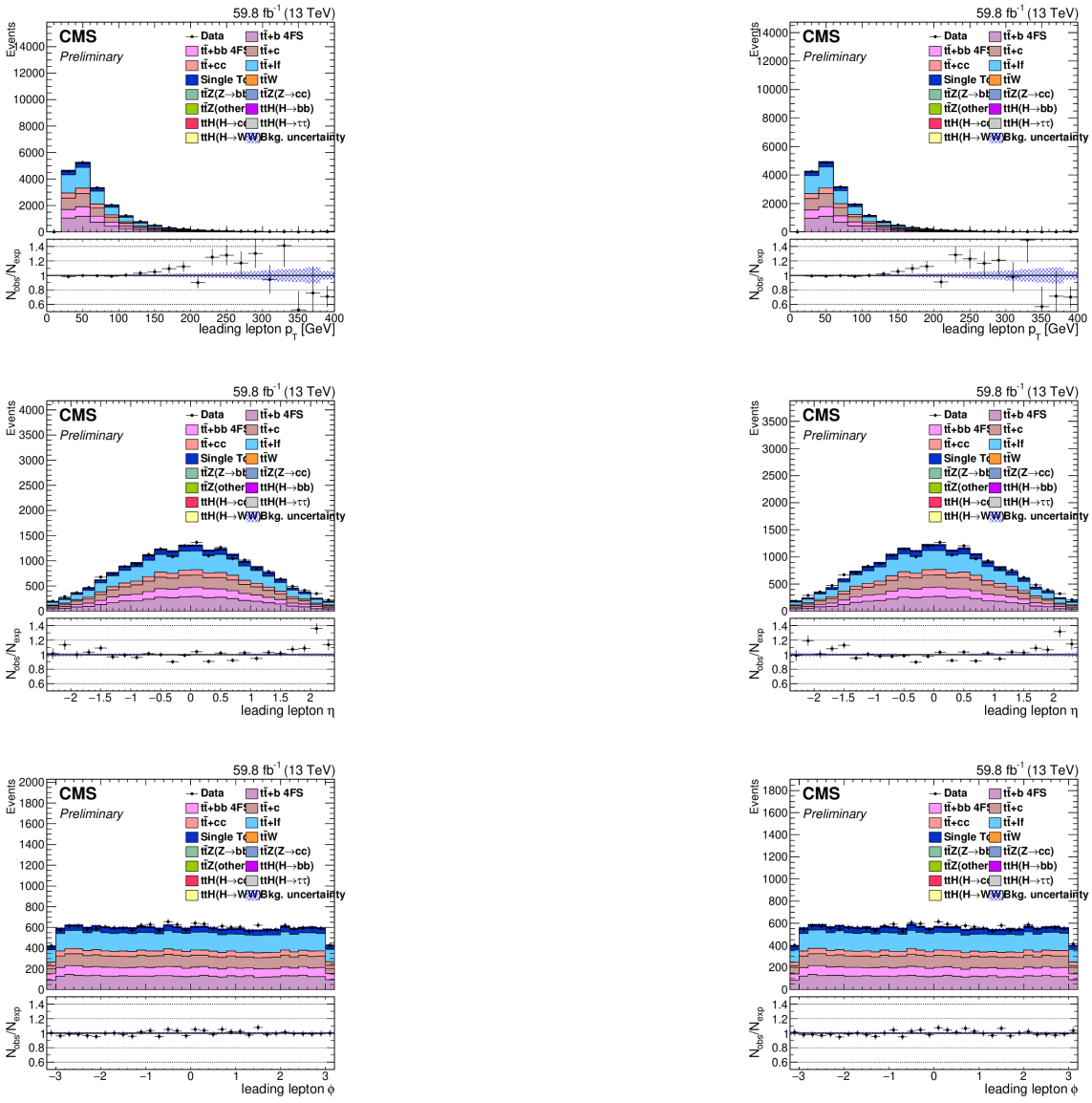


Figure 91: Comparison for Data/MC plots of muon transverse momentum (top row), muon pseudorapidity (middle row) and muon azimuthal angle (bottom row) in the muon channel. **New: left. Old: right**

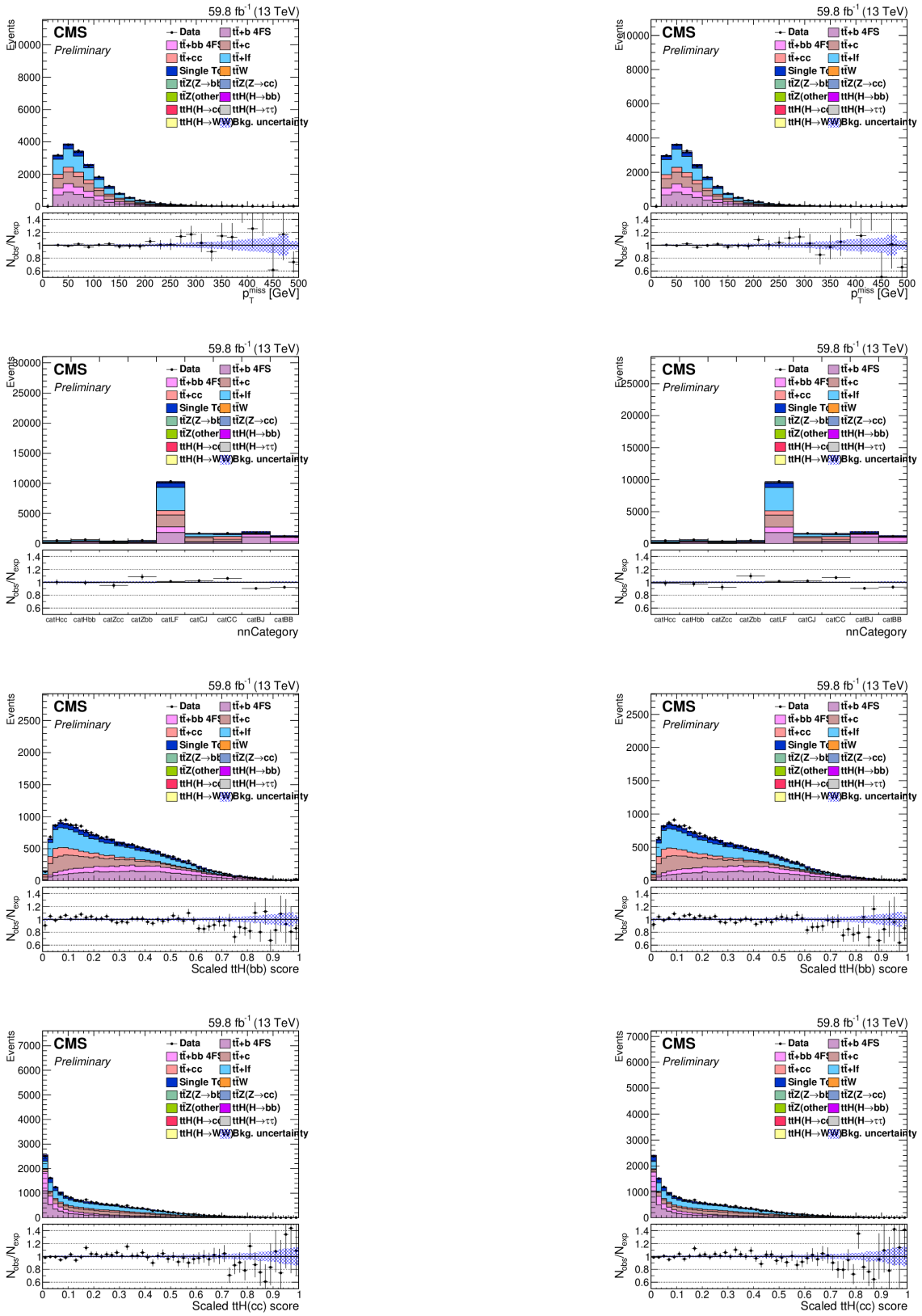


Figure 92: Comparison for Data/MC plots of MET (top row), NN category (second row), ParT score ratio for ttHbb (third row), and ParT score ratio for $t\bar{t}H(c\bar{c})$ (bottom row) in the muon channel. **New: left. Old: right**

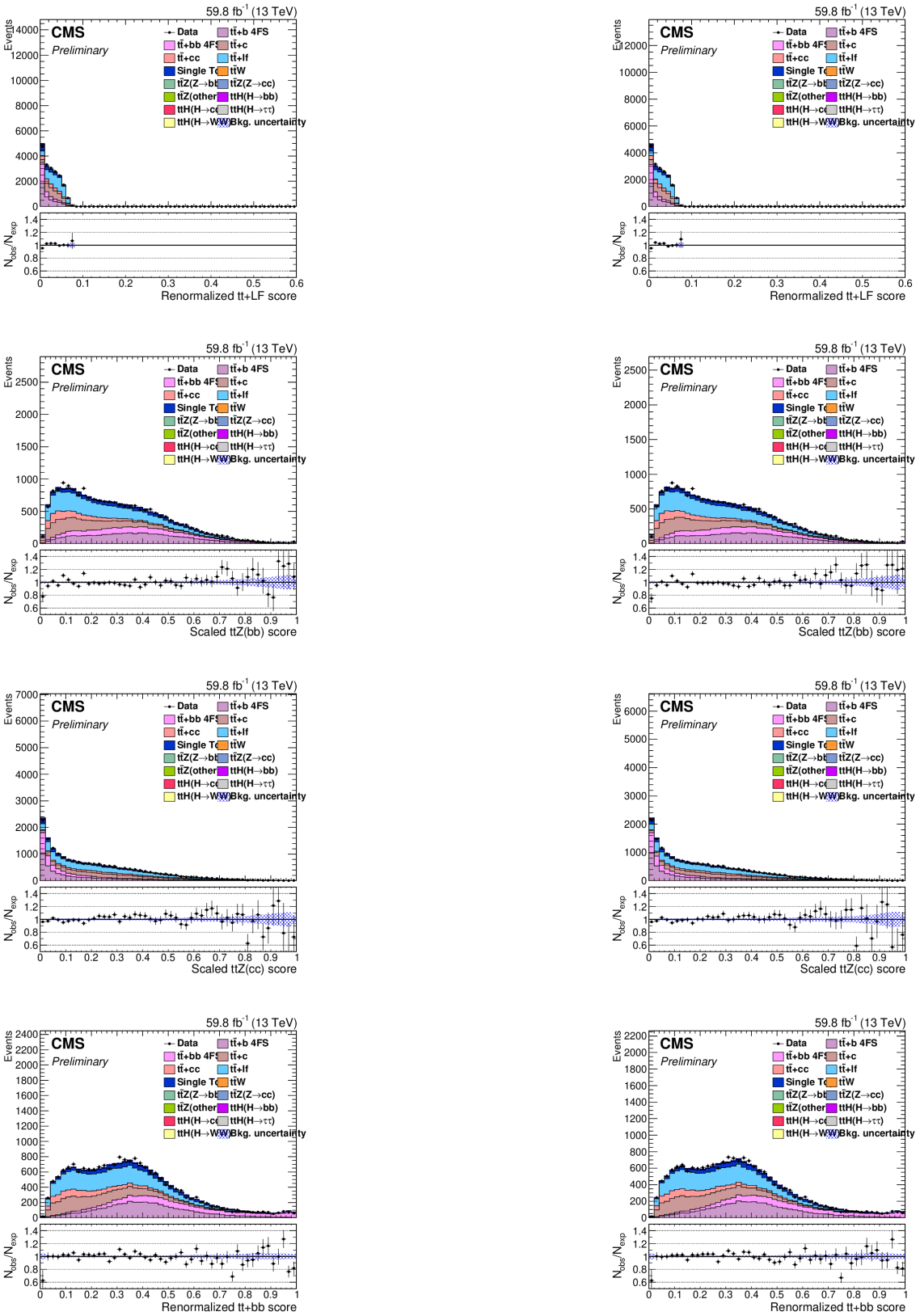


Figure 93: Comparison for Data/MC plots of ParT score ratios for $t\bar{t}$+LF (top row), $t\bar{t}Z(b\bar{b})$ (second row), $ttZ(c\bar{c})$ (third row) and $t\bar{t}+b\bar{b}$ (bottom row) in the muon channel. **New: left. Old: right**

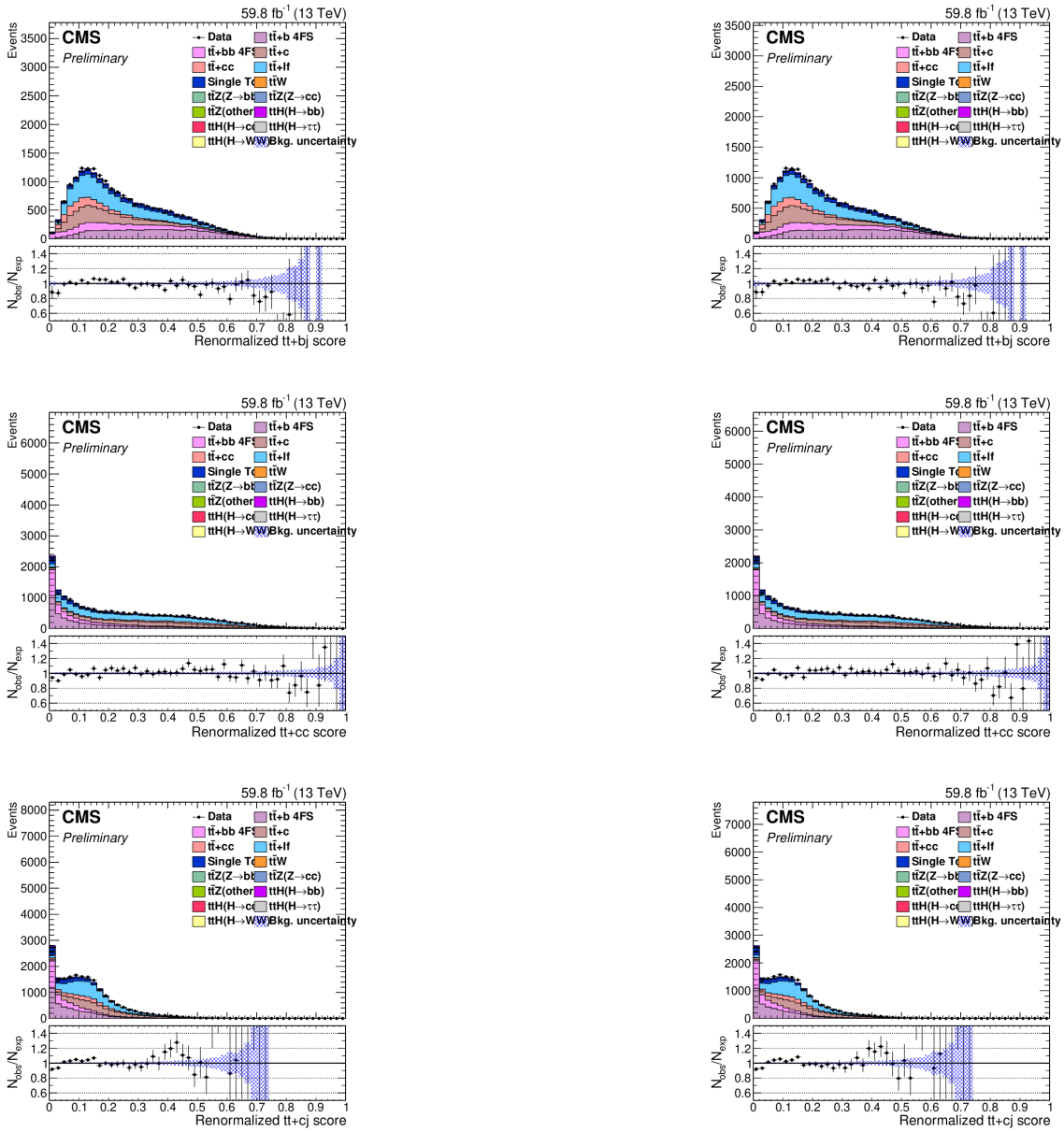


Figure 94: Comparison for Data/MC plots of ParT score ratios for $t\bar{t}$+bj (top row), $t\bar{t}+c\bar{c}$ (second row), $t\bar{t}$+cj (bottom row) in the muon channel. **New: left. Old: right**

The conclusions for the SL channel are consistent with those for the DL channel. The single-electron channel is more affected than the single-muon channel. In both cases, the score distributions remain unchanged, showing good agreement between Data and MC. This leads to the next, more critical test: assessing the impact of the ID change on statistical metrics such as significance and asymptotic limits.

6.4 Sensitivity and upper limits

In order to calculate the significance and the limits for the samples selected from the two different IDs to assess their strength, the COMBINE framework [96] was used to compute the asymptotic limits and the expected significance for Asimov data. COMBINE is a powerful statistical tool developed by the CMS Collaboration that integrates various techniques for hypothesis testing and confidence interval estimation in High-Energy Physics. Its primary aim is to facilitate the combination of different datasets and the implementation of complex statistical models to derive robust conclusions from experimental data.

In our case, the Asimov dataset represents an idealized scenario where all observations perfectly match the expected distributions predicted by the 4FS (Four-Flavor Scheme) framework. The 4FS is a theoretical scheme used in particle physics calculations that includes contributions from top quarks but neglects the bottom quark in the initial state. No additional systematic uncertainties were included as nuisance parameters in the fit, as these would shift the significance and the limits in the same direction for both cases. Significance, in this context, refers to the measure of how strongly the data deviates from the null hypothesis, typically expressed in units of standard deviations (σ). Limits, on the other hand, define the range within which the ratio r_{Hcc}, representing the observed Hcc rate relative to the Standard Model expectation, is expected to lie with a certain confidence level.

The metrics were calculated solely for the 2018 data, and the fit was performed in the signal region (SR). The limits and significance for the single lepton (SL) and dilepton (DL) channels are summarized in Tables 15, 16, 17, and 18 for the two different IDs.

Asymptotic Limits (CLs)	**Value of r_{Hcc}**	**Significance Value**
Expected 2.5%	$r_{\mathrm{Hcc}} < 13.0532$	Hcc: 0.0854424
Expected 16.0%	$r_{\mathrm{Hcc}} < 17.6027$	Hbb: 1.4561
Expected 50.0%	$r_{\mathrm{Hcc}} < 24.9375$	Zcc: 0.605751
Expected 84.0%	$r_{\mathrm{Hcc}} < 35.7723$	Zbb: 0.916042
Expected 97.5%	$r_{\mathrm{Hcc}} < 49.3861$	

Table 15: Asymptotic Limits (CLs) and Significance for DL - New ID

Asymptotic Limits (CLs)	**Value of $r_{\mathbf{Hcc}}$**	**Significance Value**
Expected 2.5%	$r_{\mathrm{Hcc}} < 14.0923$	Hcc: 0.079352
Expected 16.0%	$r_{\mathrm{Hcc}} < 19.0814$	Hbb: 1.35083
Expected 50.0%	$r_{\mathrm{Hcc}} < 27.1250$	Zcc: 0.558156
Expected 84.0%	$r_{\mathrm{Hcc}} < 39.0184$	Zbb: 0.844874
Expected 97.5%	$r_{\mathrm{Hcc}} < 53.9594$	

Table 16: Asymptotic Limits (CLs) and Significance for DL - Old ID

Asymptotic Limits (CLs)	**Value of $r_{\mathbf{Hcc}}$**	**Significance Value**
Expected 2.5%	$r_{\mathrm{Hcc}} < 7.0059$	Hcc: 0.154603
Expected 16.0%	$r_{\mathrm{Hcc}} < 9.4206$	Hbb: 3.72218
Expected 50.0%	$r_{\mathrm{Hcc}} < 13.1875$	Zcc: 1.05798
Expected 84.0%	$r_{\mathrm{Hcc}} < 18.6018$	Zbb: 2.23632
Expected 97.5%	$r_{\mathrm{Hcc}} < 25.2364$	

Table 17: Asymptotic Limits (CLs) and Significance for SL - New ID

Asymptotic Limits (CLs)	**Value of $r_{\mathbf{Hcc}}$**	**Significance Value**
Expected 2.5%	$r_{\mathrm{Hcc}} < 7.5256$	Hcc: 0.144986
Expected 16.0%	$r_{\mathrm{Hcc}} < 10.0791$	Hbb: 3.41288
Expected 50.0%	$r_{\mathrm{Hcc}} < 14.0625$	Zcc: 1.0013
Expected 84.0%	$r_{\mathrm{Hcc}} < 19.8921$	Zbb: 2.11085
Expected 97.5%	$r_{\mathrm{Hcc}} < 26.9496$	

Table 18: Asymptotic Limits (CLs) and Significance for SL - Old ID

From these results, it is evident that the new ID effectively suppresses the expected limits while simultaneously increasing the anticipated significance. In light of these outcomes, it appears that the new ID offers superior efficiency for this analysis, suggesting its potential to replace the existing ID.

7 Summary

In summary, the objective of this thesis was to develop a novel identification algorithm for leptons associated with tt+X production. Building upon an existing MVA model, we initially examined the input features to assess their standalone effectiveness for the identification method. Subsequently, by incorporating an additional variable for muons and created a new ID apart from the classifier's score, we evaluated the efficiency and purity of our proposed algorithm against an existing identification algorithm using an independent sample. Our findings indicated that our algorithm yielded superior results.

Furthermore, we explored the integration of our algorithm into an ongoing analysis, specifically a search for $H \to c\bar{c}$ associated with $t\bar{t}H$ production mode. Leveraging state-of-the-art techniques such as the ParticleNet algorithm for AK4 jets and the ParticleTransformer (ParT) architecture, we effectively captured correlations among physics objects, enabling robust discrimination of the $H \to c\bar{c}$ signal from backgrounds.

Upon observing a gain in event yields with the use of our new identification algorithm, particularly in channels associated with leptons, and implementing a set of scale factors to account for non-uniform gains, we found consistent Data/MC plots and significant improvements from a statistical perspective.

Future steps include the implementation of our identification algorithm in the ongoing analysis, either as a replacement for the existing algorithm or as an adjustment for Run3 data. Additionally, we may explore modifying the training features to create a new identification algorithm that relies solely on the MVA score, particularly for muons. Despite being an ongoing work with ample room for improvement, our results thus far are highly encouraging, indicating the strength and potential of our approach.

References

[1] Gaillard, Mary K., Paul D. Grannis, and Frank J. Sciulli. "The standard model of particle physics." Reviews of Modern Physics 71.2 (1999): S96.

[2] Giddings, Steven B. "The deepest problem: some perspectives on quantum gravity." arXiv preprint arXiv:2202.08292 (2022).

[3] Thomson, Mark, *"Modern particle physics"*, Cambridge University Press, 978-1-107-03426-6, 2013.

[4] Halzen, F. and Martin, Alan D. , *"QUARKS AND LEPTONS: An Introductory Course in Modern Particle Physics"*, 978-0-471-88741-6, 1984.

[5] C. S. Wu, E. Ambler, R. W. Hayward, D. D. Hoppes, and R. P. Hudson, "Experimental Test of Parity Conservation in Beta Decay," *Phys. Rev.* **105** (4), 1413–1415 (1957), doi: 10.1103/PhysRev.105.1413, url: https://link.aps.org/doi/10.1103/PhysRev.105.1413.

[6] Makoto Kobayashi, Toshihide Maskawa, CP-Violation in the Renormalizable Theory of Weak Interaction, Progress of Theoretical Physics, Volume 49, Issue 2, February 1973, Pages 652–657, https://doi.org/10.1143/PTP.49.652

[7] Jakob Schwichtenberg, *Physics from Symmetry*. Springer International Publishing, Cham, 2018.

[8] Huong Nguyen, *Search for the Standard Model Higgs Boson in Leptons plus Jets Final States*. Ph.D. thesis, University of Somewhere, Year of Publication: 2014. Available online: https://www.osti.gov/biblio/1155186.

[9] G. Aad et al. (ATLAS Collaboration), *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*. *Physics Letters B*, vol. 716, no. 1, Elsevier BV, ISSN: 0370-2693, 2012. http://dx.doi.org/10.1016/j.physletb.2012.08.020.

[10] S. Chatrchyan et al. (CMS Collaboration), *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*. *Physics Letters B*, vol. 716, no. 1, pp. 30-61, ISSN: 0370-2693, 2012. https://doi.org/10.1016/j.physletb.2012.08.021.

[11] S. Chatrchyan et al. (CMS Collaboration), *Observation of a new boson with mass near 125 GeV in pp collisions at* $\sqrt{s}$ *= 7 and 8 TeV*. *Journal of High Energy Physics*, vol. 2013, no. 6, Springer Science and Business Media LLC, ISSN

[12] G. Aad et al. (ATLAS Collaboration), *Measurements of the Higgs boson production and decay rates and coupling strengths using pp collision data at $\sqrt{s}$ = 7 and 8 TeV in the ATLAS experiment*. *The European Physical Journal C*, vol. 76, no. 1, Springer Science and Business Media LLC, ISSN: 1434-6052, 2016. http://dx.doi.org/10.1140/epjc/s10052-015-3769-y.

[13] G. Aad et al. (ATLAS Collaboration), *Evidence for the spin-0 nature of the Higgs boson using ATLAS data*. *Physics Letters B*, vol. 726, no. 1–3, Elsevier BV, ISSN: 0370-2693, 2013. http://dx.doi.org/10.1016/j.physletb.2013.08.026.

[14] V. Khachatryan et al. (CMS Collaboration), *Precise determination of the mass of the Higgs boson and tests of compatibility of its couplings with the standard model predictions using proton collisions at 7 and 8 TeV*. *The European Physical Journal C*, vol. 75, no. 5, Springer Science and Business Media LLC, ISSN: 1434-6052, 2015. http://dx.doi.org/10.1140/epjc/s10052-015-3351-7.

[15] V. Khachatryan et al. (CMS Collaboration), *Constraints on the spin-parity and anomalous HVV couplings of the Higgs boson in proton collisions at 7 and 8 TeV*. *Physical Review D*, vol. 92, no. 1, American Physical Society (APS), ISSN: 1550-2368, 2015. http://dx.doi.org/10.1103/PhysRevD.92.012004.

[16] M. Tanabashi et al. (Particle Data Group), *Review of Particle Physics*. *Phys. Rev. D*, vol. 98, no. 3, pp. 030001, American Physical Society, August 2018. https://link.aps.org/doi/10.1103/PhysRevD.98.030001.

[17] CMS Collaboration (Sirunyan, A. M. et al.), *Observation of $t\bar{t}H$ Production*, *Phys. Rev. Lett.*, vol. 120, no. 23, p. 231801, Jun. 2018. http://dx.doi.org/10.1103/PhysRevLett.120.231801

[18] CERN, *Higgs Cross Section Working Group (LHC HXSWG)*. CERN TWiki, URL: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/HiggsXSBR.

[19] A.M. Sirunyan et al. (CMS Collaboration), *Combined measurements of Higgs boson couplings in proton–proton collisions at $\sqrt{s}$ = 13 TeV*. *The European Physical Journal C*, vol. 79, no. 5, Springer Science and Business Media LLC, ISSN: 1434-6052, 2019. http://dx.doi.org/10.1140/epjc/s10052-019-6909-y.

[20] G. Aad et al. (ATLAS and CMS Collaborations), *Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s}$ = 7 and 8 TeV with the ATLAS and CMS Experiments*. *Physical Review Letters*, vol. 114, no. 19, American Physical Society (APS), ISSN: 1079-7114, 2015. http://dx.doi.org/10.1103/PhysRevLett.114.191803.

[21] S. Abachi et al., "Observation of the Top Quark," Physical Review Letters, vol. 74, pp. 2632–2637, Apr. 1995.

[22] S. W. Herb et al., "Observation of a Dimuon Resonance at 9.5-GeV in 400-GeV Proton-Nucleus Collisions," Physical Review Letters, vol. 39, pp. 252–255, 1977.

[23] M. Kobayashi and T. Maskawa, "CP Violation in the Renormalizable Theory of Weak Interaction," Progress in Theoretical Physics, vol. 49, pp. 652–657, 1973.

[24] K. Lannon, F. Margaroli, and C. Neu, *Measurements of the production, decay and properties of the top quark: a review*, *The European Physical Journal C*, vol. 72, no. 8, pp. 2120, Aug. 29, 2012.

[25] Behera, P. (2017). Measurement of the semileptonic $t\bar{t} + \gamma$ production cross section in pp collisions at $\sqrt{s} = 8$ TeV. *Journal of High Energy Physics*, 2017(10). doi:10.1007/JHEP10(2017)006

[26] Abdullin, Salavat, et al. "The fast simulation of the CMS detector at LHC." Journal ofPhysics: Conference Series. Vol. 331. No. 3. IOP Publishing, 2011

[27] CERN Bulletin, *The particle suppliers. Les fournisseurs de particules*. In: BUL-NA-2010-077. 14/2010 (2010), p. 03. URL: https://cds.cern.ch/record/1255151

[28] I. Efthymiopoulos, *Overview of the ATLAS detector at LHC*. Tech. rep. ATL-CONF-99-002. 7. version revised com-conf-99-002 version1. Geneva: CERN, 1999. URL: http://cds.cern.ch/record/409257

[29] S. Chatrchyan et al., *The CMS Experiment at the CERN LHC*. In: JINST 3 (2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004

[30] Jr. Alves A. Augusto et al., *The LHCb Detector at the LHC*. In: JINST 3 (2008), S08005. DOI: 10.1088/1748-0221/3/08/S08005

[31] K. Aamodt et al., *The ALICE experiment at the CERN LHC*. In: JINST 3 (2008), S08002. DOI: 10.1088/1748-0221/3/08/S08002

[32] CERN, *LS3 Schedule Change*, Webpage. Available at: https://hilumilhc.web.cern.ch/article/ls3-schedule-change.

[33] CMS Collaboration, *CMS Public TWiki: Luminosity Public Results - Multi-year plots*. https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults.

[34] Adolphi, Roman. "The CMS experiment at the CERN LHC." Jinst 803 (2008): S08004.

[35] T. Sakuma, *Cutaway diagrams of CMS detector*, General Photo, 2019. https://cds.cern.ch/record/2665537

[36] Unknown, *LaTeX Example: Spherical Coordinates in CMS*, Webpage. Available at: https://wiki.physik.uzh.ch/cms/latex:example_spherical_coordinates.

[37] Christian Thomay and Rudolf Frühwirth, *Überprüfung und Kalibrierung der Vertexrekonstruktion mit ersten CMS Daten*, Journal: Unknown, Volume: 32, Year: 2011.

[38] F. Kircher, Bruno Levesy, Y. Pabot, D. Campi, Benoit R. Cure, Alain Herve, I.L. Horvath, P. Fabbricatore, and Riccardo Musenich, "Status report on the CMS superconducting solenoid for LHC," IEEE Transactions on Applied Superconductivity, vol. 9, pp. 837–840, July 1999.

[39] CMS Collaboration, K.W. Bell, Caleb Brew, R.M. Brown, B. Camanzi, D.J.A. Cockerill, Jack Coughlan, N.I. Geddes, K. Harder, Shernice Harper, B.W. Kennedy, P. Murray, C.H. Shepherd-Themistocleous, I.R. Tomalin, J.H. Williams, W.J. Womersley, S.D. Worm, and Ryszard Romaniuk, "Precise mapping of the magnetic field in the CMS barrel yoke using cosmic rays," Journal of Instrumentation, vol. 5, p. T03021, Mar. 2010.

[40] R. Adolphi et al. (CMS Collaboration), *The CMS experiment at the CERN LHC*, JINST **3**, S08004 (2008), doi: 10.1088/1748-0221/3/08/S08004.

[41] CMS Collaboration, *Identifying Tracks with Silicon Pixels*, Webpage. Available at: https://cms.cern/detector/identifying-tracks/silicon-pixels.

[42] CMS Collaboration, *Silicon Strips - Identifying Tracks*, Webpage. Available at: https://cms.cern/detector/identifying-tracks/silicon-strips.

[43] Axer, Markus, *Development of a Test System for the Quality Assurance of Silicon Microstrip Detectors for the Inner Tracking System of the CMS Experiment*, Journal Article.

[44] Benaglia, Andrea. "The CMS ECAL performance with examples." Journal of Instrumentation 9.02 (2014): C02008.

[45] CMS Collaboration, *Performance and operation of the CMS electromagnetic calorimeter*, Journal of Instrumentation, 5(03):T03010, March 2010.

[46] C. D. Barney, *The CMS Electromagnetic Calorimeter: Its Performance and Role in the Discovery of the Higgs Boson and Perspectives for the Future*, CMS Conference Report, CMS CR-2013/410.

[47] Q. Ingram, *Energy Resolution of the Barrel of the CMS Electromagnetic Calorimeter*, Journal of Instrumentation, 2(04):P04004–P04004, April 2007.

[48] Freeman, J. "Innovations for the CMS HCAL." At the Leading Edge: The ATLAS and CMS LHC Experiments (2010): 259.

[49] CMS Collaboration, *Identification and Filtering of Uncharacteristic Noise in the CMS Hadron Calorimeter*, Journal of Instrumentation, 5(03):T03014–T03014, March 2010.

[50] CMS Collaboration, *Calibration of the CMS Hadron Calorimeters Using Proton-Proton Collision Data at* $\sqrt{s} = 13$ *TeV*, September 2019.

[51] Emlyn Corrin, *Development of Digital Readout Electronics for the CMS Tracker*, ResearchGate, April 2003.

[52] G. Abbiendi et al., "Study of the effects of radiation on the CMS Drift Tubes Muon Detector for the HL-LHC," CERN, arXiv:1912.06178 [physics.ins-det], CMS CR-2019/159, https://cds.cern.ch/record/2705998, doi:10.1088/1748-0221/14/12/C12010.

[53] CMS Collaboration et al., "Performance of the CMS drift tube chambers with cosmic rays," Journal of Instrumentation, vol. 5, no. 03, article T03015, 2010.

[54] CMS Collaboration, "Muon Drift Tubes," Webpage. Available at: https://cms.cern/detector/detecting-muons/muon-drift-tubes.

[55] CMS Experiment, "Resistive Plate Chambers," Webpage. Available at: https://cmsexperiment.web.cern.ch/index.php/detector/detecting-muons/resistive-plate-chambers.

[56] CMS Collaboration, "Performance of the CMS cathode strip chambers with cosmic rays," Journal of Instrumentation, vol. 5, no. 03, pp. T03018, March 2010.

[57] M. Ripert and M. Hohlmann, "Calibration of analog sensors for the alignment of muon chambers in the CMS experiment," 2005.

[58] CMS Collaboration, "Gas Electron Multiplier (GEM)," Retrieved from https://cms.cern/detector/detecting-muons/gas-electron-multiplier.

[59] Paolo Gunnellini, "The CASTOR calorimeter at the CMS experiment," arXiv preprint arXiv:1304.2943 (2013).

[60] Benoit Roland, "Forward Physics Capabilities of CMS with the CASTOR and ZDC detectors," arXiv preprint arXiv:1008.0592 (2010).

[61] Manfred Jeitler, "Upgrade of the trigger system of CMS," Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 718:11–15, 2013. Proceedings of the 12th Pisa Meeting on Advanced Detectors.

[62] Vardan Khachatryan et al., "The CMS trigger system," Journal of Instrumentation, 12(01):P01020, 2017. DOI: 10.1088/1748-0221/12/01/P01020. arXiv:1609.02366 [physics.ins-det].

[63] Sergio Cittolin, Attila Rácz, and Paris Sphicas, "CMS The TriDAS Project: Technical Design Report, Volume 2: Data Acquisition and High-Level Trigger," CMS trigger and data-acquisition project. Technical Design Report CMS. Geneva: CERN, 2002. http://cds.cern.ch/record/578006.

[64] CMS Collaboration, *Particle-flow reconstruction and global event description with the CMS detector*, Journal of Instrumentation, 12(10):P10003–P10003, Oct. 2017, https://doi.org/10.1088/1748-0221/12/10/P10003.

[65] A. M. Sirunyan et al., *Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC*, Journal of Instrumentation, vol. 16, no. 05, p. P05014, May 2021, doi: 10.1088/1748-0221/16/05/P05014.

[66] The CMS Collaboration, *Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at* $\sqrt{s}$ = 13 *TeV*, Journal of Instrumentation, 13(06):P06015–P06015, Jun. 2018, https://doi.org/10.1088/1748-0221/13/06/P06015.

[67] R. Frühwirth, *Application of Kalman filtering to track and vertex fitting*, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 262(2):444–450, 1987, https://doi.org/10.1016/0168-9002(87)90887-4.

[68] CMS Collaboration. (2018). Enhancing the muon with the Compact Muon Solenoid. Retrieved from https://cms.cern/news/enhancing-muon-compact-muon-solenoid#:~:text=An%20algorithm%20called%20%22TuneP%22%20picks,some%20of%20the%20muon%20measurements

[69] CMS Collaboration. (2012). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, 716(1). doi:10.1016/j.physletb.2012.08.021

[70] CMS Collaboration. (2020). A measurement of the Higgs boson mass in the diphoton decay channel. *Physics Letters B*, 805, 135425. doi:10.1016/j.physletb.2020.135425

[71] CMS Collaboration. (2021). Evidence for Higgs boson decay to a pair of muons. *Journal of High Energy Physics*, 2021(1). doi:10.1007/JHEP01(2021)148

[72] ATLAS Collaboration. (2018). Search for the Decay of the Higgs Boson to Charm Quarks with the ATLAS Experiment. *Physical Review Letters*, 120(21), 211802. doi:10.1103/PhysRevLett.120.211802. arXiv:1802.04329

[73] CMS Collaboration. (2020). A search for the standard model Higgs boson decaying to charm quarks. *Journal of High Energy Physics*, 2020(03), 131. doi:10.1007/JHEP03(2020)131. arXiv:1912.01662

[74] H. Q. Loukas Gouskos and J. Incandela, *Search for the Higgs boson decaying to a pair of charm quarks, produced in association with a vector boson, using large radius jets*, CMS Physics Analysis Note AN-18-243, 2019.

[75] M. Cacciari, G. P. Salam, and G. Soyez. (2008). The anti-k_T jet clustering algorithm. *Journal of High Energy Physics*, 2008(04), 063. doi:10.1088/1126-6708/2008/04/063. arXiv:0802.1189

[76] Loukas Gouskos, Huang Huang, Joseph Incandela, Jan Kieseler, Qiang Li, Matthias Mozer, Huilin Qu, Paris Sphicas, Markus Stoye, Mauro Verzetti, *Boosted jet identification with particle-level information and deep neural networks*, CMS Physics Analysis Note AN-18-107, 2018.

[77] H. Qu and L. Gouskos, *ParticleNet: Jet Tagging via Particle Clouds*, Phys. Rev. D 101 (2020), no. 5, 056019, doi: 10.1103/PhysRevD.101.056019, arXiv:1902.08570.

[78] Wang, Yue, Sun, Yongbin, Liu, Ziwei, Sarma, Sanjay E., Bronstein, Michael M., Solomon, Justin M. (2019). Dynamic Graph CNN for Learning on Point Clouds. arXiv:1801.07829 [cs.CV]

[79] Yang, Dongxu, Liu, Yadong, Sun, Jianxiang. (2020). GDCNN: A Gated CNN with dilated convolution for decoding steering directions from EEG. In *2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)* (Vol. 9, pp. 952-957). doi:10.1109/ITAIC49862.2020.9338860

[80] Huilin Qu, Congqiao Li, and Sitian Qian, *Particle Transformer for Jet Tagging*, arXiv preprint arXiv:2202.03772, 2024.

[81] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. (2021). Going deeper with Image Transformers. arXiv:2103.17239 [cs.CV]

[82] CMS Collaboration, "Baseline muon selections for Run-II," *Twiki*, https://twiki.cern.ch/twiki/bin/view/CMS/SWGuideMuonIdRun2, r54.

[83] CMS Collaboration, "Rochester Correction," *Twiki*, https://twiki.cern.ch/twiki/bin/view/CMS/RochcorMuon, r29.

[84] CMS Collaboration, "Cut Based Electron ID for Run 2," *Twiki*, https://twiki.cern.ch/twiki/bin/view/CMS/CutBasedElectronIdentificationRun2, r60.

[85] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth, 1984.

[86] Y. Coadou, *Boosted Decision Trees*, Artificial Intelligence for High Energy Physics, WORLD SCIENTIFIC, 2022, pp. 9–58. DOI: http://dx.doi.org/10.1142/9789811234033_0002. ISBN: 9789811234033.

[87] C. Gini, *Variabilità e mutabilità*, (reprinted in *Memorie di Metodologica Statistica*, eds. E. Pizetti and T. Salvemini, Libreria Eredi Virgilio Veschi, Rome, 1955), 1912.

[88] GeeksforGeeks, "Boosting in Machine Learning - Boosting and AdaBoost", URL: https://www.geeksforgeeks.org/boosting-in-machine-learning-boosting-and-adaboost/

[89] L. Breiman, *Prediction Games and Arcing Algorithms*, Neural Comput., vol. 11, pp. 1493, 1999.

[90] J. H. Friedman, *Greedy function approximation: A gradient boosting machine*, Ann. Statist., vol. 29, pp. 1189, 2001.

[91] L. Breiman, *Arcing the Edge*, Ann. Prob., vol. 26, pp. 1683, 1998.

[92] CMS Collaboration, *Observation of four top quark production in proton-proton collisions at* $\sqrt{s} = 13\,TeV$, Physics Letters B, vol. 847, Dec. 2023, p. 138290, ISSN: 0370-2693, DOI: 10.1016/j.physletb.2023.138290, URL: http://dx.doi.org/10.1016/j.physletb.2023.138290.

[93] ROOT Collaboration. (n.d.). TMVA User's Guide. Retrieved from https://root.cern.ch/download/doc/tmva/TMVAUsersGuide.pdf

[94] T. Chen and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, ACM, Aug. 2016. doi: 10.1145/2939672.2939785.

[95] W. Adam, R. Frühwirth, A. Strandlie, and T. Todorov. (2005). Reconstruction of electrons with the Gaussian sum filter in the CMS tracker at LHC. *J. Phys. G*, 31(9), N9. doi:10.1088/0954-3899/31/9/N01. arXiv:0306087

[96] Hayrapetyan, Aram et al., "The CMS statistical analysis and combination tool: Combine", arXiv:2404.06614, arXiv.org, 2024, CMS-CAT-23-001, CERN-EP-2024-078, Submitted to *Comput. Softw. Big Sci.*