Università degli Studi di Bari Aldo Moro
Dipartimento interateneo di fisica Michelangelo Merlin
PhD in Physics

Development of jet identification and flavor tagging algorithms for CMS experiment

Supervisors:
Chiar.ma Prof.ssa Anna Colaleo
Dott. Luigi Longo

PhD Coordinator:
Chiar.mo Prof. Vincenzo Spagnolo

PhD candidate:
Donato Troiano

Firmatario: Donato Troiano
Data: 21/07/2026

XXVIII cicle
Bari - 22/05/2026

Contents

Introduction

At the Large Hadron Collider (LHC), heavy resonances such as top quarks, W and Z bosons and the Higgs boson are frequently produced with large Lorentz boosts. In such conditions, when a boosted resonance decays hadronically, its decay products become highly collimated in the laboratory frame and are often reconstructed as a single, large-radius jet (or *fat* jet). The identification of the fat jet mother particle plays a central role in improving the sensitivity to physics processes. Indeed, a large number of Standard Model (SM) measurements, as well as searches for physics beyond the SM, rely on the reconstruction of hadronically decaying particles produced with large transverse momentum (p_{T}). In order to address this challenge, dedicated algorithms—commonly referred to as taggers—have been developed and continuously refined over the years. These algorithms aim to exploit the distinctive features of jets originating from boosted heavy particles. In particular, heavy-resonance taggers are designed not only to discriminate jets produced by the decay of boosted resonances from generic jets arising from Quantum Chromodynamics (QCD) processes, but also to classify different decay modes. For instance, they allow the identification of heavy-flavor decays involving $b-$ and $c-$quarks and their separation from light-quarks decays (u, d, and s). To achieve this, tagging algorithms combine information from the global kinematic properties of the jet with detailed features of its internal structure, such as the distribution of its constituents and the presence of reconstructed secondary vertices (SVs).

CMS jet reconstruction is based on clustering algorithms applied to reconstructed particle-flow (PF) candidates made of calorimeter clusters and charged-particle tracks. Small-radius jets typically use the anti-k_T algorithm with radius $R \sim 0.4$ (AK4), while fat jets require $R \sim 0.8$ (AK8) to contain the full decay products. In recent years, AK8 tagging strategies evolved from cut-based approaches, using a limited number of substructure variables, to multivariate and machine learning (ML) based discriminants that combine many inputs. Modern taggers often merge high-level observables with low-level information about PF constituents and SVs to reach superior performance.

In CMS, ParticleNet (PNet) is a deep-learning tagger that treats a jet as an unordered set of constituent particles, a *point cloud*, and operates on per-constituent feature vectors (e.g. p_T, pseudorapidity, azimuthal angle, electric charge, and SV-related features). Architecturally, PNet uses graph- and edge-convolution layers that are permutation-invariant and can learn local neighborhood relations among constituents, effectively capturing the multi-prong patterns expected from boosted decays. The AK8 PNet implementation used in CMS (PNet-MD) is furthermore trained on flat p_T-mass distributions to minimize dependence on the jet p_T and mass and, consequently, to not bias the selection towards a specific resonance. PNet-MD is, in particular, configured to separate QCD jets from jets originating from hadronic decays of massive resonances (X), with a further classification on decay mode products ($X \to b\bar{b}$, $X \to c\bar{c}$, $X \to$light quarks). Thanks to its innovative architecture and constituent-level processing, PNet-MD proved to be the most performant AK8 tagging algorithm during Run 2 of CMS and, at the start of Run 3, became the state of the art for AK8 tagging both at the fully reconstructed (*offline*) level and at the trigger (*online*) level.

At the beginning of Run 3, the offline PNet-MD tagger was retrained using simulated samples reflecting the final Run 2 detector and running conditions. Before being deployed in physics analyses, the new Run 3 PNet-MD discriminator must therefore be carefully validated to ensure that its response in collision data is accurately reproduced by simulation. During my PhD, I was responsible for validating the PNet-MD score using the first available Run 3 data. The agreement between data and simulation was studied in a region enriched in events with a topology compatible with $Z \to b\bar{b}$ production in association with a recoil jet. In this region, the SM expectation is dominated by $Z \to q\bar{q}$ and $W \to q\bar{q}'$ processes, together with a large background contribution from QCD multijet production. This topology is particularly well suited for validating the PNet-MD discriminator, as applying increasingly tighter selections on the PNet-MD score is expected to enhance the Z boson mass peak. Events were therefore categorized into four regions corresponding to increasing values of the PNet-MD score. The progressive enhancement of the $Z \to b\bar{b}$ signal was studied together with the agreement between the SM prediction and the observed data. While the $Z \to q\bar{q}$ and $W \to q\bar{q}'$ contributions were taken from simulation, the QCD multijet background was estimated using data-driven techniques.
Concerning the online PNet-MD model, although PNet-MD significantly mitigated the dependence on the jet mass and p_T, residual correlations were still present and a new training campaign was conducted to minimize remaining biases and to ensure stable performance across an extended phase space. In this context, I was responsible for training a new online implementation of PNet-MD designed to be as independent as possible from the jet mass and p_T. The model was trained using QCD simulation as background against a signal model where a graviton-like resonance X decays into a Higgs bosons pair ($X \to HH$), with X and H masses spanning on a wider range that allow to not let the training learn a specific signal mass for the AK8 jet associated to the Higgs boson mass. In particular the model was trained to identify AK8 from QCD background generated from 2, 1, 0 heavy flavor quarks and fat jet from Higgs decaying into $b\bar{b}$, $c\bar{c}$, gg, $\tau\bar{\tau}$ and light quarks. The new trained model performances were compared with the model available at trigger level on a CMS data collected in 2025. This comparison was carried out to evaluate the performance improvements, investigate the potential inclusion of the new model in the CMS trigger menu, and establish a strategy for mass-p_T model-independent training of next High-Luminosity LHC era.
The thesis is structured in 5 chapters. **Chapter 1** describes the theoretical framework of the Standard Model and the CMS state of art at CMS for H and Z boson measurements. **Chapter 2** describes the experimental frameworks: the LHC and the detectors of the CMS apparatus. **Chapter 3** explains the algorithms used in CMS for the reconstruction and identification of the main physics objects, focusing in particular on algorithms for jet-flavor tagging. **Chapter 4** describes the offline PNet-MD validation, describing the event selection, the QCD estimate and showing the comparison of data - expected SM. **Chapter 5** describes the new trained online PNet-MD model, how the training sample was refined in order to reduce the model dependence on p_T and mass, the comparison with the current online model and the tests carried out to include the model in a trigger selection for the CMS 2026 data-taking.

Chapter 1

The standard model

The Standard Model (SM) of particle physics is a gauge field theory that was developed between the late 1960s and the early 1970s [1]. It provides a unified description of three of the four known fundamental forces of nature (the electromagnetic (EM), weak, and strong interactions) and classifies all known elementary particles. Gravity, the fourth fundamental force, is not included in the SM and remains described by general relativity. The formalism of the SM was completed with the experimental discovery of the Higgs boson in 2012 [2, 3] by the Compact Muon Solenoid (CMS) [4] and A Toroidal LHC ApparatuS (ATLAS) [5] collaborations, which confirmed the mechanism responsible for the generation of particle masses.

1.1 Standard model of particle physics

1.1.1 Elementary particles

The SM of particle physics classifies all known elementary particles into two broad categories according to their intrinsic spin: fermions, which carry half-integer spin and constitute matter, and gauge bosons, which carry integer spin and mediate the fundamental interactions. Elementary fermions, in turn, are divided into quarks and leptons. Leptons interact only via electroweak (EW) interactions and can be found in nature as free particles. Quarks interact via both strong and EW interactions and are observed in hadrons, either mesons which are bound states of one quark and one anti-quark or baryons which are bound states of three (anti)quarks. The 3 massive gauge bosons Z, W^+ and W^- are the responsible of the weak interactions, the massless photon (γ) for the EM force, and the eight massless gluons (g) for the strong force. Table 1.1 shows a summary of the gauge bosons and their properties.
Ordinary matter is composed of the lightest quarks and the lightest charged lepton. In the SM these building blocks are the up and down quarks, u and d, together with the electron e. Historically, experiments revealed additional species of fermions beyond u, d and e. The observed fermions arrange themselves into three families (or generations), each generation containing a charged lepton, its associated neutrino (ν), and a pair of quarks. Only the first generation appears with any appreciable abundance in ordinary matter because the heavier charged leptons (μ, τ) and the strange, charm, bottom and top quarks (s, c, b, t) are unstable and decay via the weak interaction into lighter states. The t quark mean lifetime is $\sim 10^{-25}$ s, much shorter than typical hadronization timescales of $\sim 10^{-24}$ s, and for this reason it was traditionally assumed not to form bound states. However, recent analyses by the CMS and ATLAS collaborations observed an excess in the $t\bar{t}$ production rate

Gauge bosons				
Name	Mass [GeV]	Charge $[e]$	Spin	Mediates
γ	0	0	1	EM
W	80.37	± 1	1	Weak
Z	91.19	0	1	Weak
g	0	0	1	Strong

Table 1.1: Overview of the gauge bosons in the SM, reporting their masses, electric charges, spin, and the fundamental interaction they mediate [6].

near threshold that is consistent with the formation of a short-lived, color-singlet quasi-bound $t\bar{t}$ system [7, 8]. Fermion classifications are reported in Tables 1.2 and 1.3.

Leptons			
Flavour	Mass $[MeV]$	Charge $[e]$	Spin
ν_e	$< 1.1\ 10^{-6}$	0	1/2
e	0.511	-1	1/2
ν_μ	< 0.19	0	1/2
μ	105.66	-1	1/2
ν_τ	< 18.2	0	1/2
τ	1777	-1	1/2

Table 1.2: Overview of the leptons in the SM, grouped by families, reporting their masses, electric charges and spin [6].

Quarks			
Flavour	Mass $[GeV]$	Charge $[e]$	Spin
u	0.002	$+2/3$	1/2
d	0.005	$-1/3$	1/2
c	1.27	$+2/3$	1/2
s	0.093	$-1/3$	1/2
t	173	$+2/3$	1/2
b	4.18	$-1/3$	1/2

Table 1.3: Overview of the quarks in the SM, grouped by families, reporting their masses, electric charges and spin [6].

The interactions between elementary particles are described by a Lagrangian density $\mathscr{L}$ (called Lagrangian for simplicity) invariant under the symmetry group $SU(3)_C \bigotimes SU(2)_L \bigotimes U(1)_Y$, where the indices stand for color, weak isospin, and weak hypercharge respectively. The strong interaction is described by the quantum chromodynamics (QCD), a theory based on the colour group $SU(3)_C$. The electromagnetism and the weak interactions are unified in the EW theory based on the weak isospin and weak hypercharge group $SU(2)_L \bigotimes U(1)_Y$.

1.1.2 Quantum chromodynamics

The QCD theory defines the interactions between quarks and gluons, governed by the symmetry group $SU(3)_C$. It is a non-abelian group with 8 generators representable by the 8 Gell-Mann matrices $\frac{\lambda_a}{2}$, $a = 1, ..., 8$ [1]. Quantum number associated to this group is called colour and it can take 3 values: blue, red and green. The Lagrangian describing not-interacting quarks (represented as a four-vector), for the different quark flavour j, can be written as:

$$\mathscr{L}_{QCD} = \sum_{j=1}^{f} \bar{\mathbf{q}}^j(x)(i\gamma^\mu \partial_\mu - m_j)\mathbf{q}^j(x) \tag{1.1}$$

where γ^μ are the Dirac matrices, $\mathbf{q}^j$ are the quark fields and m_j are the quark masses. Under this local gauge symmetry, the quark fields linear transformation can be expressed as follow:

$$\mathbf{q}^j(x) \to \mathbf{U}(x)\mathbf{q}^j(x) \quad (j = 1, ..., f) \tag{1.2}$$

$$\text{with } \mathbf{U}(x) = exp\left[i \sum_{a=1}^{8} \alpha^a(x) \frac{\lambda_a}{2}\right] \tag{1.3}$$

where $\alpha^a(x)$ are the eight local parameters of the transformation. The Lagrangian (1.1) is not invariant under (1.2). In order to make $\mathscr{L}_{QCD}$ invariant under a local transformation, ∂_μ has to be replaced with the covariant derivative:

$$\mathbf{D}_\mu = \partial_\mu + ig_s\mathbf{G}_\mu(x) \tag{1.4}$$

$$\text{with } \mathbf{G}_\mu(x) = G^a_\mu(x)\frac{\lambda_a}{2} \tag{1.5}$$

where g_s is the strong coupling constant and the 8 gauge vector fields G^a_μ correspond to 8 gauge bosons (gluons). The gluon fields are transformed as

$$\mathbf{G}_\mu \to \mathbf{U}(x)\mathbf{G}_\mu\mathbf{U}^\dagger(x) - \frac{i}{g_s}\mathbf{U}(x)\partial_\mu\mathbf{U}^\dagger(x) \tag{1.6}$$

Finally, with the above substitutions, the gauge invariant $\mathscr{L}_{QCD}$ can be written as:

$$\mathscr{L}_{QCD} = -\frac{1}{4}G^a_{\mu\rho}(x)G^{a\mu\rho}(x) + \sum_{j=1}^{f} \bar{\mathbf{q}}^j(x)\left[i\gamma^\mu\left(\partial_\mu + ig_sG^a_\mu(x)\frac{\lambda_a}{2}\right) - m_j\right]\mathbf{q}^j(x) \tag{1.7}$$

$$\text{with } G^a_{\mu\rho}(x) = \partial_\mu G^a_\rho(x) - \partial_\rho G^a_\mu(x) - g_s f_{abc}G^b_\mu(x)G^c_\rho(x) \tag{1.8}$$

where f_{abc} are the structure constants of the $SU(3)$ group[1]. The first term in (1.7) introduces trilinear and quadrilinear terms that correspond to interactions among three and four gluons, the second term is the free Lagrangian from (1.1) with the interaction term between quark and gluon fields.

[1] $f_{123} = 1$, $f_{147} = f_{246} = f_{257} = f_{345} = 1/2$, $f_{156} = f_{367} = -1/2$, $f_{458} = f_{678} = \sqrt{3}/2$

1.1.3 Jet factorization in proton-proton collisions

The proton is not an elementary particle since it is constituted by partons (quarks and gluons), each one carrying a fraction of the energy of the proton. The cross section for proton A + proton B $\rightarrow$ parton c + anything else (X) can be expressed (when produced at high energy) as [9]:

$$d\sigma_{AB\rightarrow c+X}(P_A, P_B, P_c) = \sum_{a,b} \int_0^1 dx_a \int_0^1 dx_b \; f_{a/A}(x_a, Q^2)\, f_{b/B}(x_b, Q^2) \\ \times\; d\hat{\sigma}_{ab\rightarrow c+X}\big(x_a P_A,\; x_b P_B,\; P_c,\; Q^2\big) \tag{1.9}$$

where P_A, P_B, P_c are respectively the four-momenta of the incident protons and the scattered parton, x_a (x_b) is the fraction of momentum carried by the parton a (b) of the proton A (B), Q^2 is the "hard scale" and corresponds to the square of the invariant mass of the scattering partons, $f_{a/A}(x_a, Q^2)$ ($f_{b/B}(x_b, Q^2)$) is the parton distribution function (PDF) corresponding to the probability of having a parton a (b) with x_a (x_b) and $d\hat{\sigma}_{ab\rightarrow c+X}\big(x_a P_A,\; x_b P_B,\; P_c,\; Q^2\big)$ (parton cross section) is the perturbative cross section for the scattering of partons a and b producing a parton c (plus other products indicated by X). For hard parton scatterings $d\hat{\sigma}_{ab\rightarrow c+X}$ can be evaluated via a perturbative expansion at different accuracy in $\alpha_s \equiv \frac{g_s^2}{4\,\pi}$.
However, the strong interaction does not allow coloured final-state objects: quarks and gluons produced in a hard scattering are never observed in isolation. To describe the colored particles confinement, a commonly used phenomenological model for the static quark–antiquark potential is the Cornell potential

$$V(r) = -\frac{4\alpha_s(Q^2)}{3r} + kr \tag{1.10}$$

where the short-distance Coulomb-like term $-\frac{4\alpha_s(Q^2)}{3r}$ encodes one-gluon exchange and the long-distance linear term kr (with string tension $k \simeq 1GeV/fm$) models confinement. This potential is a useful approximation for static or heavy quark systems (*e.g.* charmonium and bottomonium) and is supported by lattice QCD calculations [10]; it becomes numerically accurate when relativistic corrections and light-quark pair creation contribution are small. The corrections are important for light quarks, relativistic dynamics and pair creation and the simple potential picture must be replaced by full dynamical (lattice or effective-field) treatments.
The full space–time picture relevant for collider physics is the following: a hard scattering creates a high-energy parton at scale Q^2; that parton radiates additional quarks and gluons in a perturbative parton shower. An example of parton shower is shown in Figure 1.1. The perturbative shower description is valid while the transverse momenta of emissions remain well above the non-perturbative scale $\sim O(0.2GeV)$; when the shower evolution reaches scales of order a few hundred MeV the coupling becomes strong and perturbation theory breaks down. At that point non-perturbative hadronization models (string fragmentation, cluster hadronization, etc.) take over and convert colored partons into colourless hadrons [12]. The whole sequence hard process $\rightarrow$ perturbative shower $\rightarrow$ non-perturbative hadronization produces a collimated spray of observable particles that we call a jet. A jet is therefore not a single particle but a collection of final-state hadrons (charged pions, kaons, protons, neutral hadrons, photons from π^0 decay and occasionally leptons) whose momenta cluster in direction and whose energy flow traces back to an initiating parton. Details about jet reconstructions are provided in Section 3.8.

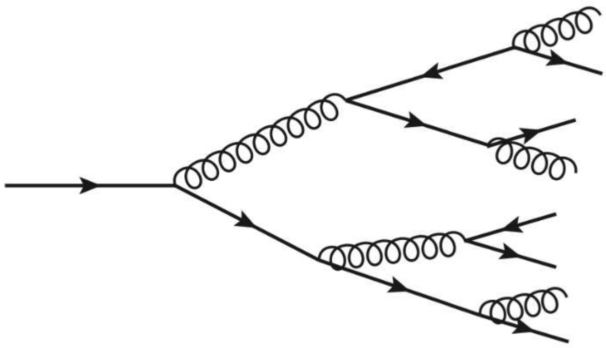

Figure 1.1: Schematic illustration of a typical parton shower in a high-energy limit of QCD, showing successive branchings of an energetic parton into lower-energy quarks and gluons. This diagram visualizes the multiple emission processes that contribute to the development of a jet [11].

1.1.4 Electroweak interaction

The electroweak theory defines the interactions between fermions and four gauge bosons (Z, W^+,W^- and γ), governed by the $SU(2)_L \bigotimes U(1)_Y$ symmetry group [1]. Quantum numbers associated to this group are the weak isospin (I_3) and the weak hypercharge (Y), directly linked to the electric charge by the relation $Q = I_3 + \frac{Y}{2}$. In the EW theory, fermions of opposite chirality[2] have different interactions. The $SU(2)_L$ left-handed doublets are denoted as $\mathbf{L} = \begin{pmatrix} \psi_L(x) \\ \psi'_L(x) \end{pmatrix}$ while the right-handed singlets as $\psi_R(x)$, $\psi'_R(x)$ where

$$\begin{aligned} \psi_L(x) &= \frac{1}{2}\,(1-\gamma_5)\,\psi(x) \\ \psi_R(x) &= \frac{1}{2}\,(1+\gamma_5)\,\psi(x) \end{aligned} \tag{1.11}$$

where $\psi(x)$ and $\psi'(x)$ are the fermionic fields. A single family of lepton can be described (assuming only massless and left-handed neutrinos) as

$$\mathbf{L} = \begin{pmatrix} \nu_\ell(x) \\ \ell^-(x) \end{pmatrix}_L \qquad \psi_R = 0 \quad \psi'_R = \ell^-_R(x) \tag{1.12}$$

where $\ell(x)$ indicates leptons e, μ, τ; for quarks

$$\mathbf{L} = \begin{pmatrix} u(x) \\ d(x) \end{pmatrix}_L \qquad \psi_R = u_R(x) \quad \psi'_R = d_R(x) \tag{1.13}$$

where $u(x)$ indicates quarks u, c, t and $d(x)$ indicates quarks d, s, b. The $SU(2)_L$ left-handed doublets carry weak isospin, in particular their members are assigned values $I_3 = +1/2$ (upper) and $I_3 = -1/2$ (lower), whereas $SU(2)_L$ right-handed singlets have $I_3 = 0$. The quantum numbers associated to the $SU(2)_L$ fermionic fields are reported in Table 1.4.

[2] The left and right chiral components of a field are defined from the $\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3\gamma^4$ matrix. This is used to define left and right chirality projection operators as $\frac{1-\gamma^5}{2}$ and $\frac{1+\gamma^5}{2}$, respectively.

Field	I_3	Q	Y
$(\nu_\ell, \ell)_L$	$(+\frac{1}{2}, -\frac{1}{2})$	$(0, -1)$	$(-1, -1)$
ℓ_R	0	-1	-2
$(u, d)_L$	$(+\frac{1}{2}, -\frac{1}{2})$	$(+\frac{2}{3}, -\frac{1}{3})$	$(+\frac{1}{3}, +\frac{1}{3})$
u_R	0	$+\frac{2}{3}$	$+\frac{4}{3}$
d_R	0	$-\frac{1}{3}$	$-\frac{2}{3}$

Table 1.4: Electroweak quantum numbers for the SM fermionic fields. The table lists the weak isospin I_3, the electric charge Q, and the weak hypercharge Y (evaluated from the formula $Q = I_3 + Y/2$); for doublets the upper/lower components are given.

The not-interacting EW Lagrangian can be written as:

$$\mathscr{L}_{EW} = \bar{\mathbf{L}} i\gamma^\mu \partial_\mu \mathbf{L} + \bar{\psi}_R(x) i\gamma^\mu \partial_\mu \psi_R(x) + \bar{\psi}'_R(x) i\gamma^\mu \partial_\mu \psi'_R(x) \tag{1.14}$$

As in the previous case $\mathscr{L}_{EW}$ must be invariant under a local gauge transformation. Local transformations can be written as follows:

$$\begin{aligned} \begin{pmatrix} \psi_L(x) \\ \psi'_L(x) \end{pmatrix} &\to \mathbf{U}(x) e^{i\chi(x)y_L} \begin{pmatrix} \psi_L(x) \\ \psi'_L(x) \end{pmatrix} \\ \psi_R(x) &\to e^{i\chi(x)y_R} \psi_R(x) \end{aligned} \tag{1.15}$$

$$\text{with } \mathbf{U}(x) = exp\left[i \sum_{k=1}^{3} \alpha^k(x) \frac{\sigma_k}{2} \right] \tag{1.16}$$

where σ_k are the three Pauli matrices, $\alpha^k(x)$ are the three local parameters of the transformation and y_L (y_R) is the left (right) weak hypercharge. Also in this case (1.14) is not invariant under the transformations (1.15). In order to make $\mathscr{L}_{EW}$ invariant under a local transformation, ∂_μ has to be replaced differently for fermions of opposite chirality:

$$\begin{aligned} L: \quad & \mathbf{D}_\mu = \partial_\mu + ig\mathbf{W}_\mu(x) + ig'\mathbf{B}_\mu(x) y_L \\ R: \quad & \mathbf{D}_\mu = \partial_\mu + ig'\mathbf{B}_\mu(x) y_R \end{aligned} \tag{1.17}$$

$$\text{with } \mathbf{W}_\mu(x) = \mathbf{W}^a_\mu(x) \frac{\sigma_a}{2} \tag{1.18}$$

where g and g' are the coupling constants respectively of $SU(2)_L$ and $U(1)_Y$, $\mathbf{W}^a_\mu$ and $\mathbf{B}_\mu$ are four gauge fields. The gauge fields are transformed as

$$\begin{aligned} \mathbf{W}_\mu(x) &\to \mathbf{U}(x) \mathbf{W}_\mu(x) \mathbf{U}^\dagger(x) - \frac{i}{g} \mathbf{U}(x) \partial_\mu \mathbf{U}^\dagger(x) \\ \mathbf{B}_\mu(x) &\to \mathbf{B}_\mu(x) - \frac{1}{g'} \partial_\mu \chi(x) \end{aligned} \tag{1.19}$$

Finally, the invariant EW Lagrangian can be written as:

$$\begin{aligned} \mathscr{L}_{EW} = -\frac{1}{2} Tr\left[\mathbf{W}_{\mu\rho}(x) \mathbf{W}^{\mu\rho}(x)\right] \quad &- \quad \frac{1}{4} \mathbf{B}_{\mu\rho}(x) \mathbf{B}^{\mu\rho}(x) \quad + \\ \bar{L} i\gamma^\mu [\partial_\mu + ig\mathbf{W}_\mu(x) \quad &+ \quad ig'\mathbf{B}_\mu(x) y_L] L \quad + \\ \bar{\psi}_R(x) i\gamma^\mu \left[\partial_\mu + ig'\mathbf{B}_\mu(x) y_R\right] \psi_R(x) \quad &+ \quad \bar{\psi}'_R(x) i\gamma^\mu \left[\partial_\mu + ig'\mathbf{B}_\mu(x) y_R\right] \psi'_R(x) \end{aligned} \tag{1.20}$$

$$\text{with } \mathbf{W}_{\mu\rho}(x) = \left[\partial_\mu \mathbf{W}^a_\rho(x) - \partial_\rho \mathbf{W}^a_\mu(x) - g\epsilon_{abc}\mathbf{W}^b_\mu(x)\mathbf{W}^c_\rho(x)\right]\frac{\sigma_a}{2} \tag{1.21}$$

$$\text{and } \mathbf{B}_{\mu\rho}(x) = \partial_\mu \mathbf{B}_\rho(x) - \partial_\rho \mathbf{B}_\mu(x) \tag{1.22}$$

where ϵ_{abc} are the structure constants of the $SU(2)_L$ group[3]. The first two terms in (1.20) describe trilinear and quadrilinear EW interactions among gauge bosons while the third and fourth term are the free Lagrangian from (1.14) with the interactions between fermions and EW bosons.

The connections between $\mathbf{W}^a_\mu$ and $\mathbf{B}_\mu$ and physical weak boson fields ($\mathbf{W}^\pm_\mu$ and $\mathbf{Z}_\mu$) and the photon field $\mathbf{A}_\mu$ are given by the following linear combinations:

$$\begin{aligned} \mathbf{W}^\pm_\mu &= \frac{\mathbf{W}^1_\mu \mp i\mathbf{W}^2_\mu}{\sqrt{2}} \\ \mathbf{Z}_\mu &= \cos\theta_W \mathbf{W}^3_\mu - \sin\theta_W \mathbf{B}_\mu \\ \mathbf{A}_\mu &= \sin\theta_W \mathbf{W}^3_\mu + \cos\theta_W \mathbf{B}_\mu \end{aligned} \tag{1.23}$$

$$\text{with } \cos\theta_W = \frac{g}{\sqrt{g^2 + g'^2}} \tag{1.24}$$

$$\text{and } e = \frac{gg'}{\sqrt{g^2 + g'^2}} \tag{1.25}$$

where θ_W, called Weinberg angle, is the weak mixing angle and e is the electric charge.

1.1.5 Spontaneous symmetry breaking and the Higgs sector

Although the presented $\mathscr{L}$s are able to predict interactions between elementary particles to an unprecedented scale [13], it predicts massless EW bosons, which is in contrast with the experimental results [14, 15]. A way out of this problem, ensuring the gauge simmetry, is the spontaneous symmetry breaking with the Brout-Englert-Higgs (BEH) mechanism which was first proposed in 1964 in two independent papers from Englert and Brout [16] and Higgs [17]. The mechanism requires an additional scalar (Higgs) field:

$$\phi(x) = \begin{pmatrix} \phi_1(x) \\ \phi_2(x) \end{pmatrix}, \tag{1.26}$$

a $SU(2)_L$ doublet of complex scalar fields introduced in the Lagrangian of the SM with the term:

$$\mathscr{L}_{Higgs} = \left[\partial_\mu \phi^\dagger(x)\right]\left[\partial^\mu \phi(x)\right] - V\left(\phi(x)\right) \tag{1.27}$$

where

$$V\left(\phi(x)\right) = -\mu^2\left[\phi^\dagger(x)\phi(x)\right] + \lambda\left[\phi^\dagger(x)\phi(x)\right]^2 \tag{1.28}$$

with μ^2 and λ positive constant terms. The minimum of $V(\phi(x))$ is reached when

$$\sqrt{\phi^\dagger\phi} = \frac{\rho_0}{\sqrt{2}} = \frac{1}{\sqrt{2}}\sqrt{\frac{\mu^2}{\lambda}} \tag{1.29}$$

The interaction terms between the Higgs field and the fermions are scalar under Lorentz transformations, invariant under $SU(2)_L \bigotimes U(1)_Y$ transformations and contain minimal, non-derivative

[3] $\epsilon_{123} = \epsilon_{231} = \epsilon_{312} = 1$, $\epsilon_{132} = \epsilon_{213} = \epsilon_{321} = -1$, otherwise 0

interaction terms (no derivative couplings), in order to keep the theory renormalizable. They can be therefore written as:

$$\begin{aligned}\mathscr{L}_{Yuk} = &- c\left[\phi_1^\dagger(x)\bar{\psi}'_R(x)\psi_L(x) + \phi_2^\dagger(x)\bar{\psi}'_R(x)\psi'_L(x)\right]\\ &- c'\left[\phi_1^\dagger(x)\bar{\psi}_R(x)\psi_L(x) + \phi_2^\dagger(x)\bar{\psi}_R(x)\psi'_L(x)\right] + h.c.\end{aligned} \tag{1.30}$$

where c and c' are coupling constants and $h.c.$ stands for "hermitian conjugate". The invariance under hypercharge transformations is guaranteed assigning a hypercharge y_H to the Higgs field satisfying the relation

$$y_H = y_L - y_R \tag{1.31}$$

where y_L and y_R have been introduced in (1.15).
If we add to (1.20) the Yukawa term (1.30) and the Higgs Lagrangian (1.27), the full Lagrangian remains gauge invariant provided that the ordinary derivatives acting on the Higgs field are replaced by the appropriate covariant derivative as follows:

$$\partial_\mu\phi(x) \to \mathbf{D}_\mu\phi(x) = \left[\partial_\mu + ig\mathbf{W}_\mu(x) + ig'y_H\mathbf{B}_\mu(x)\right]\phi(x) \tag{1.32}$$

The invariance under $SU(2)_L \times U(1)_Y$ transformation allows to write the Higgs field as

$$\phi(x) = \begin{pmatrix} 0 \\ \frac{1}{\sqrt{2}}\,\rho(x) \end{pmatrix}, \tag{1.33}$$

with $\rho(x) > 0$ and vacuum expectation value $\langle 0|\rho(x)|0\rangle = \rho_0$. Even if not invariant under $SU(2)_L \bigotimes U(1)_Y$, the vacuum expectation value is broken down to a residual $U(1)_{\text{EM}}$ subgroup invariance, whose generator is the electric charge operator Q. This means that only those gauge transformations corresponding to Q leave the ground state invariant, while the other combinations of $SU(2)_L$ and $U(1)_Y$ do not.
If the fluctuation ρ' around the vacuum expectation value of the Higgs fields is considered

$$\rho'(x) = \rho(x) - \rho_0 \;\text{ with }\; \langle 0|\rho'(x)|0\rangle = 0, \tag{1.34}$$

it is possible to write the Higgs field as

$$\phi(x) = \begin{pmatrix} 0 \\ \frac{\rho_0}{\sqrt{2}}\left[1 + \frac{\rho'(x)}{\rho_0}\right] \end{pmatrix} \tag{1.35}$$

and the complete SM Lagrangian can be expressed in the following form

$$\begin{aligned}\mathscr{L}_{SM} = &-\frac{1}{4}\mathbf{G}^a_{\mu\rho}(x)\mathbf{G}^{a\mu\rho}(x) - \frac{1}{2}Tr\left[\mathbf{W}_{\mu\rho}(x)\mathbf{W}^{\mu\rho}(x)\right] - \frac{1}{4}\mathbf{B}_{\mu\rho}(x)\mathbf{B}^{\mu\rho}(x)\\ &+ \mathbf{W}^+_\mu\mathbf{W}^{-\mu}m_W^2\left[1 + \frac{\rho'(x)}{\rho_0}\right]^2 + \frac{1}{2}\mathbf{Z}_\mu\mathbf{Z}^\mu m_Z^2\left[1 + \frac{\rho'(x)}{\rho_0}\right]^2\\ &+ \sum_\ell\left\{\bar{\nu}_{\ell L}(x)i\gamma^\mu\partial_\mu\nu_{\ell L}(x) + \bar{\ell}\left[i\gamma^\mu\partial_\mu - m_\ell\left[1 + \frac{\rho'(x)}{\rho_0}\right]\right]\ell\right\}\\ &+ \sum_q\bar{\mathbf{q}}(x)\left\{i\gamma^\mu\left(\partial_\mu + ig_s\mathbf{G}^a_\mu(x)\frac{\lambda_a}{2}\right) - m_q\left[1 + \frac{\rho'(x)}{\rho_0}\right]\right\}\mathbf{q}(x)\\ &+ \mathscr{L}_{int} + \frac{1}{2}\partial_\mu\rho'(x)\partial^\mu\rho'(x) - \frac{1}{2}m_H^2\rho'(x)^2\left\{1 + \frac{\rho'(x)}{\rho_0} + \frac{1}{4}\left[\frac{\rho'(x)}{\rho_0}\right]^2\right\}\end{aligned} \tag{1.36}$$

The indices $\ell = e, \mu, \tau$ and $q = u, c, t, d, s, b$ are to be summed over. The first three terms, the same in (1.7) and (1.20) are the self-interaction terms between gauge bosons. The fourth and the fifth terms are the interaction terms between the Higgs boson and the weak bosons, written in order to explicate the terms of boson masses evaluated with the (1.37). The sixth and the seventh terms are the fermionic free Lagrangians from (1.7) and (1.20) with the interactions between fermions and gluons and between fermions and Higgs boson, the masses are evaluated with (1.37). The last two terms are the free Lagrangian and the self-interactions of the Higgs boson, where the mass of the Higgs boson is evaluated with the (1.37).

$$
\begin{aligned}
m_W^2 &= \frac{e^2\rho_0^2}{4\sin^2\theta_W} \\
m_Z^2 &= \frac{e^2\rho_0^2}{4\sin^2\theta_W\cos^2\theta_W} \\
m_f &= c_f\frac{\rho_0}{\sqrt{2}} \\
m_H &= 2\lambda\rho_0^2
\end{aligned}
\tag{1.37}
$$

where $f = e, ..., t$ and c_f are coupling terms that must be determined experimentally. $\mathscr{L}_{int}$ is the interaction Lagrangian between fermions and EW bosons and is given by:

$$
\mathscr{L}_{int} = -e\left[\mathbf{A}_\mu\mathscr{J}^\mu_{em} + \frac{1}{\sin\theta_W\cos\theta_W}\mathbf{Z}_\mu\mathscr{J}^\mu_{NC} + \frac{1}{\sqrt{2}sin\theta_W}\left(\mathbf{W}^+_\mu\mathscr{J}^\mu_{CC} + \mathbf{W}^-_\mu\mathscr{J}^{\mu\dagger}_{CC}\right)\right]
\tag{1.38}
$$

The currents which appears here are the electromagnetic current $\mathscr{J}^\mu_{em}$, the neutral current $\mathscr{J}^\mu_{NC}$ and the charged current $\mathscr{J}^\mu_{CC}$. $\mathscr{J}^\mu_{em}$ is

$$
\mathscr{J}^\mu_{em} = \sum_\ell -\ell\gamma^\mu\ell \quad + \quad \sum_q Q_q\bar{\mathbf{q}}\gamma^\mu\mathbf{q}
\tag{1.39}
$$

where Q_q are the quark charges. $\mathscr{J}^\mu_{NC}$ is

$$
\begin{aligned}
\mathscr{J}^\mu_{NC} =& (\overline{\nu}_e, \overline{\nu}_\mu, \overline{\nu}_\tau)\gamma^\mu\frac{1}{2}\frac{1-\gamma_5}{2}\begin{pmatrix}\nu_e\\ \nu_\mu\\ \nu_\tau\end{pmatrix} \\
&+ (\bar{e}, \bar{\mu}, \bar{\tau})\gamma^\mu\left(-\frac{1}{2}\frac{1-\gamma_5}{2} + \sin^2\theta_W\right)\begin{pmatrix}e\\ \mu\\ \tau\end{pmatrix} \\
&+ (\bar{\mathbf{u}}, \bar{\mathbf{c}}, \bar{\mathbf{t}})\gamma^\mu\left(\frac{1}{2}\frac{1-\gamma_5}{2} - \frac{2}{3}\sin^2\theta_W\right)\begin{pmatrix}\mathbf{u}\\ \mathbf{c}\\ \mathbf{t}\end{pmatrix} \\
&+ (\bar{\mathbf{d}}, \bar{\mathbf{s}}, \bar{\mathbf{b}})\gamma^\mu\left(-\frac{1}{2}\frac{1-\gamma_5}{2} + \frac{1}{3}\sin^2\theta_W\right)\begin{pmatrix}\mathbf{d}\\ \mathbf{s}\\ \mathbf{b}\end{pmatrix}
\end{aligned}
\tag{1.40}
$$

$\mathscr{J}^\mu_{CC}$ is

$$
\mathscr{J}^\mu_{CC} = (\overline{\nu}_{eL}, \overline{\nu}_{\mu L}, \overline{\nu}_{\tau L})\gamma^\mu\begin{pmatrix}e_L\\ \mu_L\\ \tau_L\end{pmatrix} + (\bar{\mathbf{u}}_L, \bar{\mathbf{c}}_L, \bar{\mathbf{t}}_L)\gamma^\mu\mathbf{V}_{CKM}\begin{pmatrix}\mathbf{d}_L\\ \mathbf{s}_L\\ \mathbf{b}_L\end{pmatrix}
\tag{1.41}
$$

$\mathbf{V}_{CKM}$, called Cabibbo-Kobayashi-Maskawa matrix, is introduced in the theory because the quark mass eigenstates $\mathbf{d}_L, \mathbf{s}_L, \mathbf{b}_L$ differ from the quark flavour eigenstates $\mathbf{d}'_L, \mathbf{s}'_L, \mathbf{b}'_L$. It is defined by three angles $\theta_1, \theta_2, \theta_3$ and a complex phase $e^{i\delta}$, all determined experimentally.
Because of the absence of an interaction term between leptons and gluons in (1.36) it is assumed that leptons do not have strong interactions. Because of the absence of an interaction term in (1.36) between Higgs boson and γ/g it is assumed that the photon and the gluons do not have mass, the same thing is also assumed for neutrinos although there are experimental results of neutrinos oscillations which implies that neutrinos are massive [18]. Including neutrinos masses in (1.36) can be done assuming that left-handed neutrinos have mass eigenvalues different from flavour eigenvalues and therefore introducing the Pontecorvo–Maki–Nakagawa–Sakata matrix [19].

1.2 Higgs and Z boson at CMS

Within the physics program of the CMS experiment, both the Higgs boson and the Z boson play a central role, as they are key elements of the SM of particle physics and provide opportunities for both precision measurements and searches for new phenomena.
The Higgs boson is associated with the mechanism responsible for generating the masses of fundamental particles through electroweak symmetry breaking. Consequently, precise measurements of its production cross sections and decay branching fractions allow stringent tests of the SM predictions and offer sensitivity to possible contributions from physics beyond the Standard Model (BSM).
The Z boson, on the other hand, is produced with large cross sections and features clean and well-understood decay signatures. For this reason, it is widely used as a reference process for detector calibration, alignment, and validation of reconstruction and analysis techniques. Measurements involving Z bosons also enable precision studies of electroweak parameters and provide important constraints on PDFs.
Owing to these properties, both the Higgs and Z bosons are extensively exploited across CMS analyses, ranging from detector performance studies and calibration procedures to precision measurements and searches for new physics. Currently the most precise measurement of the mass of the Higgs boson was obtained by ATLAS combining the results of $H \to ZZ^* \to 4\ell$ and $H \to \gamma\gamma$ with the Run 2 and the Run 1 data [20]:

$$m_H = 125.11 \pm 0.09\,(stat) \pm 0.06\,(syst)\ \text{GeV}.$$

CMS does not provide an independent determination of the Z-boson mass that supersedes the measurement obtained at the Large Electron-Positron Collider (LEP) [21, 22]; instead, CMS adopts the PDG value [6]

$$m_Z = 91.1880 \pm 0.0020\ \text{GeV}$$

as the reference for calibrations.

1.2.1 Production mechanisms

For center-of-mass energy $\sqrt{s}$=13.6 TeV (the dataset configuration used in this analysis) Higgs production is dominated by gluon fusion, which proceeds via a virtual top-quark loop (Figure 1.2(a)). The next-largest contribution comes from vector-boson fusion, mediated by W or Z exchange with the Higgs radiated from the weak-boson propagator (Figure 1.2(b)). In decreasing order of cross section there are the associated production with W and Z bosons (Figure 1.2(c)), the production with $\mathrm{t\bar{t}}$ or with $\mathrm{b\bar{b}}$ (Figure 1.2(d)). The smallest cross section among these is the single-top-associated production (Figure 1.2(e)), which is nonetheless valuable for probing quantities such as the sign of the top Yukawa coupling.

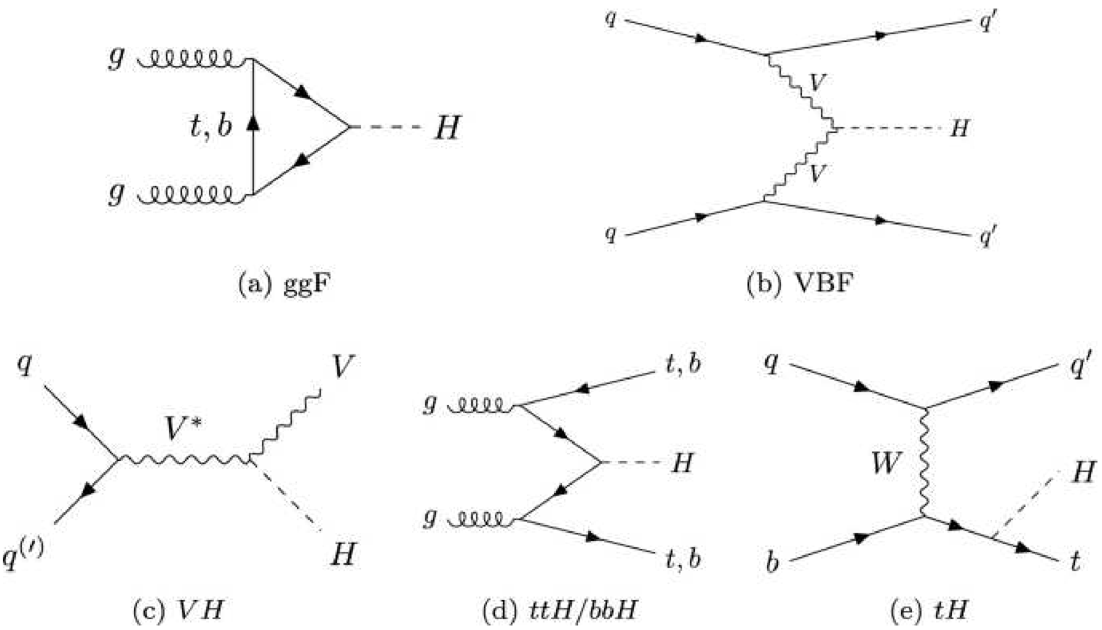

Figure 1.2: Feynman diagrams for the Higgs boson production in (a) gluon-gluon fusion, (b) vector boson fusion, (c) associated production with a W or Z (V) boson, (d) associated production with a top or bottom quark pair, (e) associated production with a single top quark [23].

Table 1.5 summarizes the theoretical cross sections for the dominant Higgs production processes in pp collisions for $m_H = 125$ GeV, together with the relative uncertainties arising from sources such as higher-order perturbative QCD corrections, uncertainties in the PDFs, and EW corrections.
At CMS, the Z boson is mainly produced via the Drell–Yan process (quark–antiquark annihilation). When a Z boson is highly boosted it very frequently appears recoiling against a hard parton (quark or gluon) that hadronizes into a jet. Conservation of transverse momentum demands a balancing object: producing a Z with large momentum in a pp collision almost always involves a hard recoil from the hard-scattering subprocess ($qg \to Zq$ or $q\bar{q} \to Zg$) or from high-momentum initial (final) state radiation. If the Z decays hadronically, its two decay quarks become collimated at high boost and often appear as a single "fat jet" with a two-prong substructure, which is morphologically different from a single-parton QCD jet that shows wider-angle radiation. CMS measured at $\sqrt{s}$=13.6 TeV the inclusive production cross section with the Z boson decaying into two μ, resulting into:

$$\sigma(pp \to Z + X) \times \mathcal{B}(Z \to \mu^+\mu^-) = 2.021 \pm 0.009\,(\text{syst}) \pm 0.028\,(\text{lumi})^{+0.011}_{-0.013}\,(\text{acceptance})\text{ nb}$$

including contributions from the uncertainty in the integrated luminosity (lumi), the uncertainty associated with the fiducial acceptance and the residual systematic effects (syst), while the statistical uncertainty is negligible [24].

1.2.2 Decay channels

Since the SM predicts a mean lifetime of about $1.6 \cdot 10^{-22}$ s and $2.6 \cdot 10^{-25}$ s respectively for a Higgs boson with a mass of 125 GeV and the Z boson [6], both bosons fail to reach the detector before

Production mode	Cross section (pb)
ggF	$52.2^{+5.6\%}_{-7.4\%}$
VBF	$4.1^{+2.1\%}_{-1.5\%}$
WH	$1.46^{+1.8\%}_{-1.9\%}$
ZH	$0.95^{+4.0\%}_{-3.6\%}$
ttH	$0.57^{+6.9\%}_{-9.9\%}$
total	$59.2^{+5\%}_{-7\%}$

Table 1.5: Theoretical Higgs boson production cross sections for each production mode and the total theoretical production cross section at $\sqrt{s} = 13.6$ TeV for $m_H = 125$ GeV [6].

Decay channel	Predicted branching fraction (%)	Measured branching fraction (%)
$b\bar{b}$	58.24 ± 0.70	53 ± 8
WW^*	21.37 ± 0.33	25.7 ± 2.5
gg	8.19 ± 0.42	-
$\tau^+\tau^-$	6.27 ± 0.09	$6.0^{+0.8}_{-0.7}$
$c\bar{c}$	2.89 ± 0.09	-
ZZ^*	2.71 ± 0.04	2.80 ± 0.30
$\gamma\gamma$	0.227 ± 0.005	$(2.50 \pm 0.20)\times10^{-3}$
$Z\gamma$	0.153 ± 0.009	$(3.4 \pm 1.1)\times10^{-3}$

Table 1.6: Predicted and measured branching fractions for the decay channels of an Higgs boson with $m_H = 125$ GeV [25, 6].

decaying so they must be searched for through their decay products which are detectable in the experiment. Table 1.6 and 1.7 summaries respectively the Higgs boson and the Z boson branching ratios and the relative uncertainty for the decay channels. $b\bar{b}$ is the most probable Higgs boson decay with a branching ratio of $\sim 58\%$ and one of the most probable Z boson decay with a branching ratio of $\sim 15\%$. Despite the presence of many bosons that decay in this way, search in this channel is very challenging due to the overwhelming QCD background. This channel is mainly exploited in boosted regimes, mainly produced in association with a recoil jet, leaving a signal that is somewhat easier to distinguish from the QCD background. Over the years, the CMS experiment has developed several algorithms to distinguish jets originating from the hadronization of a resonance decaying into $b\bar{b}$ from those produced inclusively by QCD processes (see Section 3.10). During my PhD, I worked with algorithms that distinguish boosted jets, by validating the performance in the data (see Chapter 4) and training one of them for the trigger level (see Chapter 5).

Decay channel	Z boson branching fraction (%)
hadrons	69.911 ± 0.056
$\nu\bar{\nu}$	20.000 ± 0.055
$\tau^+\tau^-$	3.3696 ± 0.0066
$\mu^+\mu^-$	3.3662 ± 0.0066
e^+e^-	3.3632 ± 0.0042
$b\bar{b}$	15.12 ± 0.05
$c\bar{c}$	12.03 ± 0.21

Table 1.7: Observed branching fractions for the Z boson main decay channels. The hadronic fraction (hadrons) includes the specific contributions from $b\bar{b}$ and $c\bar{c}$ [6].

Chapter 2

The CMS detector at LHC

The conseil européen pour la recherche nucléaire (CERN) [26], officially founded on 29 September 1954, stands as one of the world's foremost centers for particle physics research, driving progress at the frontiers of science and technology. Among its facilities there is the Large Hadron Collider (LHC) [27], the largest and most powerful particle accelerator ever built. The LHC is a 27-kilometre circular tunnel equipped with superconducting magnets and a complex system of accelerating structures that increase the energy of two circulating particle beams in opposite directions, typically protons or heavy ions up to a nominal beam energy of 7 TeV. These high-energy beams are brought to collide at four dedicated interaction points, each instrumented with a large experimental apparatus.

The CMS experiment is placed in one of these interaction points and plays a crucial role in testing the predictions of the SM and searching for signatures of new physics. Its design enables the precise reconstruction of a wide variety of final states, making it a central tool to understand the building blocks of matter and the forces that govern the universe.

2.1 The Large Hadron Collider

The LHC is located in the same 27-km circumference underground tunnel that previously hosted the Large Electron Positron Collider until 2000. It is designed as both a proton–proton (pp) and a heavy-ion collider, with a nominal center-of-mass energy of $\sqrt{s} = 14\,\text{TeV}$. Under its design conditions, particle bunches cross every 25 ns, reaching a luminosity of $O(10^{34})\,\text{cm}^{-2}\text{s}^{-1}$. These extreme conditions allow the LHC to explore SM processes and also probing scenarios of new physics.

Before injection into the LHC, proton beams undergo a carefully staged pre-acceleration process through a chain of smaller accelerators, illustrated in Figure 2.1.

The process begins with hydrogen gas, whose electrons are stripped by an electric field, resulting in free protons. These protons are first accelerated by Linac4 (Linear Accelerator 4) to an energy of 50 MeV, and then injected into the Proton Synchrotron Booster (PSB). The PSB increases their energy up to 2 GeV, after which the beam is transferred to the Proton Synchrotron (PS). With a circumference of 628 m, the PS boosts the beam energy further to 26 GeV, preparing it for injection into the Super Proton Synchrotron (SPS). The SPS, a circular accelerator with circumference of 7 km, raises the beam energy to 450 GeV, which is the injection energy for the LHC. Once inside the LHC, the beams are accelerated to their target energies and then brought to collide at four dedicated interaction points. These collision points host four LHC experiments: ALICE (A Large Ion Collider Experiment) [29], ATLAS, CMS and LHCb (Large Hadron Collider beauty) [30]. Alongside them,

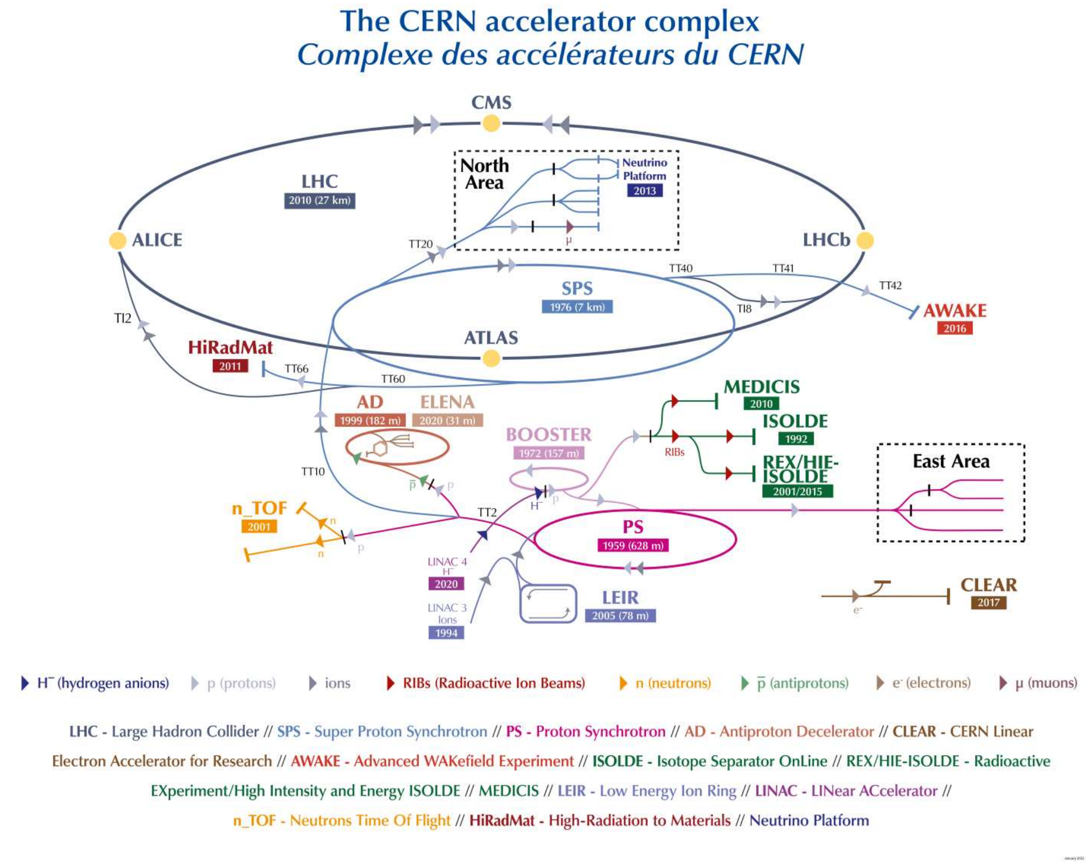

Figure 2.1: CERN accelerator complex: a sequence of machines (Linac4 → PSB → PS *to* SPS → LHC) that progressively boost particle beams before final injection into the LHC for high-energy collisions [28].

three additional specialized detectors extend the physics reach of the collider: TOTEM (TOTal Elastic and diffractive cross-section Measurement) [31] located near CMS; MoEDAL (Monopole and Exotics Detector at the LHC) [32] close to LHCb; LHCf (Large Hadron Collider forward) [33] positioned near ATLAS; FASER (ForwArd Search ExpeRiment) [34] installed in the far-forward region downstream of ATLAS.
Unlike a continuous beam, the LHC uses bunched beams, with each bunch containing around 10^{11} protons. Under nominal operating conditions, a beam may contain up to 2808 bunches. The shaping of these proton bunches and their stable acceleration are ensured by a system of eight radio-frequency cavities per beam, where oscillating electromagnetic fields at 400 MHz compress and accelerate the particles. Importantly, the bunch size is not uniform around the ring: it is deliberately reduced, or "squeezed", at the interaction points to maximize the probability of collisions between protons. A comparison between the original design parameters of the LHC and the achieved performance in 2025 is reported in Table 2.1.

	Designed	2025
Centre of mass energy	14 TeV	13.6 TeV
Time spacing	25 ns	25 ns
Number of bunches	2808	2460
Number of protons per bunch	$1.15 \cdot 10^{11}$	$1.15 \cdot 10^{11}$

Table 2.1: Comparison between the LHC design parameters and the operational conditions for *pp* collisions in 2025, including center-of-mass energy, bunch spacing, number of bunches and protons per bunch.

The parameter which defines the performances of an accelerator is the instantaneous luminosity (L) defined as:

$$L = \frac{R_i}{\sigma_i} = \frac{1}{\sigma_i}\frac{dN_i}{dt} \tag{2.1}$$

where σ_i and R_i are respectively the cross section and the rate, defined as the number of events produced (N_i) per unit of time, of a given process i. L can be expressed in terms of the beam parameters as:

$$L = \gamma \frac{f n_b N^2}{4\pi\epsilon_n\beta^*}R;\ R = 1/\sqrt{1+\left(\frac{\theta_c\sigma_Z}{2\sigma}\right)^2} \tag{2.2}$$

where $\gamma = E/m$ is the relativistic Lorentz factor of the colliding particles, n_b is the number of bunches per beam, N is the number of particles per bunch, f is the revolution frequency, β^* is the beam beta function[1] at the collision point, ϵ_n is the normalized transverse beam emittance[2], R is a reduction factor due to the beam crossing angle at the interaction point, θ_c is the full crossing angle between colliding beams, σ and σ_z are the transverse and longitudinal RMS beam sizes, respectively [27]. That expression is valid under the following idealized assumptions: gaussian transverse and longitudinal bunch distributions, identical beams ($N_1 = N_2 = N$) with equal emittances and transverse sizes at the interaction point and round beams ($\sigma_x = \sigma_y = \sigma$). Figure 2.2 shows the integrated luminosity as a function of time, delivered by the LHC and recorded by the CMS experiment during stable beams for *pp* collisions at 13.6 TeV for Run 3 (2022-2026) [35].

[1] The beta function describes how the beams are focused along the accelerator. In particular, it determines the transverse size of the beams at the collision point.

[2] The normalized transverse beam emittance describes how much the particles of the beams are spread out in the transverse phase space (position and angle with respect to the beam direction).

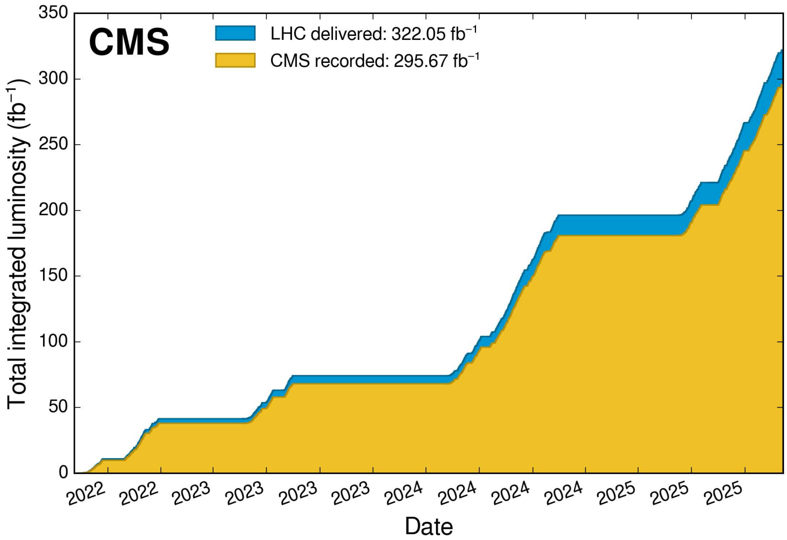

Figure 2.2: Cumulative day-by-day integrated luminosity from 2022 to 2025. The plot shows the measured luminosity delivered by the LHC to CMS (blue) and recorded by the CMS (orange) during stable beams [35].

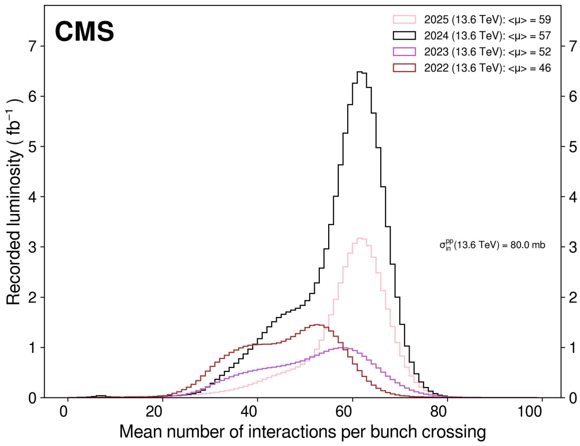

Figure 2.3: Distribution of the average pileup with the corresponding mean values for each year of Run 3 *pp* collisions: 2022 (brown), 2023 (fuchsia), 2024 (black), 2025 (pink) [35].

During LHC operation, along with the luminosity, the mean number of proton-proton collisions per bunch crossing ($< \mu >$) increased from ≈ 46 in 2022 up to ≈ 59 in 2025. Figure 2.3 reports interactions per crossing, denoted as pileup (PU), for different data taking periods [35].

2.2 The Compact Muon Solenoid

CMS is one of the general purpose experiments which takes data at the LHC with a cylindrical structure with a diameter of 15 m and a length of 21.6 m (28.7 m considering the external structural parts). Its physics goals range from the search for the Higgs boson to the searches for physics beyond the SM, to the precision measurements of already known particles and phenomena. CMS has a so called "onion structure", as illustrated in Figure 2.4, where each layer is designed to stop or track a different type of particle emerging from *pp* and heavy ion collisions.
The detector is built around a huge solenoidal magnet which has the form of a cylindrical coil of superconducting cable of 5.9 m inner diameter. Particles emerging from collisions first pass through the tracker system which detects charged particles tracks and allows to measure their momentum and determine their charge. Outside the tracker, the calorimeters stop the particles and measure their energy: the electromagnetic calorimeter (ECAL) is a sub-detector designed to measure the energy of γs and electrons, the hadronic calorimeter (HCAL) is a sub-detector designed to detect hadrons. Both tracker and calorimeters are placed inside the solenoid. Muons are detected in gas-ionization chambers embedded in the steel flux-return yoke outside the magnet. The only particles from the SM that are not detectable by the CMS detector are neutrinos. In the following, different CMS subdetectors have been described starting from the tracker, that is the closest to the interaction point, and moving outwards.

2.2.1 Coordinate system

The coordinate system adopted by CMS has the origin centered at the nominal collision point inside the experiment, the x axis points to the center of the LHC ring, the y axis points upwards and the

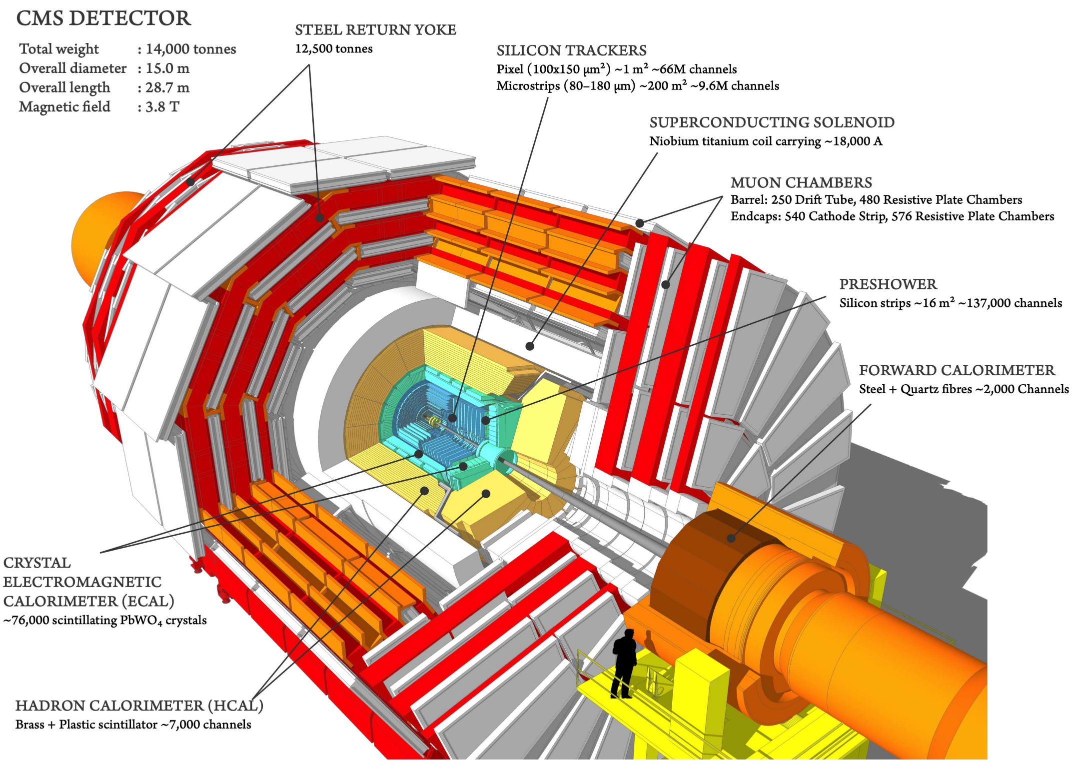

Figure 2.4: Perspective view of the CMS detector with major sub-detectors indicated [36].

z axis coincides with the proton beam direction. Cylindrical coordinates are used. The transverse plane is given by the x-y plane. The azimuthal angle ϕ is measured from the x axis in the x-y plane and takes values in $[-\pi, \pi]$. The polar angle θ is measured from the z axis and takes values in $[0;\pi]$. An other variable used to describe particles is the pseudorapidity (η) defined as:

$$\eta = -\ln \tan \frac{\theta}{2} \tag{2.3}$$

The value $\eta = 0$ corresponds to a direction perpendicular to the beamline, while the limit $\eta = \infty$ gives a direction parallel to the beamline. For very energetic particles ($|\mathbf{p}| \gg m$) the mass term can be neglected and η coincides with the rapidity (y) defined as:

$$y = \frac{1}{2} \ln \frac{E + p_z}{E - p_z} = \frac{1}{2} \ln \frac{\sqrt{m^2 + |\mathbf{p}|^2} + |\mathbf{p}| \cos\theta}{\sqrt{m^2 + |\mathbf{p}|^2} - |\mathbf{p}| \cos\theta} \tag{2.4}$$

The angular distance between two particles with azimuthal angles $\phi_{1(2)}$ and pseudorapidities $\eta_{1(2)}$ is commonly expressed as ΔR defined as:

$$\Delta R = \sqrt{(\phi_1 - \phi_2)^2 + (\eta_1 - \eta_2)^2} \tag{2.5}$$

2.2.2 The tracking system

The CMS tracking system, or tracker, is the innermost active sub-detector surrounding the beam pipe and is designed to reconstruct the trajectories of charged particles by measuring their curvature in the magnetic fields [37]. Due to the geometry of CMS, the particle trajectories are often described in the transverse plane. Indeed, thanks to the presence of the magnetic field, charged tracks curve in the transverse plane; the curvature directly gives the particle transverse momentum (p_T) by

$$p_T[GeV] \simeq 0.3\; q\; B[T]\; R[m] \tag{2.6}$$

where q is the electric charge, B is the magnetic field and R is the radius of curvature in the x-y plane. In addition, measuring curvature is simpler and more accurate than measuring the longitudinal component. The entire tracker is made of silicon sensors to ensure high spatial granularity and large hit redundancy, features that are essential for robust pattern recognition, precise momentum measurement and the reconstruction of primary and secondary vertices. When a charged particle crosses a silicon sensor cell, it produces electron–hole pairs that drift in the presence of an applied electric field, resulting in a detectable signal pulse. By having a precise knowledge of the spatial position of the sensors hit by a charged particle, the particle trajectory can be reconstructed. Two different technologies are used in CMS: silicon pixel and silicon strip sensors, respectively for the innermost and the outermost layers of the tracking system.

The pixel detector

The pixel detector [38], covers up to $|\eta| < 2.5$ with an active area of approximately 1.9 m^2 and about 124 million pixels of size $100 \times 150\ \mu\mathrm{m}^2$, providing typical spatial resolutions of $\approx 9.5\ \mu$m in the transverse plane and $\approx 22\ \mu$m along z. The high resolution is essential for the reconstruction of the primary vertex and the secondary vertices from b hadrons or τ lepton decays. The layout adopted for Run 3 comprises four concentric barrel layers (BPix) at radii of 29, 68, 109 and 160 mm and three endcap disks per side (FPix) located at $z \approx 291$, 396 and 516 mm, implements a carbon fiber mechanical support, CO_2 cooling and a digital readout to cope with higher rates. The layout of the current pixel detector is compared to the one of the original pixel detector in Figure 2.5.

During technical stop in June 2023, 27 modules in BPix layers 3 and 4 became non-operational due to a failure in the distribution of the LHC clock signals to those modules [39]. Since then, the affected modules have remained switched off. The affected area corresponds to a sector of about 0.4 radians (≈23°) in ϕ centered around $\phi \approx -1.0$ at negative η. Because the affected regions in η–ϕ space overlap across both layers, this results in a localized acceptance hole in the formation of track seeds from standard pixel-hit combinations. Consequently, the year 2023 is treated in two periods for tracking studies: pre-BPix, before the modules became non-operational, and post-BPix, after the affected modules were switched off.

The silicon strip detector

The silicon strip detector consists of a total of about 9.3 million strips, covering an area of about 198 m^2 up to $|\eta| < 2.5$. The layout of the the silicon strip detector can be seen in Figure 2.6. It consists of 10 cylinder layers in the barrel region around the pixel detector:

- four-layer tracker inner barrel (TIB) cover r < 55 cm and |z| < 118 cm and provide a single-point resolution of 13-38 μm in the transverse plane and 23 μm in the z-direction;

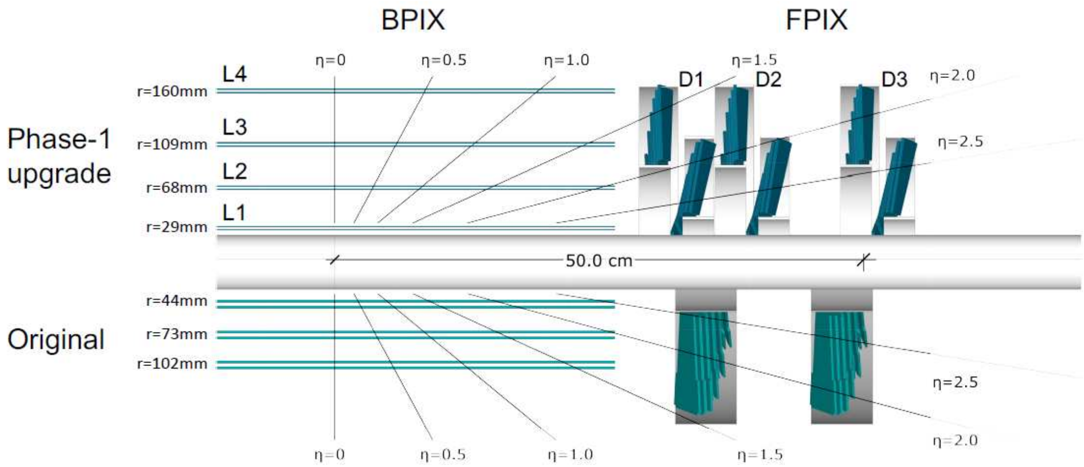

Figure 2.5: Layout of the current pixel detector (Phase-1) compared to the original detector layout, in longitudinal view. The original detector consisted of three barrel layers at radii of 44, 73, and 102 mm and two endcap disks per side located at 345 and 465 mm from the interaction point [38].

- six-layer tracker outer barrel (TOB) covers r $>$ 55 cm and $|$z$|$ $<$ 118 cm providing a resolution of 18-47 μm in the transverse plane and 47μm in z-direction.

The endcap disks are populated by concentric rings of wedge-shaped silicon strip modules, arranged radially around the beam axis in petal-like structures. 12 disks are installed at each end of the barrel cylinders:

- three tracker inner disks (TIDs) at approximately $|z| \approx 70$, 85, and 105 cm providing the same resolution and coverage as the TIB detectors;
- nine the tracker endcap (TEC) disks with the same resolution as TOB detectors. They cover the region 124 $<$ $|$z$|$ $<$ 282 cm.

The track p_T resolution varies between 0.7% to 1.5% for tracks with 1 GeV to 100 GeV respectively.

2.2.3 Electromagnetic calorimeter

ECAL is designed to measure energy of electrons and γ. A schematic view of the region covered by ECAL is shown in Figure 2.7. All ECAL is made of lead tungstate ($PbWO_4$) crystals.
The energy measurement is based on the conversion of the incident electron, positron or γ inside the crystals to an EM shower, *i.e.* a particle cascade made of electrons/positrons and γ where the γ produces an e^+e^- pair and the electron (positron) emits a γ via bremsstrahlung until the energy of the γ is below the threshold to produce a further e^+e^- pair. The scintillation light produced by the EM shower is detected in the barrel by photomultipliers (PMT) and in the endcap by a vacuum phototriode. The $PbWO_4$ crystals show a high density (8.28 g/cm^3), small radiation length (X_0 = 0.89 cm), short Molière radius (R_M = 2.2 cm)[3] and short scintillation decay time (25 ns). The

[3] R_M approximates the radius of a cylinder coaxial with the shower axis which contains 90% of the EM shower energy.

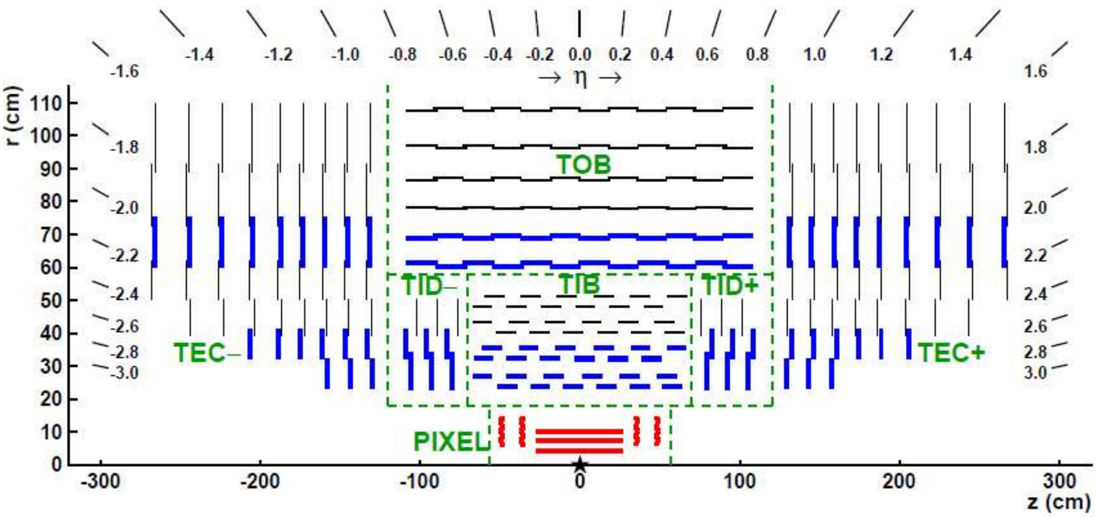

Figure 2.6: Schematic cross-section of the CMS tracker in the r–z plane. Green dashed lines indicate the boundaries of the different tracker subsystems. Strip tracker modules providing 2-D hit information are shown as thin black lines, while those enabling 3-D hit reconstruction are represented by thick blue lines. The latter consist of two back-to-back strip modules, with one rotated by a 100 mrad angle. Pixel modules, shown in red, also provide 3-D hit measurements. Modules are slightly staggered in r or z to ensure overlap and avoid acceptance gaps [40].

disadvantage of this material is the relatively low light yield, corresponding to about 30 γ per MeV of deposited energy, which calls for the usage of photo-detectors with internal amplification. The ECAL is constituted of a barrel and two endcap regions.

The Barrel Calorimeter

The ECAL barrel (EB) covers the region $|\eta| < 1.479$ and is instrumented with 61,200 crystals. The EB front faces lie at a radius of 1.29 m. Each crystal has a roughly square cross section, 22×22 mm^2 at the front and 26×26 mm^2 at the rear, and a length of 230 mm ($\approx 25.8X_0$). To minimize the small inter-module gaps (so-called cracks) aligned with particle trajectories, the crystal axes are tilted by about 3° with respect to the line from the nominal interaction point. The barrel is split into two half-cylinders, EB+ and EB-; each half contains 18 supermodules, each covering 20° in ϕ. A supermodule is formed from four modules (one with 500 crystals and three with 400 each). For mechanical simplicity, the crystals are assembled in 2×5 arrays housed in a thin (≈ 200 μm) alveolar structure that defines the submodule units.

The Endcap Calorimeter

The ECAL endcaps (EEs) extend the electromagnetic calorimeter coverage to the forward region, spanning in $|\eta|$ from 1.479 to 3. Each EE contains 7,324 $PbWO_4$ crystals (front face $28.62 \times 28.62\,\mathrm{mm}^2$, rear face $\approx 30 \times 30\,\mathrm{mm}^2$, length $220\,\mathrm{mm} \simeq 24.7\,X_0$). Each half-endcap (called Dee) comprises 3,662 crystals arranged as 138 full 5×5 supercrystals plus 18 partial supercrystals on the inner/outer perimeters. These partial supercrystals contain less than 25 crystals (on average ≈12 crystals each in the installed detector) and have irregular shapes to fill the available peripheral space

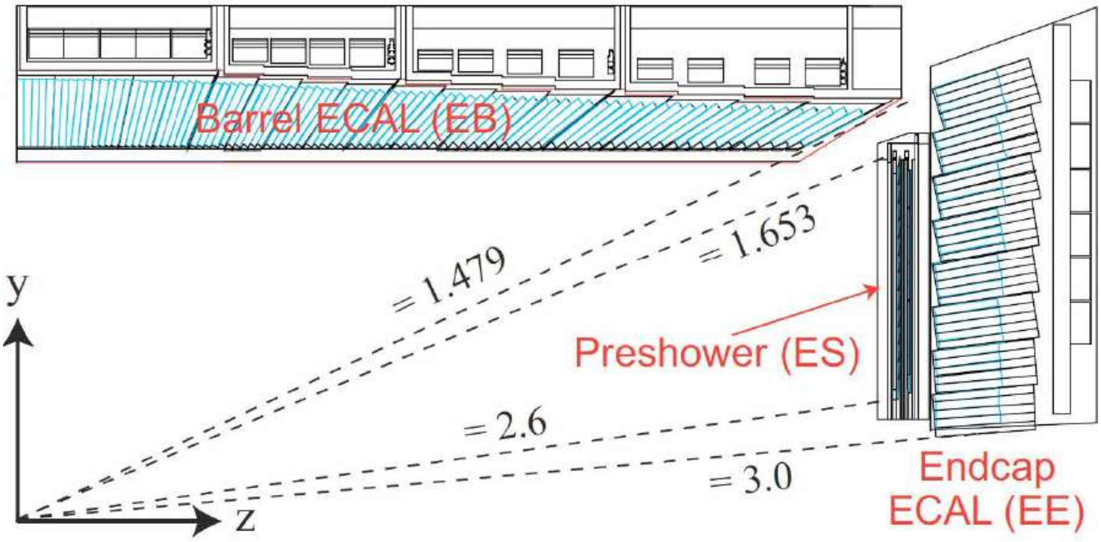

Figure 2.7: Longitudinal view of an ECAL quadrant. The pseudorapidity coverages of the barrel, endcap, and preshower systems are indicated; the preshower system enhances electron–photon separation in the endcap region [41].

while maintaining continuous coverage.

During Run 3 data taking in 2022 a water leak developed in the cooling circuit of the EE front-end electronics. The issue affected 491 crystals (centered at approximately $\eta \simeq 2$ and $\phi \simeq 2.2$) which were therefore masked from 17 September until the end of the 2022 run, leaving about 7% of the EE unavailable for data collection. Consequently, the year 2022 is treated in two periods for tracking studies: pre-EE and post-EE. The leak was repaired during the 2022 year-end technical stop and normal operations were resumed for the 2023 data taking [42].

The Preshower Detector

An EM PreShower detector is installed in front of the endcaps to assist in γ/π^0 separation and to improve the estimation of the γ direction, thanks to a higher spatial resolution. It covers a pseudorapidity range $1.65 \leq |\eta| \leq 2.61$. It contains two lead converters of thickness 2 X_0 and 1 X_0 respectively, followed by detector planes of silicon strips with a pitch less then 2 mm. The accuracy is typically 300 μm at 50 GeV.

ECAL calibration

The energy resolution of an electromagnetic calorimeter can be expressed as the sum in quadrature of three terms:

$$\frac{\sigma_E}{E} = \frac{A}{\sqrt{E(GeV)}} \oplus \frac{B}{E(GeV)} \oplus C \tag{2.7}$$

The first one (A) is a stochastic term that depends on intrinsic shower fluctuations, photoelectron statistics, dead material at the front of the calorimeter and sampling fluctuations for minimum ionizing particles. The second term (B) accounts for the PU and the noise due to the electronic of the readout chain which depends on the features of the readout circuit (detector capacitance, cables, etc.). The third term (C) is related to detector inhomogeneities (due to detector geometry, temperature gradients, radiation damage, etc.), instrumental effects and calibration uncertainties. The parameters that describe the CMS ECAL calibration are the following: A=2.8%, B=12% and C=0.3% [43].

2.2.4 Hadronic Calorimeter

HCAL provides the energy measurement of hadrons traversing the tracker and ECAL, providing the only direct energy measurement for neutral hadrons and complementing tracker momentum measurements for charged hadrons [4]. It is a compact, hermetic sampling calorimeter located between ECAL and the solenoid, constructed from alternating layers of dense absorber and plastic scintillator tiles. When a hadron interacts with the absorber material, both inelastic and elastic scattering processes occur between the incoming particle and the nucleons in the atomic nuclei of the material. The high-momentum transfer in these interactions leads to the production of a large number of secondary hadrons, which in turn continue to interact, creating a cascade of particles. The hadronic shower gradually ceases as the energy of the secondary hadrons decreases, until it stops completely due to energy loss through ionization or nuclear absorption. Incident hadrons interact in the brass absorber (70% Cu / 30% Zn, density ≈ 8.53 g $\times$ cm^{-3}) and the charged secondary particles, produced in the hadronic cascades, pass through the scintillator tiles and generate blue−violet scintillation light. Wavelength-shifting fibers embedded in each tile (diameter < 1 mm) absorb and re-emit this light in green, which is transported by clear optical fibers to the silicon photomultipliers (SiPMs), in turn interfaced with the front-end electronics. The HCAL system consists of a barrel (HB) and two endcap (HE) subsystems. The HB has readout cells of size $\Delta\eta \times \Delta\phi = 0.087 \times 0.087$, covering $|\eta| < 1.3$. Instead, the HE extends the coverage up to $|\eta| < 3.0$, and features a finer granularity of $\Delta\eta \times \Delta\phi = 0.017 \times 0.017$.
An additional hadronic calorimeter layer, the HCAL Outer (HO), is placed outside the solenoid magnet to increase the total number of interaction lengths to approximately 10, thereby reducing the leakage of high-energy hadronic jets into the muon system, decreasing non-Gaussian tails in the energy resolution, and lowering the probability of hadronic jets being misidentified as muons. Moreover, beyond the endcaps, the forward calorimeter (HF) is located at a distance of 11.2 m from the interaction point on each side of the detector and extends the coverage up to $|\eta| \approx 5.2$, ensuring good hermeticity. In the HF, quartz fibers embedded in a steel absorber are used to collect the energy of the developing showers and generate the signal through the Cherenkov effect. The structure of the HCAL is illustrated in Figure 2.8.

2.2.5 Magnet

The tracking system and calorimeters are installed within the solenoid, which is in turn enclosed by the segmented iron return yoke, a dodecagonal, three-layer structure extending to approximately 14 m in diameter, which guides the magnetic flux and houses the muon detectors between its layers. The CMS superconducting solenoid produces an approximately uniform axial magnetic field used to bend charged particle trajectories, thereby enabling the determination of the charge sign and p_T from the track curvature. The magnet is composed of refrigerated niobium–titanium superconducting coils, with an overall length of about 13 m and an internal diameter of roughly 6 m. While the solenoid was originally specified for 4.0 T, it is operated at 3.8 T to improve longevity and operational stability [45]. The nominal intensity of 3.8 T is established in the volume of the inner tracker and calorimeters, while it is reduced to 1.5 - 2 T in the volume defined by the return yoke. Figure 2.9

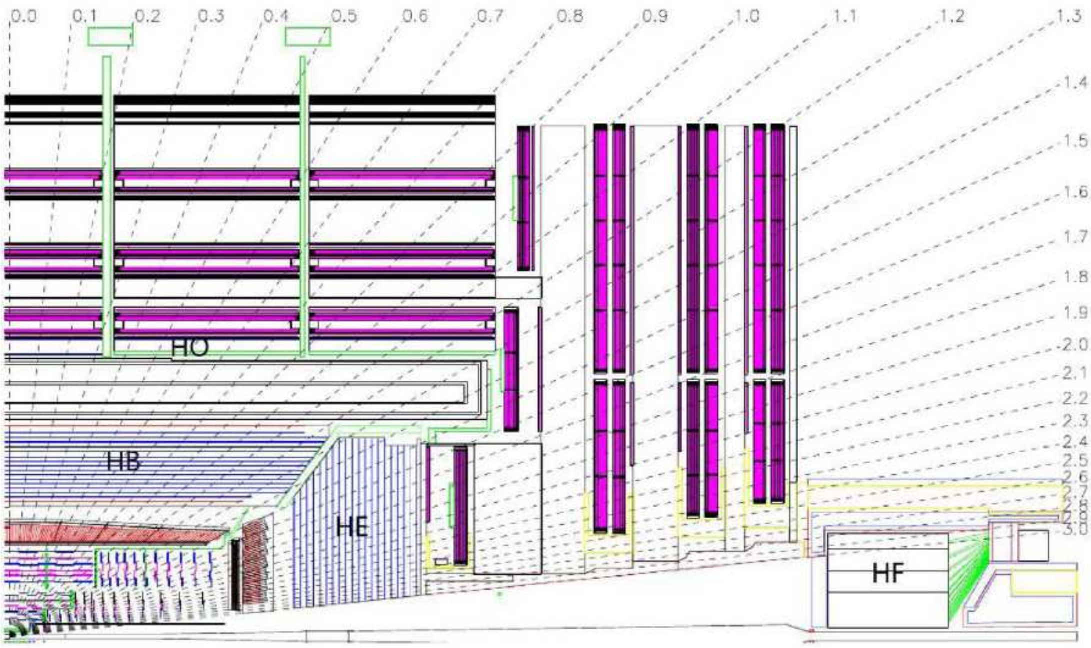

Figure 2.8: Longitudinal view of the geometry in a HCAL quadrant. The location and η coverage of HB, HO, HE, and HF are illustrated [44].

shows a representation of the CMS magnet field [45] while Table 2.2 summaries the parameters of the solenoid.

Parameter	Value
Field	3.8 T
Inner diameter	5.9 m
Length	12.5 m
Number of solenoid turns	2168
Current	18.16 kA
Stored energy	2.3 GJ

Table 2.2: Parameters of the CMS superconducting solenoid.

The charged particles momentum is estimated from the particles trajectory inside the solenoid, and the momentum resolution can be approximated, under the assumption of an approximately uniform magnetic field and in the high-p_T regime ($\geq$ 20 GeV), where the sagitta measurement uncertainty dominates over multiple scattering and energy-loss effects, by

$$\frac{\sigma_{p_T}[GeV]}{p_T[GeV]} = s[m]\frac{8\,p_T[GeV]}{0.3\,|B|[T]R^2[m]} \tag{2.8}$$

where |B| is the magnetic field intensity, s is the sagitta of the particle trajectory and R is the solenoid radius. Therefore strong field and large radius are an efficient approach to reach optimal momentum resolution.

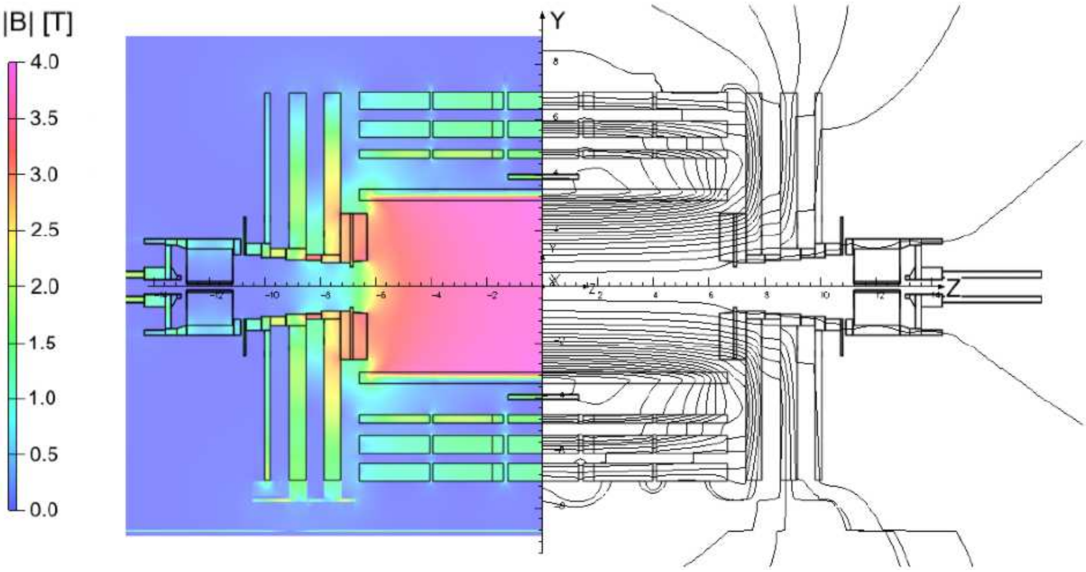

Figure 2.9: Predicted magnetic field intensity |B| (left) and magnetic field lines (right) on a longitudinal section of the CMS detector for the underground configuration at a central magnetic flux density of 3.8 T. The field-line density increases with coordinate |Y|, reflecting the stronger magnetic field, and each line represents a flux increment of 6 Wb [45].

2.2.6 Muon System

Muons produced in collisions at the LHC have minimal energy loss rates. As a consequence, they traverse the calorimeters and the solenoid volumes without being stopped and are identified and measured in the muon system located inside the return yoke. The main goals of the muon system are the identification of muons and a precise measurement of their momentum, with the help of the information coming from the tracker. The muon system also works as trigger for events which involve muons and it provides a precise time measurement of the bunch crossing [46].
The muon system relies on four kinds of gaseous detectors: drift tubes (DT), cathode strip chambers (CSC), resistive plate chambers (RPC) and gaseous electron multiplier (GEM). The longitudinal view of a quarter of the muon system is given in Figure 2.10. The barrel extends up to $|\eta| < 1.2$, the endcaps up to $|\eta| < 2.4$. All the detectors are organized in 4 different stations placed at different R or η, depending if they are placed in the barrel or in the end-cap regions.

Drift tubes

The DT fundamental sensing element is a rectangular drift cell (typical section 42×13 mm^2) with a central anode wire (length $\approx 2-3$ m) and cathode strips on the short sides; the cathode strips are conductive metal segments that collect the induced charge from the drifting electrons and allow precise measurement of the coordinate along the wire. The cells are filled with an Ar$-$CO$_2$ mixture (85%$-$15%) and operated at a gas gain of order 10^5. When an ionizing particle traverses the gas it produces electron$-$ion pairs. The liberated electrons drift toward the anode wire under the chamber's strong electric field and trigger secondary ionization, producing an avalanche. The resulting charge cloud induces a spatially distributed signal on the segmented cathode strips. The measured drift

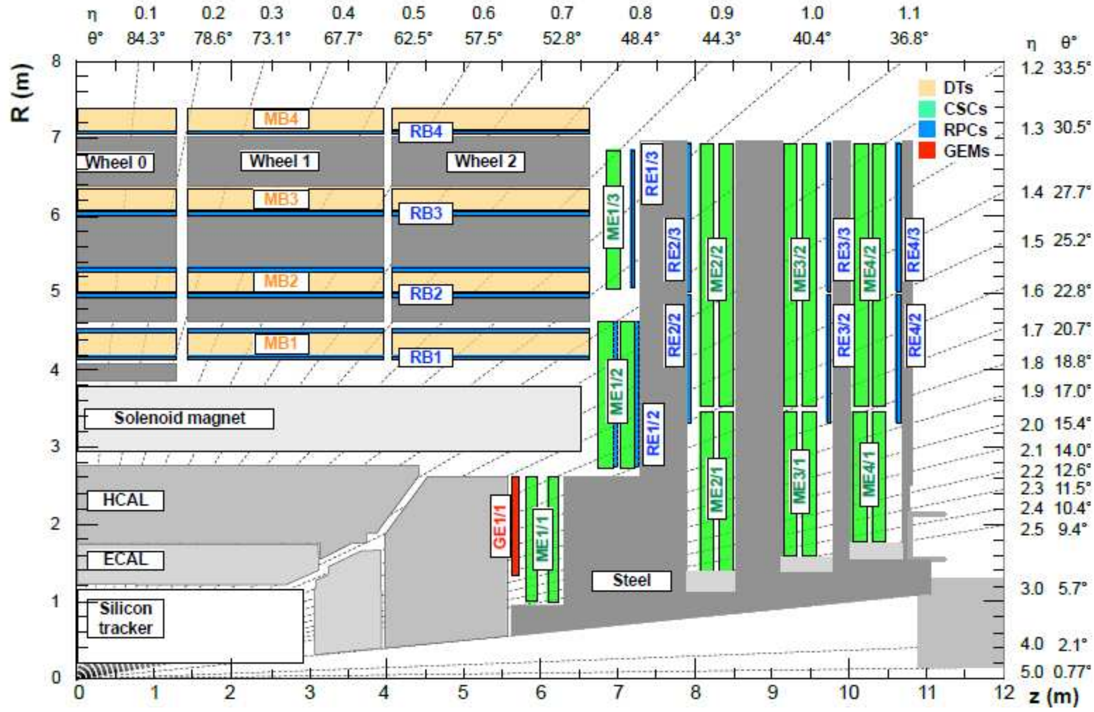

Figure 2.10: A quadrant of CMS muon system. The four different sub-detectors are highlighted: DT (in yellow) are installed in the muon barrel, CSC (in green) and GEM (in red) are placed in the muon endcap and Resistive Plate Chambers (in blue) are present in both, barrel and endcaps (labeled as RB and RE). The dark gray areas are the steel flux-return disks of the magnet [47].

time provides the transverse distance from the wire, while longitudinal segmentation/readout yields the coordinate along the wire. Maximum drift times are of order 400 ns, with a time resolution of 2 ns; single-hit resolutions are typically $\approx 200\ \mu$m [48]. The section of a drift cell of a DT detector is shown in Figure 2.11.

In the CMS muon barrel (MB), the DT system is organized into chambers, which are grouped into four concentric stations around the beam line. DT chambers are further subdivided into two or three superlayers (SL), each comprising four layers of cells staggered by half a cell. The two outer SLs have wires parallel to the beam axis (providing $r-\phi$ measurements), while the inner SL, when present, has wires orthogonal to the beam for the z measurement. Chamber segment resolutions is of order $80 - 120\ \mu$m [48]. Each of the three inner stations contains 60 chambers, distributed across five wheels, while the outermost station contains 70 chambers, also arranged on five wheels. The outermost station lacks the z SL and therefore measures only the $r-\phi$ coordinate. The DT configuration in one of the barrel wheels is shown in Figure 2.12.

Cathode Strip Chambers

The CSC are trapezoidal multi-wire proportional chambers whose cathode plane is segmented into strips that run perpendicular to the anode wires. When an ionizing particle traverses the gas volume it creates electron–ion pairs; the liberated electrons drift toward the anode wires and, in the cham-

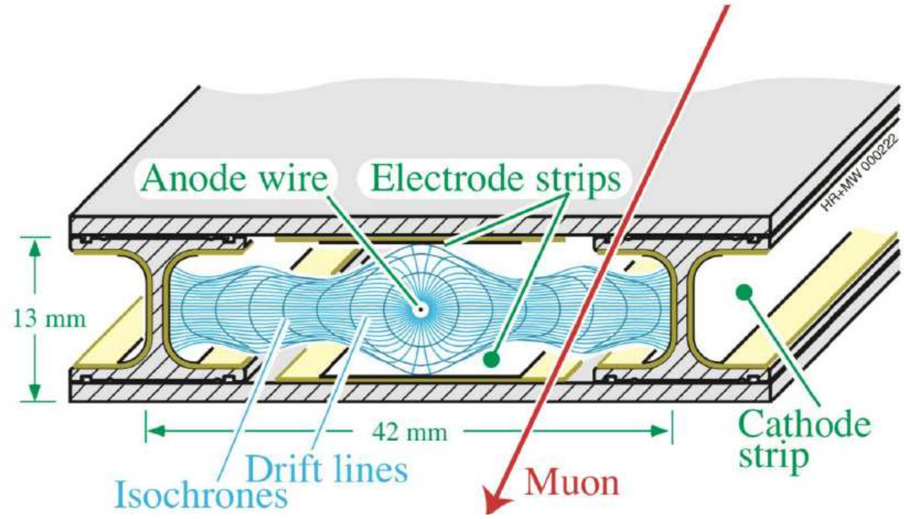

Figure 2.11: Section of a DT drift cell showing the anode wire, cathode strips (which collect the induced charge and provide the longitudinal coordinate), the drift lines and the isochrones (contours of equal drift time). The plates at the top and bottom of the cell are at ground potential. Electrode strips below and above the anode wire of the cell are needed to shape the electric field lines. Voltages: wires +3600 V, strips +1800 V, cathodes -1200 V [49].

ber's strong electric field, initiate avalanches of secondary ionization. The avalanche charge induces a distributed signal on the cathode strips, and the particle position along the wire is reconstructed by interpolating the fractions of charge collected on neighboring strips, giving a precise measurement of the coordinate orthogonal to the wires, as shown in Figure 2.13. The CSC instrument the two muon endcaps (ME) and are mounted on four station disks per endcap (ME1−ME4), installed between the iron plates of the return yoke, where the residual magnetic field is intense ($\approx 1.4-2.3$ T) as seen in Figure 2.9, and the particle rate ranges from few tens Hz/cm^2 to several hundred Hz/cm^2. CSC chambers are arranged in concentric rings in each station (ME1 contains three rings, subsequent stations two rings), with individual rings formed from either 18 or 36 trapezoidal chambers, ensuring complete coverage in the $r-\phi$ plane by using narrower ($\sim 10°$) chambers (36 per ring) in the inner rings and wider ($\sim 20°$) chambers (18 per ring) in the outer rings. The use of narrower chambers in the inner rings provides finer ϕ granularity to improve angular resolution and trigger performance in regions of higher particle rates, while wider chambers are used in the outer rings. The cathode plane is segmented into finely pitched radial strips (80 strips per cathode plane) with strip pitches varying approximately between 2.2 mrad and 4.7 mrad in ϕ, while the orthogonal anode wires are spaced by about $2.5-3.16$ mm along the plane. The nominal operating gas mixture is $Ar-CO_2-CF_4$ (40%–50%–10%), with typical gas gains of order 7×10^4. The chamber-level $r-\phi$ spatial resolution ranges from approximately 70 μm to 210 μm moving from the innermost to the outermost chamber, while the complementary coordinate measured from the anode wires is much coarser, typically of order a few millimeters. Timing capabilities at the few-nanosecond level ($\sim 2-3$ ns) are achieved by reading out the timing information from the anode wires. Both the spatial and timing resolutions allow unambiguous bunch-crossing assignment and fast trigger signals [48].

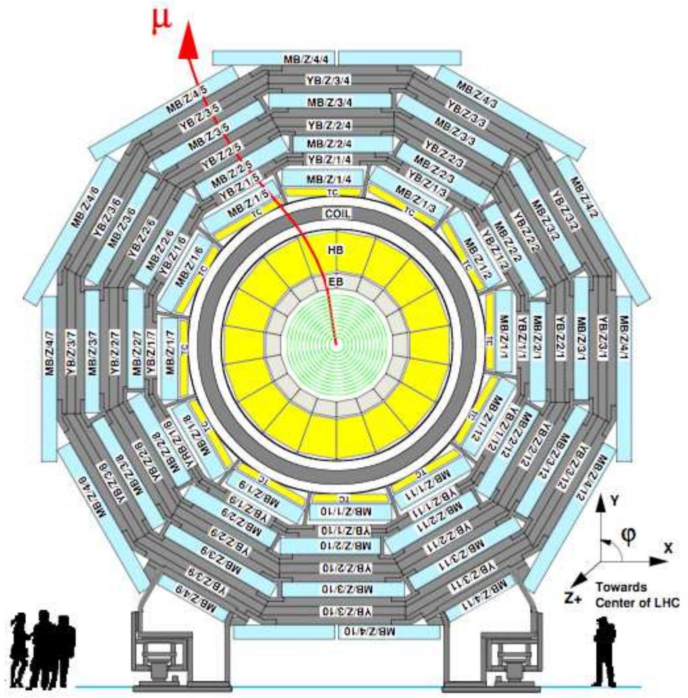

Figure 2.12: Layout of the CMS barrel muon DT chambers within a single wheel [50].

Resistive Plate Chambers

RPC are installed in both barrel and end-cap to provide a very fast timing response that complements the DT and CSC for trigger purposes. An RPC is a planar gaseous detector formed by two bakelite plates coated with a thin graphite film that serve as the electrodes, separated by a narrow gas gap (≈ 2 mm) filled with a gas mixture of $C_2H_2F_4$, i-C_4H_{10} and SF_6 ($96.2\% - 3.5\% - 0.3\%$). Typically humidity ($0.7\% - 0.8\%$) is added to control bakelite resistivity and stability. CMS uses a double-gap chamber design in which two gas gaps face a common readout plane segmented into strips; a high voltage of order 9.6 kV is applied to the outer graphite surfaces. When a muon passes through the chamber, electrons are knocked out of gas atoms. These electrons in turn hit other atoms causing an avalanche of electrons. The design is shown in Figure 2.14.
In the barrel, the RPCs are segmented as the DT chambers, ensuring geometrical correspondence and straightforward matching between the two detector systems. In particular RPC are embedded in six layers within the iron yoke, four layers adjacent to the first two DT stations, and two layers at the inner side of the third and fourth DT stations. Each barrel chamber typically carries 96 readout strips oriented parallel to the beam, with strip pitches increasing from ≈ 2.1 cm to ≈ 4.1 cm. In the endcaps the RPC are arranged in four stations (matching the CSC layout); trapezoidal chambers employ radial readout strips whose lengths range from ≈ 25 cm to ≈ 80 cm with strip pitches of order $0.7 - 3.0$ cm. RPCs deliver spatial resolution of order 1 cm and timing resolution ≈ 2 ns, well suited to fast trigger signals [48].

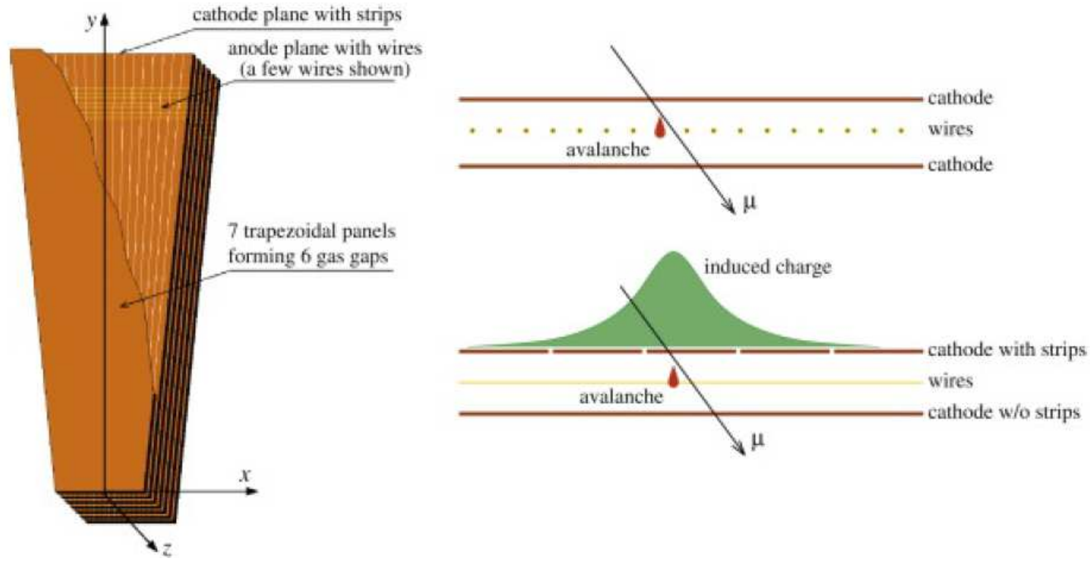

Figure 2.13: Schematic view of cathode strip chamber (left) and the principle of operation (right) seen from strips (right-top) and wires (right bottom) view [46].

Gaseous Electron Multiplier

During the High-Luminosity phase, the muon flux in the forward region is expected to reach $5kHz/cm^2$ in the innermost layer, motivating the deployment of high-rate and radiation-tolerant detectors [52]. To this end, large-area triple-GEM chambers were installed in CMS: the GE1/1 stations. It covers approximately $1.55 < |\eta| < 2.18$ and comprises two layers of 36 triple-GEM chambers per endcap, each chamber subtending a $10°$ azimuthal sector, are located at $z \approx 566 - 574$ cm and spans a radial region $\approx 145 - 230$ cm. Chambers are matched to form a superchamber, providing two measurement planes and maximizing the detection efficiency for the station. The chambers are manufactured in two different sizes, as shown in Figure 2.15. The odd-numbered chambers in GE1/1 are slightly longer to optimize η coverage while fitting within the spatial limitations set by the support structure.

The basic building block of a GEM detector is the GEM foil: a thin polyamide substrate coated with copper on both faces and etched with a regular array of microscopic holes. Typical dimensions are ≈ 50 μm for the polyamide plus ≈ 5 μm of copper cladding; the hole pattern has a pitch of ≈ 140 μm [47]. A single-GEM detector consists of a gas volume bounded by a cathode electrode on one side and a readout anode on the other, with a single GEM foil placed between them. A uniform electric field is established between the cathode and the top surface of the GEM foil. When a charged particle traverses the detector, it ionizes the gas in the drift region, producing primary electrons that drift toward the GEM under the influence of this field. A voltage applied across the GEM foil creates very strong electric fields inside the holes. As the primary electrons enter these holes, they undergo avalanche multiplication. Part of the avalanche charge is collected on the GEM foil electrodes, while the remaining electrons move towards the anode thanks to an additional electric field, applied between the bottom of the GEM foil and the anode. The charge is then collected by readout strips or pads and converted into an electrical signal.

A CMS triple-GEM is a micro-pattern gaseous detector formed from a drift electrode, three GEM

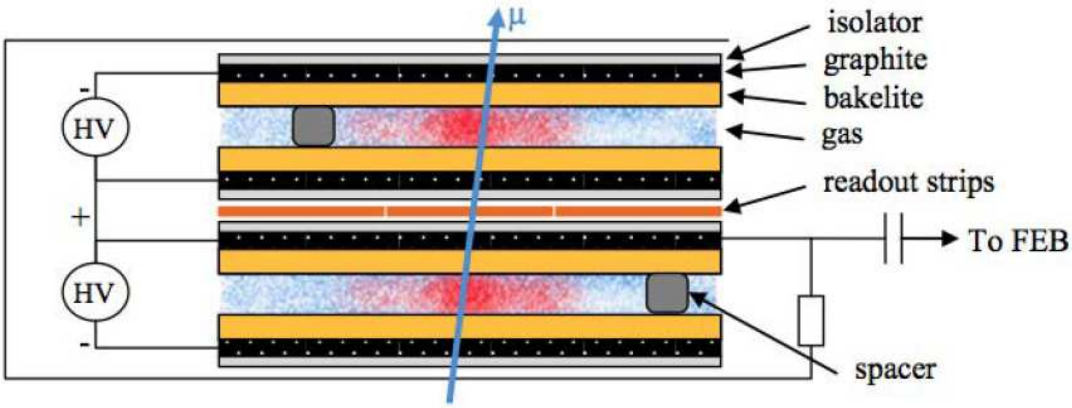

Figure 2.14: Cross-sectional schematic of a two-gap RPC. Two resistive electrodes are arranged symmetrically around a central pickup printed circuit board, forming two gas gaps. The pickup board hosts parallel readout strips that run along the length of the detector and are read out from both ends. The figure indicates the high-voltage connections applied to the electrodes to polarize the gas gaps [51].

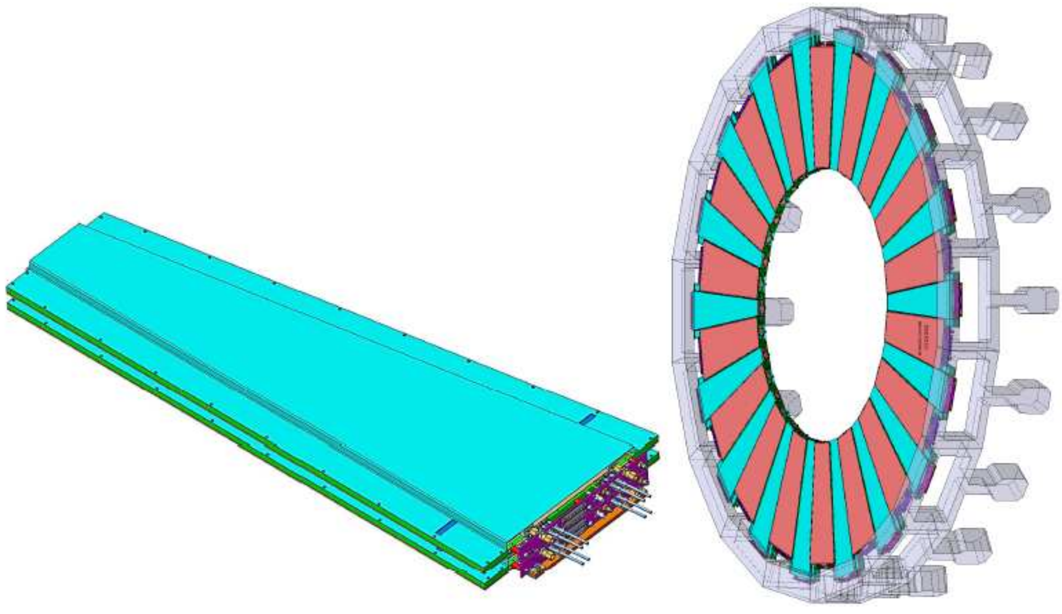

Figure 2.15: Left: two GEM chambers paired to form a superchamber. Right: long and short chambers are combined to optimize coverage within the mechanical constraints of the endcap region [47].

foils and a readout board. A scheme is shown in Figure 2.16. The readout plane carries radially oriented strips (strip pitch $\approx 0.6-1.2$ mm) segmented into up to 8×3 $\eta-\phi$ partitions with 128 strips per partition. Cascading three foils yields a high total gain (stable operation in CMS at 10^4-10^5) while keeping the charge density per hole low and reducing discharge risk. GE1/1 chambers operate with an Ar/CO_2 (70%−30%) mixture.
GEM technology is well suited to the forward environment: it tolerates very high rates, provides per-layer spatial resolution of order $250-500$ μm and timing below 10 ns; the GE1/1 station provides an angular resolution of $\approx 300\mu$rad, significantly improving endcap muon reconstruction.

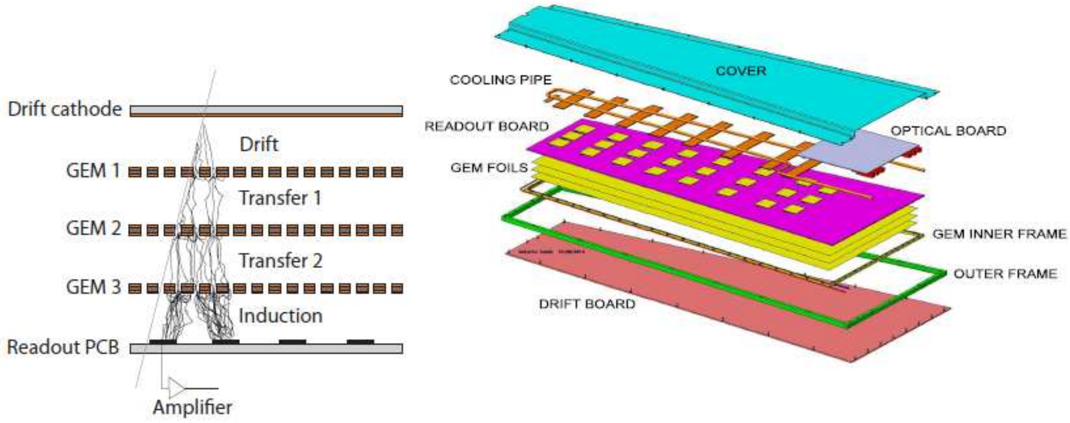

Figure 2.16: Left: three GEM foils detector scheme. Avalanche multiplication take place inside the holes of the three foils ensuring high gains. Right: schematic exploded view showing the structural layout of a Triple-GEM chamber [47].

2.3 Trigger and Data Acquisition

pp collisions take place at the center of the CMS detector every 25 ns, producing an enormous amount of information, corresponding to about 100 terabytes of data per second [53]. At present, no technology is capable of reading out, storing, and analyzing such data volumes in real time. Fortunately, the vast majority of collisions involve low-energy interactions that are not relevant for the CMS physics program. Under these extreme conditions, a highly efficient trigger system is required to reduce the event rate from 40 MHz in an optimal way. Its purpose is to discard uninteresting low-energy processes while retaining as many potentially interesting high-energy events as possible. This task is accomplished through a two-level system: the Level-1 (L1 [54]) trigger and the High-Level Trigger (HLT [55]). Both L1 and HLT rates can also be tuned through the use of prescale factors, which randomly accept only a fraction of the events passing a given selection.

2.3.1 Level-1 Trigger

The L1 trigger is a fully hardware-based system designed to provide the first selection stage in CMS. Its primary role is to reduce the input event rate from the LHC bunch crossing frequency of 40 MHz down to about 100 kHz, which can then be handled by the subsequent High-Level Trigger system. The L1 decision is made within 4 μs, based on coarse-grained information from the calorimeters and

muon detectors. Due to the stringent latency requirements, tracker information cannot be used at this stage, since full track reconstruction is too time-consuming.
The L1 system identifies so-called trigger objects, corresponding to physics signatures such as high-p_T leptons, photons, hadronic jets, and missing transverse momentum. These objects are reconstructed from trigger primitives (TPs), which are basically energy deposits in the calorimeters or hit patterns and track segments in the muon chambers.
The outputs from the calorimeter and muon trigger subsystems are then combined in the L1 Global Trigger, which applies a programmable set of selection criteria to decide whether an event is sufficiently interesting to be kept. If the decision is positive, the full event record is transferred to HLT farm for further processing; otherwise, the event is permanently discarded. A schematic overview of the L1 trigger architecture is shown in Figure 2.17.

2.3.2 High Level Trigger

The output rate from the L1 trigger is further reduced by HLT to about 2.6 kHz. Unlike the hardware-based L1, the HLT is a software system that exploits full event reconstruction to select the subset of events relevant for the wide CMS physics program. Starting with 2022, the HLT computing farm has been enhanced with graphical processing units (GPUs), enabling many reconstruction algorithms to run on both central processing units (CPUs) and GPUs. These algorithms are executed on GPUs when available, otherwise they fall back on CPUs.
The HLT processing itself is organized in two stages. In the first step, calorimeter and muon detector information is used to reduce the event rate by roughly an order of magnitude. In the second step, tracking information from the silicon tracker is reconstructed and more refined selections are applied. Thanks to its flexibility, the HLT allows the definition of customized trigger selections targeting specific physics signatures. A complete chain of L1 and HLT selections, together with their prescales, defines a so-called trigger path. The reconstruction algorithms employed are largely the same as those used offline, though optimized for speed to meet the stringent latency requirements of the online system.
If an event satisfies one of the trigger paths it is marked for permanent storage and transferred to CERN Tier-0 (T0) storage system. Then Data Quality Monitoring (DQM) checks the quality of the recorded data and labeling the data-sets either good or bad. At the end of the DQM chain, a list of certified data-sets is produced to be used later for offline analysis.

2.3.3 Run 3 trigger strategies: Scouting and Parking

The overall rate at which events can be stored and reconstructed is constrained by practical limitations, such as the available bandwidth for permanent storage and the computational resources required for offline reconstruction. To enhance the number of events recorded, the CMS Collaboration introduced two complementary strategies, known as “scouting” and “data parking”, which have been further optimized and extended for the Run 3 physics program [57].
The scouting approach addresses these limitations by reducing the size of the stored events, thereby allowing data to be collected at a higher rate. Within this strategy, two modes were developed:

- **L1 scouting system:** for Run 3, a new hardware component was integrated into the L1 to enable the recording of all events reconstructed at the L1 level, bypassing the trigger selection and operating at the full collision rate of 40 MHz. The objects recorded are not subjected to a further reconstruction and therefore exhibit limited resolution. Nevertheless, the ability to capture such a large event rate provides a unique opportunity to explore and study rare physics processes.

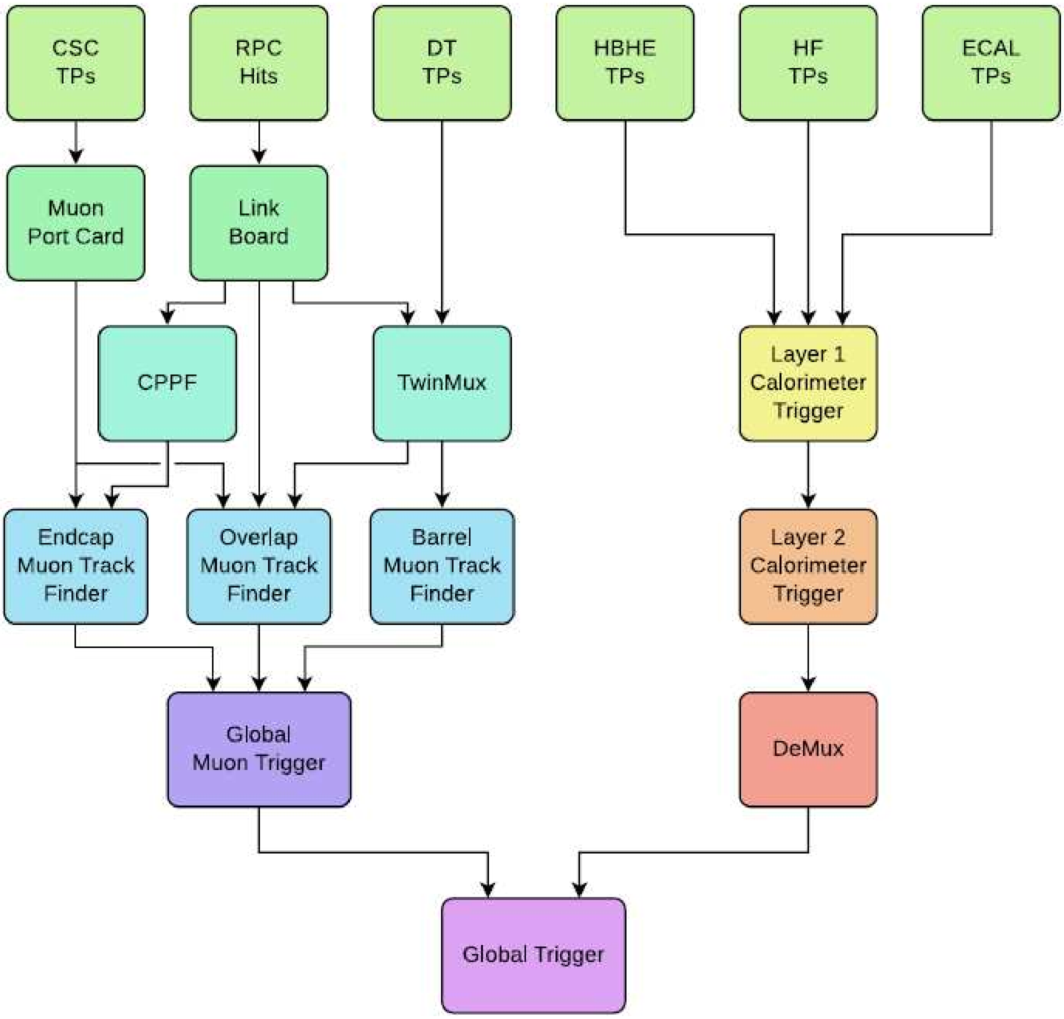

Figure 2.17: Schematic representation of the CMS L1 trigger system. A combination of TPs of CSC, RPC and DT constitutes the Global Muon Trigger while a combination of TPs of HB, HE, HF and ECAL calorimeters constitutes the Calorimeter Trigger. The Global Trigger is then obtained as combination of the Global Muon Trigger and the Calorimeter Trigger [56].

- **HLT data scouting:** the HLT data scouting strategy was first introduced during Run 1 and subsequently expanded in Run 2 and Run 3. In the current implementation, scouting events reconstructed at the HLT have an average size of only 7 kB, compared to about 1 MB for a full event. This significant reduction in event size makes it possible to record data at rates of up to 30 kHz in 2022 and 22 kHz in 2023. As for L1 scouting, the objects recorded are not subjected to a further reconstruction.

In contrast, the data parking strategy increases the amount of data written to disk by postponing the offline reconstruction step. This approach enables the storage of extra data at rates limited not by the prompt reconstruction capacity but by the CMS data acquisition bandwidth and available tape storage. Data parking was first implemented in Run 1 at a total rate of 300−350 Hz, including trigger paths for vector boson fusion (VBF) topology, Higgs boson studies, SUSY and dark matter searches, and certain di-muon triggers. Ahead of 2018, the final year of Run 2, the "B parking" strategy was introduced, focusing on B-flavor anomaly searches and collecting data at an average rate of 2−3 kHz. For Run-3, double Higgs and VBF parking strategies have been developed.

2.4 The World LHC Computing Grid

Events that pass the trigger selections and are recorded must be subsequently reprocessed for analysis. Both real and simulated data are managed within the World LHC Computing Grid (WLCG), which is also responsible for transferring and processing them [58]. This computing infrastructure consists of over 170 data centers among more than 40 countries, all interconnected through high-speed networks, as illustrated in Figure 2.18. The WLCG is arranged in four layers, or "Tiers", called 0, 1, 2 and 3. Each Tier is made up of several computer centers and provides a specific set of services.

Data produced by the CMS online acquisition (RAW) are first stored at T0. T0 does not usually host datasets for direct analysis; its main role is to organize the incoming material into Primary Datasets according to the trigger paths used for acquisition, and to convert RAW files into analysis-ready formats. These derived datasets are then sent to the 12 Tier-1 centers distributed worldwide. Tier-1 sites additionally store the data received from Tier-0 and carry out reconstruction, calibration and quality-screening of RAW data, as well as distributing datasets to Tier-2 centers. Tier-1 centers also retain and redistribute Monte Carlo (MC) simulations productions generated at Tier-2. Tier-2 facilities provide the computing resources where most users run analyses; individual researchers may also use local (Tier-3) resources such as departmental clusters or personal machines, which are not formally contracted to the WLCG.

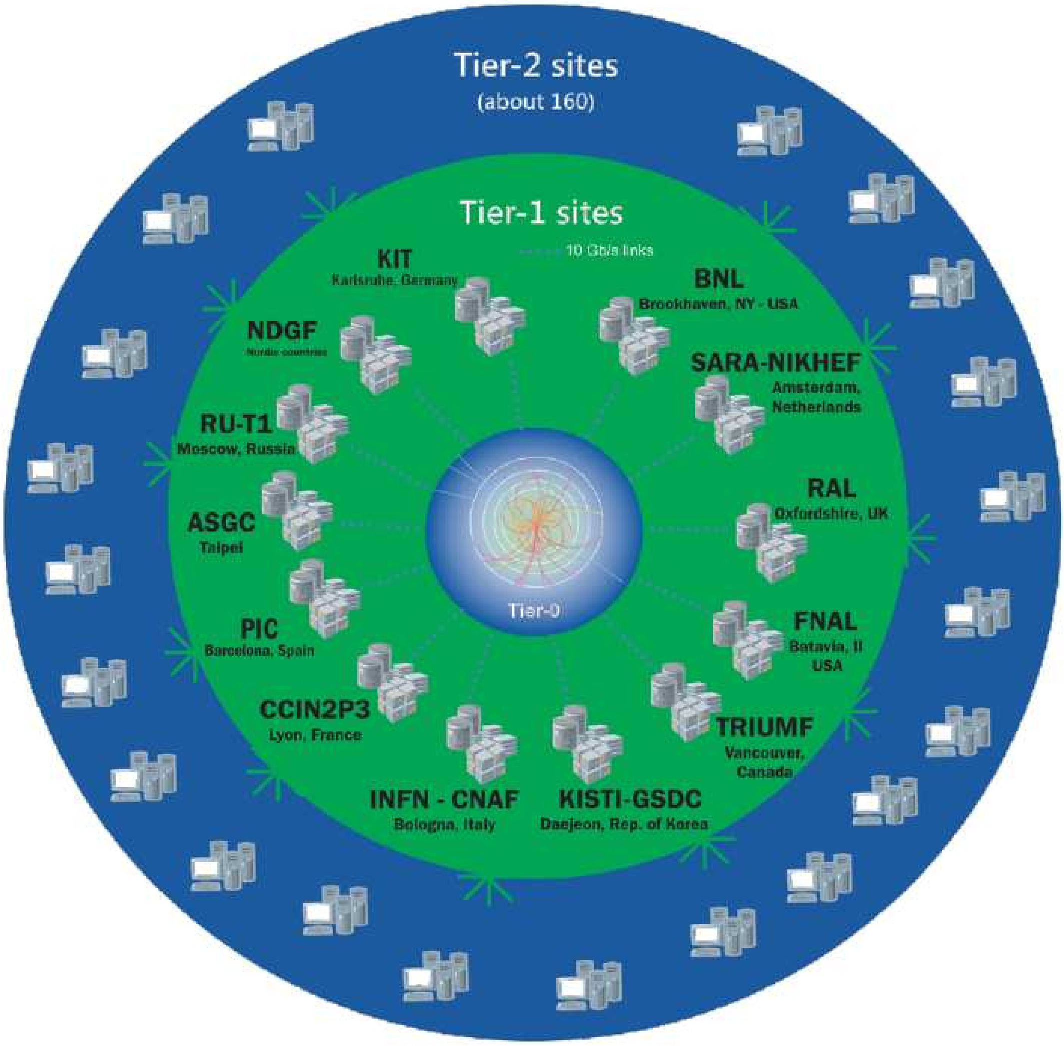

Figure 2.18: Schematic representation of the WLCG distributed infrastructure. It is organized in different levels, called Tiers, connected via high-speed networks [59]. The Russian Tier-1 site (RU-T1) is no longer participating in the WLCG.

Chapter 3

Object reconstructions in CMS

The CMS apparatus is able to detect electrons, muons, photons and hadrons as the decay products of the particles produced in collisions. Figure 3.1 shows how they are reconstructed from the different detectors. Photons are identified as ECAL energy clusters not linked to the extrapolation of any charged particle trajectory into the calorimeter. Electrons are identified as a charged particle track whose extrapolation matches at least one ECAL energy cluster. HCAL clusters are associated to charged hadrons if the clusters match at least on hit in the tracker layers, otherwise the clusters are associated to neutral hadrons. Muons are identified as a track in the tracker system compatible with several hits in the muon system. An ensemble of signals in the detectors, associated to a specific type of particle, is named physics object. In this chapter the most relevant aspects of the CMS object reconstruction will be described. Tracks and clusters are combined by the Particle Flow (PF) algorithm [60] to identify each final-state particle, referred to as PF candidates, and then reconstruct the particle properties on the basis of this identification. Jets are subsequently reconstructed by clustering the PF candidates using sequential recombination algorithms, grouping the particles originating from the hadronization of quarks and gluons.

3.1 Tracker reconstruction

Track reconstruction aims to estimate the momentum and position parameters of the charged particles from the hits collected by the tracker detector. The tracking software at CMS is commonly referred to as the Combinatorial Kalman filter (CKF) [61]. The track reconstruction consists in four stages: seed generation to find a few hits compatible with a charged-particle trajectory; track finding (or pattern recognition) to gather hits from all tracker layers compatible with charged-particle trajectory; track fitting to determine the charged-particle properties (*e. g.* origin, momentum, direction); track selection to set quality flags and discards tracks that fail certain specified criteria.

3.1.1 Seed generation

The seeds define the starting trajectory parameters of potential tracks. In the quasi-uniform magnetic field of the tracker, charged particles follow helical paths and therefore five parameters are needed to define a trajectory. A general helix would require six parameters, however the magnetic field fixes the relation between curvature and momentum, removing one degree of freedom. Extraction of these five parameters requires either three 3-D hits, or two 3-D hits and a constraint on the origin of the trajectory based on the assumption that the particle originated near the beam

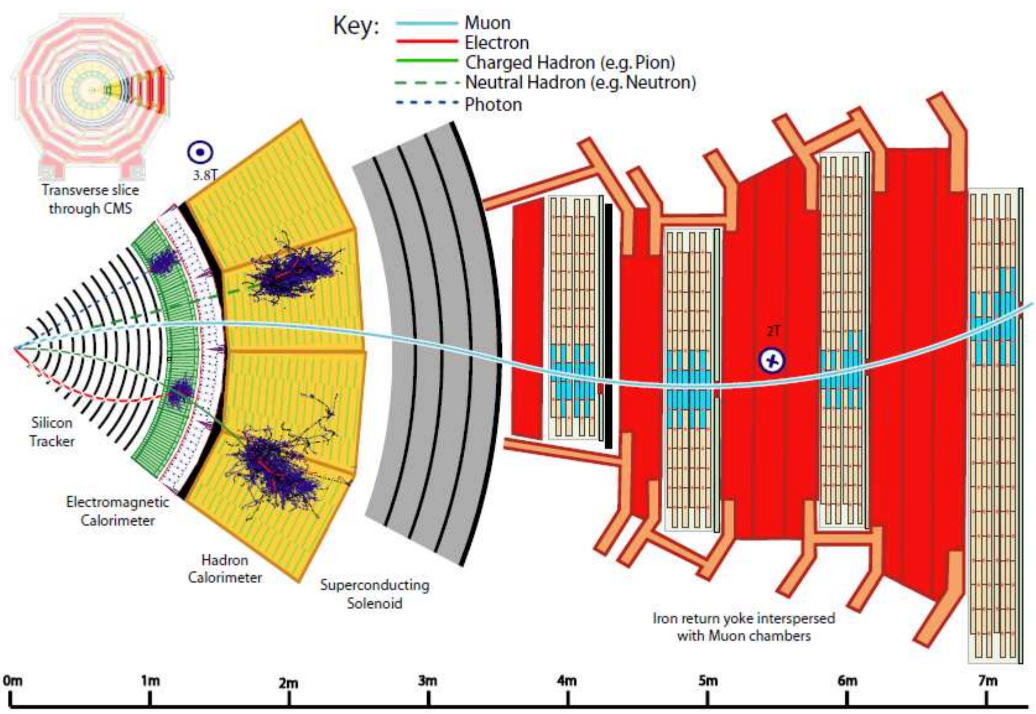

Figure 3.1: A sketch of the specific particle interactions in a transverse slice of the CMS detector, from the beam interaction region to the muon detector. The muon, electron and pion produce hits in the tracker system, the photon and the electron produce an EM shower in the ECAL, the pion and the neutron produce clusters in the HCAL, the muon is the only particle producing hits in the muon system. The muon and the charged pion are positively charged, and the electron is negatively charged [60].

spot (BS). Pixel-based seeds (pixel tracks) are built first and serve as primary seeds for both HLT and the first offline reconstruction. Complementary seeding includes triplets and pairs using pixel and inner-strip layers applying cellular-automaton (CA) based seeding [62]. The CA algorithm creates hit doublets (cells) for each pair of layers and then compute the compatibility between adjacent cells. Seed selection enforces configurable constraints on minimum p_T and on transverse/longitudinal impact parameters (minimal distance) relative to a BS or pixel-vertex constraint.

3.1.2 Track finding

Track finding in CMS uses a layer-by-layer CKF that evolves track seeds through the cycle *predict* $\rightarrow$ *search* $\rightarrow$ *update* $\rightarrow$ *branch/prune*. The seed is propagated to downstream surfaces taking into account the magnetic field presence and the effects of energy loss and multiple scattering. Modules are selected by overlapping the propagated seed uncertainty with module surfaces; hits are corrected for instrumental effects and tested with a Mahalanobis distance [63]. Each compatible hit generates a new candidate extension; to control combinations the candidates are scored (normalized χ^2, hit count, ghost-hit penalties) and only the top five ones are retained per step. Missing measurements are handled with missing-hit allowances, an outward build is often followed by an inward smoothing pass to recover inner hits, a trajectory cleaner removes duplicates before the final KF produces the track parameters and covariance.

3.1.3 Track fitting

For each trajectory, the track-finding stage yields a list of hits and an initial estimate of the track parameters. Because this estimate may be biased by constraints applied during seeding (*e.g.* a BS constraint), the trajectory is refitted with a KF–based procedure to obtain unbiased, optimal parameters and covariances. The KF is initialized at the location of the innermost hit, with the trajectory estimated from the innermost (typically four hits) on the track. The algorithm then proceeds in an iterative way through the full list of hits, from the inside outwards, updating the track trajectory estimate sequentially with each hit. For each valid hit, its estimated position uncertainty is reevaluated using the current values of the track parameters.

3.1.4 Tracking performance

The tracking performance on data and Monte Carlo (MC) simulations is evaluated on events selected by the triggers requiring only the presence of colliding bunches (ZeroBias). Figure 3.2 shows the tracks p_T distribution for both data and MC from ZeroBias events. The agreement is within 10% for tracks with $p_T < 20$ GeV, while it is within 20% for $p_T > 20$ GeV. The tracking reconstruction efficiency is measured in data applying the tag-and-probe technique, on events with a Z boson decays into a pair of muons. Figure 3.3 shows tracking efficiency as a function of ϕ using 2023 data. The tracking efficiency is about 99.9% for tracks associated with hits in the muon system in the whole muon pseudorapidity acceptance.

3.2 Vertices

Vertex reconstruction [66] is the process of inferring the spatial points at which charged particles were produced from the measured tracks recorded by the silicon tracker. Two classes of vertices are distinguished: primary vertices (PVs), which correspond to proton–proton interaction points, and secondary vertices (SVs), which arise from decays of long-lived hadrons (*e.g.* b- and c-hadrons),

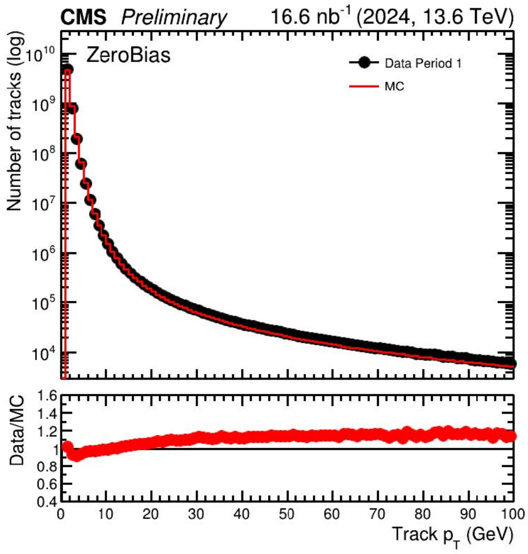

Figure 3.2: Number of tracks in data and MC as function of the track p_T for ZeroBias events corresponding to the 14/04-01/05 2024 data-taking. Lower plots shows the data/MC ratio [64].

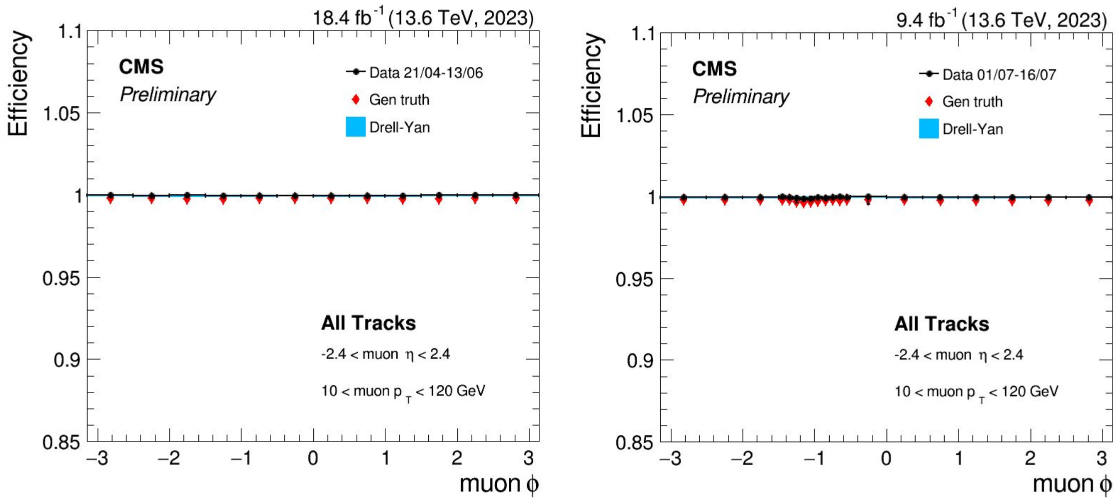

Figure 3.3: Tracking reconstruction efficiency measured for (black dots) and MC Drell-Yan (light blue rectangles) as function of ϕ for 2023 pre-BPix (left) and for 2023 post-BPix (right), see Subsection 2.2.2. The tracking efficiency measured using generator level muon particles is shown in red diamonds. The loss in performance in the ϕ region corresponding to the BPix failure is negligible. The data/MC agreement is within 1% [65].

hadronic interactions with detector material, or BSM long-lived particles. Precise vertex reconstruction is crucial for many physics tasks: identification and calibration of jets, b-tagging, reconstruction of displaced decays, and PU mitigation. Vertex reconstruction consists of the following three steps.

1. **Track selection.** The vertex reconstruction receives as input reconstructed tracks that satisfy a set of quality requirements designed to exclude poorly measured and non-prompt trajectories. Typical offline working points (WPs) used in CMS require tracks with a minimum transverse momentum ($p_T > 0.5$ GeV), a limited normalized χ^2 (< 20), and minimum hit requirements (≥ 7 total tracker hits of which ≥ 2 are pixel hits). Tracks used for the PV reconstruction are furthermore required to be compatible with the beam line, by imposing a transverse impact parameter significance with respect to the BS of $|d_0|/\sigma_{d_0} < 5$, where d_0 is the minimum distance in the transverse plane between the track and the beam line and σ_{d_0} is its associated uncertainty.

2. **Clustering.** The selected tracks are then clustered on the basis of their z-coordinates at their point of closest approach to the center of the BS by means of the deterministic annealing (DA) algorithm [67].

3. **Fitting.** Vertex candidates with at least two associated tracks are fitted using an adaptive vertex fitter (AVF) [68] which assigns per-track weights based on the track–vertex compatibility and iterates to convergence. The AVF returns the best estimate of the vertex 3D position, its covariance matrix, the fit χ^2 and the effective number of degrees of freedom; the fitted vertex is subsequently subject to quality requirements (e.g. minimal ndof, χ^2 probability, and distance from the beam line) before being accepted for physics use.

When multiple vertices are reconstructed, the PV corresponding to the hard scatter is selected using the sum of squared transverse momenta criterion: the vertex whose associated tracks have the largest $\sum_i p_{T,i}^2$ is taken as the primary hard-scatter vertex. SVs displaced from the primary interaction point are reconstructed with a dedicated inclusive algorithm. Inclusive Secondary Vertex Finder (IVF) [69] is an inclusive, seed-based algorithm that identifies tracks with large impact-parameter significance and groups them into clusters that are geometrically consistent with a common decay point. Each cluster is fitted with an adaptive fitter. IVF was developed for b-tagging and offline SV reconstruction and remains a standard offline tool because it is effective at reconstructing SVs inside jets. Tracks used for seeding are typically required to have $p_T > 0.8$ GeV, a minimum number of hits (≥ 6 hits), and a three-dimensional impact parameter significance with respect to the PV $|d_{3D}|/\sigma_{d_{3D}}>1.2$, where d_{3D} is the minimum 3-dimensional (3D) distance and $\sigma_{d_{3D}}$ is its associated uncertainty.

3.3 Photons

Photons [70] deposit the great majority of their energy in ECAL and are therefore reconstructed from ECAL energy clusters that are not matched to charged-particle tracks. The crystal signals are reconstructed with a template-fitting algorithm (multifit) that models the pulse shape including out-of-time contributions in order to mitigate PU [71]. Clusters produced by an electromagnetic shower are merged using superclustering procedure with machine learning (ML) assisted variants. The primary ML-based superclustering approach adopted for Run-3 is DeepSC [72]. The barrel–endcap transition region, $1.444 < |\eta| < 1.566$, exhibits degraded energy resolution and is frequently excluded from precision analyses. The fiducial region terminates at $|\eta| = 2.5$ where the tracker coverage ends. Photon reconstruction proceeds through three principal stages:

- Calibration and time-dependent corrections: the raw calorimeter signals are calibrated, pedestal-subtracted and corrected for changes in crystal transparency, using automated laser-monitoring information to maintain a stable energy scale over time and varying detector conditions [73].
- Superclustering: seed clusters are merged into DeepSC superclusters in order to collect the full electromagnetic shower and recover energy from photon conversions and spread along ϕ due to the magnetic field.
- Energy corrections / regression: the summed supercluster energy is corrected using parameterized or multivariate (MVA) regression techniques that account for shower containment, upstream material, cluster geometry, and local detector response, providing the best estimate of the photon energy and its associated uncertainty [74].

3.4 Electrons

Electrons [70] deposit almost all of their energy in the ECAL while also producing hits in the silicon tracker; their energy is therefore reconstructed by combining calorimetric and tracking information. Because electrons radiate bremsstrahlung inside the tracker, their ECAL energy is typically distributed among multiple crystals and clusters; dedicated superclustering algorithms merge these clusters to recover the original electron energy [72]. Electron track reconstruction uses a Gaussian-sum filter (GSF) to model the non-Gaussian bremsstrahlung energy losses and to provide robust track parameter estimates [75]. Seeding is performed either ECAL-driven (from superclusters) or tracker-driven from KF seeds, the selected seeds are fitted with the GSF; the reconstructed GSF track is then associated to the supercluster. The calibrated supercluster energy and the GSF track momentum are combined, often via a regression or weighted linear combination, to obtain the best estimate of the electron momentum and its uncertainty. A cut-based algorithm which exploit shower-shape measurements, track–cluster matching, and isolation variables is performed to identify signal electrons while rejecting background [70]. The identification WPs are shown in Figure 3.4.

3.5 Muons

Muons [48] are minimum-ionizing charged particles that traverse the full CMS detector; they are therefore reconstructed by combining independent track fits from the silicon tracker and from the muon system. Tracker muon tracks are produced with the iterative tracking sequence, while standalone muon tracks are built from muon-chamber segments. Standalone muon tracks and tracker tracks are matched and merged to form global muon candidates using KF fits while tracker-only extrapolations produce tracker muons for cases with limited muon-system information. Multiple re-fitted track hypotheses (*e.g.* tracker-only, global, or specialized refits) are compared and the Tune-P logic [48] picks the most reliable result. This procedure suppresses large outliers in the momentum resolution.
Muon isolation is estimated through the following quantity [48]:

$$I^{PF} \equiv \frac{1}{p_T^{\mu}} \left[\sum_{\substack{charged\\hadrons}} p_T + max \left(0, \sum_{\substack{neutral\\hadrons}} E_T + \sum_{photons} E_T - \frac{1}{2} p_T^{PU} \right) \right] \tag{3.1}$$

where the summation is performed on those particles laying in a cone of $\Delta R < 0.4$ around the muon, p_T^{μ} is the p_T of the muon, E_T is the transverse energy and p_T^{PU} is the neutral PU contribution

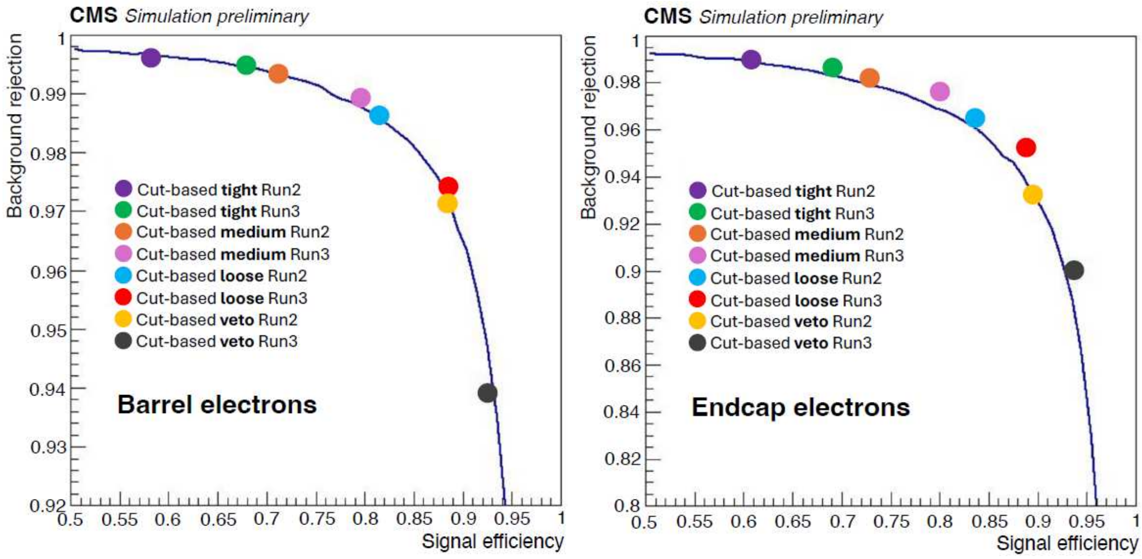

Figure 3.4: Receiver Operating Characteristic (ROC) curve for the cut-based electron identification for barrel (left) and endcap (right) electrons. Signal (background) consists of reconstructed electrons kinematically (not) matching generator-level electrons. The signal efficiency and background rejection defining the various WPs, veto (∼95%), loose (∼90%), medium (∼80%), and tight (∼70%) are also highlighted, with the numbers in parentheses representing the signal efficiency averaged over the detector acceptance [74].

to be subtracted for a correct estimate of the neutral component[1]. Muon identification classifies candidates based on track quality, hit multiplicity, segment compatibility and impact-parameter criteria, and identification efficiencies are measured in data with tag-and-probe methods [48]. As an example, Figure 3.5 shows the efficiency of muons satisfying the loose identification WP for Run-3.

3.6 Particle Flow Algorithm

The event reconstruction at CMS is based on the PF algorithm using the information from all subdetector systems for a global event description. Therefore, the objects available for the analyses are PF objects: muons, electrons, photons, charged and neutral hadrons. The reconstruction of an object first proceeds with a linking algorithm that connects pairs of nearest elements (tracks and clusters) from different subdetectors in the $\eta - \phi$ plane; the specific conditions required to link two elements depend on their nature. The link algorithm then produces PF blocks, sets of tracks and calorimeter clusters built by geometrical linking, and, in each PF block, the identification and reconstruction sequence proceeds in the following order:

1. Muon candidates are identified and reconstructed and corresponding tracks and clusters are removed from the PF block.

2. Electron candidates are identified and reconstructed with the aim of collecting the energy of all bremsstrahlung photons. Energetic and isolated photons are identified in the same step. The

[1] neutral PU $\approx 0.5\times$charged PU

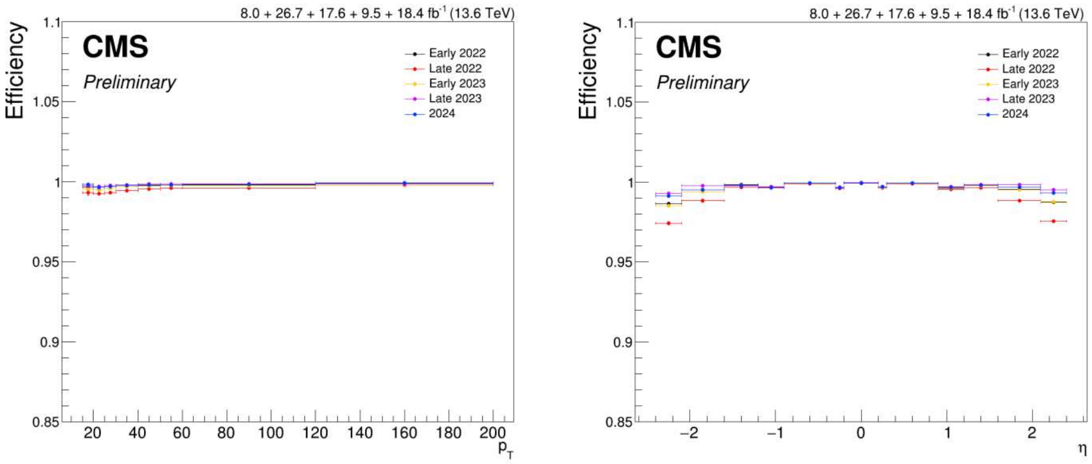

Figure 3.5: Loose identification WP efficiency as a function of p_T (left) and η (right) for 2022 pre-EE, 2022 post-EE, 2023 pre-BPix, 2023 post-BPix (see Subsection 2.2.2) and 2024. Efficiency is $\sim$1 everywhere. The small inefficiency in the forward endcaps is due to an issue in the CSC detector impacting the muon segment reconstruction efficiency [76].

corresponding tracks and ECAL or preshower clusters are excluded from further consideration.

3. Tracks with a p_T uncertainty which exceeds the calorimetric energy resolution of the cluster associated with the track are removed. This procedure effectively suppresses misreconstructed tracks not compatible with the calorimetric energy expected for that particle.

4. Clusters that are not linked to tracks are used to create PF photons and neutral hadrons. Within the tracker acceptance $|\eta| < 2.5$, all ECAL and HCAL clusters give rise to photon or neutral hadrons. Outside the tracker acceptance, all ECAL clusters linked to HCAL clusters are interpreted as neutral hadrons, while those not linked to HCAL clusters are interpreted as photons. The corresponding clusters are removed from the PF block.

At this stage the PF algorithm enters a recursive balancing logic that ensures consistency between the calorimetric energy and the sum of track momenta in the PF block. Starting from the remaining tracks and cluster energies in the PF block after the previous reconstruction steps, the algorithm repeatedly evaluates these quantities and applies the appropriate case until there are no more elements in the PF block.

1. If the sum of the cluster energies is larger than the sum of track momenta, the excess is used to create PF photons and neutral hadrons. In cases where the excess is smaller than or equal to the total ECAL energy, the excess is interpreted as a PF photon. In the remaining cases, the ECAL energy is interpreted as a PF photon and the remaining excess as a PF neutral hadron. Corresponding clusters are excluded from further consideration.

2. If the calorimetric energy is compatible with the sum of the track momenta, no neutral particle is identified. Charged hadron candidates are created for each of the remaining tracks in the PF block, with their momentum set equal to the track momenta. In cases where the sum of track momenta is compatible with the sum of cluster energies within the measured uncertainties, the

hadron momenta are redefined to the result of a fit to the tracks and clusters. At this point, all tracks and clusters in the PF block have been accounted for and the block is fully resolved.

3. If the sum of the track momenta exceeds the calorimetric energy, the remaining tracks are checked to identify and reconstruct muons inside jets, *e.g.*, those from semileptonic heavy-flavour decays or from charged hadron decays in flight, using relaxed quality criteria. Their associated tracks are then removed from the PF block. If the total track momentum of the tracks remaining in the block still exceeds the sum of cluster energies, candidate fake tracks, *e.g.*, spuriously reconstructed that do not correspond to a real charged particle, are identified and removed iteratively. Tracks are ranked (*e.g.*, by increasing transverse momentum uncertainty σ_{p_T} or by increasing σ_{p_T}/p_T) and those with σ_{p_T} above a certain threshold, such as $\sigma_{p_T} > 1$ GeV, are successively discarded until the scalar sum of the remaining track p_T is less than the calorimetric energy sum.

Once all blocks have been processed and all particles have been identified, the reconstructed event is revisited by a post-processing step such as jet identification or the calculation of the missing transverse energy (MET). MET gives indications on energy and direction of invisible particles such as neutrinos.

3.7 Pileup per particle identification

Since the huge number of PU interactions reached at CMS, as shown in Figure 2.3, it is essential to separate, as effectively as possible, the latter from the products of the primary proton–proton collision. Over the years, the CMS Collaboration has introduced several techniques to mitigate the effects of PU. The method currently in use is pileup per particle identification (PUPPI) [77, 78], which assigns each particle a probability of originating from PU and scales its energy accordingly. The algorithm computes a weight between 0 and 1 for every particle. Charged particles used in the PV fit, as well as charged particles used in the first two remaining vertices, ranked by decreasing $\sum_i p_{T,i}^2$ of associated tracks, with a longitudinal distance of closest approach to the PV smaller than 0.2 cm, are assigned a weight of 1. All other charged particles used in a vertex are assigned a weight of 0. Also, charged particles not used in any vertex with $p_T > 20$ GeV are assigned a weight of 1. Charged particles not used in any vertex with $p_T < 20$ GeV and $|\eta| > 2.4$ are assigned a weight of 1 if their longitudinal distance of closest approach to the PV is smaller than 0.3 cm; otherwise, they are assigned a weight of 0. For charge particles with $p_T < 20$ GeV and $|\eta| < 2.4$ and neutral particles, reconstructed from HCAL clusters, the weight is derived from a discriminating variable α. For a particle i, α_i is defined as:

$$\alpha_i = \log \sum_{\substack{j \neq i \\ \Delta R_{ij} < R_0}} \left(\frac{p_{T,j}}{\Delta R_{ij}} \right)^2 \qquad \begin{cases} \text{for } |\eta_i| < 2.5, & j \text{ are charged particles from PV} \\ \text{for } |\eta_i| > 2.5, & j \text{ are all kinds of reconstructed particles} \end{cases} \tag{3.2}$$

where j runs over the surrounding particles, ΔR_{ij} is the separation in $\eta - \phi$ space, and R_0 is set to 0.4. If no particles are found within a cone of radius R_0 around particle i, α_i is set to 0. For the region covered by the tracker ($|\eta| < 2.5$) only charged particles associated with the PV are used, thanks to the available tracking information. Outside this region, all reconstructed particles are considered. By construction, α tends to be larger for particles located near others belonging to the PV. A signed χ^2 value is computed by comparing the α_i of each neutral particle with the median ($\overline{\alpha}_{PU}$) and RMS (α_{PU}^{RMS}) of the α_i distribution derived from charged PU particles (the ones with

PUPPI weight 0):

$$\chi_i^2 = \frac{(\alpha_i - \overline{\alpha}_{PU})\,|\alpha_i - \overline{\alpha}_{PU}|}{\left(\alpha_{PU}^{RMS}\right)^2} \tag{3.3}$$

A larger signed χ^2 value corresponds to a higher probability that the particle originates from the PV. This quantity is then used to compute the particle weight, used to scale the particle four-momentum:

$$w_i = F_{\chi^2,NDF=1}\left(\chi_i^2\right) \tag{3.4}$$

where $F_{\chi^2,NDF=1}$ is the cumulative distribution function of the χ^2 distribution with one degree of freedom. If the weight is smaller than $p_T \cdot \frac{1}{180} - \frac{1}{9}$ it is raised to this value, up to a maximum of 1. Particles with $w_i \times p_{T,i} < (A + BN_{\text{vertices}})$ GeV, where N_{vertices} is the number of vertices in the event and A and B are tunable parameters, are also removed. Usually, A and B are chosen as a function of $|\eta|$ to minimize the differences between reconstructed jet observables (such as jet energy) and their generator-level counterparts.

3.8 Jets

Quarks and gluons produced in proton–proton collisions undergo parton showers and hadronization, forming jets of collimated particles. At CMS, jets are reconstructed by clustering the product of the hadronization process using the anti-k_T algorithm [79] implemented in the FastJet package [80]. Jets clustered with a distance parameter R=0.4 (AK4) are optimized for standard quark and gluon jets, while the decay products of highly boosted particles, such as W, Z, or Higgs bosons, can be collimated into a jet with R=0.8 (AK8). Hadronically decaying τ leptons can also be reconstructed as AK4 jets, since their visible decay products are typically highly collimated. In these decays, the τ lepton ($\tau^- \to \nu_\tau\, W^{-*} \to \nu_\tau$ hadrons) produces one- or three-prong hadronic final states (*e.g.* $\tau^- \to \nu_\tau\, \pi^-\, \pi^0$ or $\tau^- \to \nu_\tau\, \pi^-\pi^+\pi^-$), often accompanied by neutral pions, while the associated neutrino escapes undetected. The jet momentum is computed as the vector sum of the momenta of its constituent particles and is, on average, within 5–20% of the true momentum over the full p_T range and detector acceptance. AK8 jets contain not only the hard decay products but also soft, wide-angle radiation from initial-state radiation, underlying event, and PU, which can obscure the properties of the original particle. The soft-drop [81] algorithm mitigates this issue by recursively declustering the jet and removing soft and wide-angle constituents. Soft-drop declusters a jet j into two subjets j_1 and j_2 using the Cambridge/Aachen algorithm [79]. It then applies the soft-drop condition

$$\frac{\min(p_{T1},\, p_{T2})}{p_{T1} + p_{T2}} > z_{\text{cut}} \left(\frac{\Delta R_{12}}{R_0}\right)^{\beta}, \tag{3.5}$$

where p_{T1} and p_{T2} are respectively the p_T of the subjets j_1 and j_2, ΔR_{12} is the angular separation between the two subjets, R_0 is the jet radius parameter, and z_{cut} and β are algorithm parameters. If the condition is satisfied, the algorithm stops and both subjets are retained. If the condition is not satisfied, the subjet with smaller p_T is removed and the procedure is iteratively repeated on the harder branch. This process continues until a splitting satisfies the condition or until no further declustering is possible. The parameters $z_{\text{cut}} = 0.1$ and $\beta = 0$ are used in applying the soft-drop algorithm in CMS. The soft-drop mass (m_{SD}) is calculated as the invariant mass of the two subjets, a cleaner observable that better reflects the mass of the originating particle.
Jet energy corrections (JECs) are applied to bring the measured jet response closer to the true parton-level momentum and to reduce discrepancies between data and simulation [82]. JECs consist of Jet Energy Scale (JES) and Jet Energy Resolution (JER) corrections. JES corrections are derived to

correct the difference between the reconstructed jet momentum in data and simulation. To determine the JES, CMS uses data-driven calibration strategies based on clean, momentum-balanced final states such as dijet, γ+jet, and Z+jet events. The JER describes the spread of the reconstructed jet p_T around its true value, typically Gaussian, and is generally better in simulation than in data; therefore, JEC corrections are applied to simulated jets to match the observed resolution in data.

3.9 Missing transverse energy

The MET [83] (E_T^{miss}) is defined as the module of the negative vectorial sum of the transverse momenta of all final-state particles reconstructed in the detector. It represents the transverse momentum of the particle, *i.e.* neutrinos, that escape from detectors without producing any direct response in the detector elements. CMS has developed some distinct algorithms to reconstruct it.

- The primary estimator is PF E_T^{miss}, calculated from the reconstructed PF particles.
- PUPPI-weighted PF MET (PUPPI MET) uses PUPPI as baseline PU-mitigated estimator: PUPPI assigns per-particle weights that suppress contributions from PU interactions, improving MET resolution and stability in high-PU running.
- DeepMET [84] is a deep-learning MET estimator that assigns a learned weight to each reconstructed particle and forms MET from a weighted vector summation of the transverse momenta; it improves MET resolution ($\sim$10–30%) and shows greater robustness to additional interactions compared with traditional estimators.

Non-linear response of calorimeters, minimum energy thresholds in the calorimeters, p_T thresholds and inefficiencies in the tracker contribute to misestimate MET. MET is corrected through a two-step procedure to mitigate the main sources of bias. While the first step suppresses PU contribution, the second one calibrates the detector response to restore the correct energy scale. A key component of the second step is the Type-1 MET correction, which propagates jet energy corrections to MET to account for the calibrated jet energies.

3.10 Heavy-flavor tagging

In CMS, many processes of interest, such as the production of top quarks, Higgs bosons, or searches for new physics, often produce heavy-flavor quarks: bottom (b) and charm (c) quarks. These quarks hadronize into jets that carry distinct signatures compared to light-quark (uds) or gluon (g) jets [85]. Identifying (or "tagging") these heavy-flavor jets is crucial for separating heavy-flavor jets from light-flavor or gluon jets enhancing the signal-to-background ratio, allowing for more precise measurements or more sensitive searches for rare processes.
b- and c-hadrons, formed when b and c quarks hadronize, have relatively long lifetimes, respectively of the order of $\sim$1.5 ps and $\sim$0.5 ps [86]. Because of this, they can travel (boosted by their momentum) a few millimeters before decaying, producing tracks that do not originate from the PV. These displaced tracks often make it possible to reconstruct a SV. Furthermore, the larger masses of b and c quarks and their harder fragmentation, compared to light quarks and gluons, cause the decay products of heavy hadrons to carry higher p_T than typical jet constituents. In addition, about 20% of b jets and 10% of c jets contain a soft muon or electron from semileptonic decays, providing an extra handle to distinguish heavy-flavor jets from light jets. Figure 3.6 highlights these characteristics, showing the reconstructed SV, the displaced tracks, the appreciable flight distance, and

the associated impact parameter (IP), the shortest distance with respect to the PV in the plane transverse to the beam axis, of the SV.

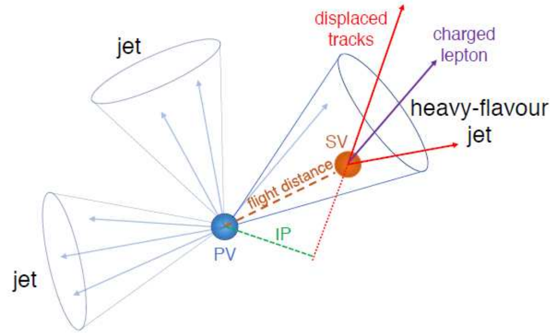

Figure 3.6: Representation of a heavy-flavor jet showing how tracks from a decaying b- or c-hadron appear displaced from the PV, since a long flight distance and a high impact parameter, and can be used to form a SV. The heavy-flavor jet also includes a charged lepton [85].

However, jet flavor tagging is intrinsically challenging: the information relevant to the jet origin is distributed across many particles, with complex correlations in kinematics, topology, and substructure. Tagging algorithms, built on cut-based approach, struggle to fully exploit this high-dimensional information. ML has become a natural and powerful solution to this problem. Modern ML models, including deep neural networks [87] and graph-based[2] [88] architectures, can learn directly from low-level detector measurements, such as track features, SVs, or PF candidates, and extract the subtle patterns that characterize different jet flavors. ML-based taggers improve both discrimination power and robustness compared to classical methods, and their flexibility allows them to incorporate additional information (*e.g.* jet substructure observables).
DeepJet [89] has been the CMS standard jet-flavor tagger for a few years. It is a convolutional–recurrent[3] neural tagger that uses as input jet global variables (*e.g.,* p_T and number of SVs) plus lists of charged and neutral PF candidates and SV features. The network processes independently the SVs within the jet radius, the charged hadrons and the neutral hadrons. The inputs pass then through 1×1 convolutional layers which transform the features of each individual object, extracting more informative representations from each jet constituent and creating new feature combinations from the original inputs. The outputs of the convolutional layers are then fed into recurrent layers which are concatenated and combined with global jet and event variables. The result is used to produce the final multiclass probabilities for jet classification. The network returns six output scores that can be interpreted as probabilities: P(b), P(bb), P(lepb), P(c), P(g), P(uds), corresponding to a single b-hadron, two b-hadrons, a leptonic b decay, charm, gluon, and light-flavor jets respectively.
More recent CMS taggers, such as ParticleNet (PNet) [90] and ParticleTransformer [91], exploit graph-neural-network and transformer-based[4] architectures to further improve jet flavor identification using low-level constituents. More details on the last taggers are reported in the next sections.

[2] A neural network designed to operate on data structured as a graph, where nodes represent entities and edges represent relationships between them. They iteratively update each node representation by aggregating information from its neighbors.

[3] Neural architectures that combine convolutional layers for feature extraction with recurrent layers that combine each input with information from previous steps.

[4] Neural network architecture designed to process sequences by allowing each element to directly consider the relevance of all other elements in the sequence.

3.10.1 ParticleNet

PNet is a dynamic graph[5] convolutional neural network designed to work directly on particle cloud, *e.g.,* an unordered and permutation-invariant set of SVs and jet constituent particles (each described by arbitrary per-particle features). Since points in a cloud are distributed irregularly (there is no fixed grid of neighbors) and any operation must be invariant under permutation of the points, an explicit definition of a local neighborhood for each particle is required, together with an aggregation that enforces permutation symmetry.

PNet is based on the EdgeConv [92] operator. The jet, seen as a cloud, is represented as a graph whose nodes are the SVs and the jet constituents and whose edges (the euclidean distance between the nodes) link each node to its k-nearest neighbors; the resulting neighborhood provides the local patch needed for enabling the node feature aggregation processing. For a point i with feature vector of dimension F ($x_i \in \mathbb{R}^F$) and neighbor indices $i_1, \ldots, i_k$, EdgeConv produces an updated feature of dimension F' ($x'_i \in \mathbb{R}^{F'}$)

$$x'_i = \square_{j=1}^{k} h_\Theta(x_i, x_{i_j}) \tag{3.6}$$

where h_Θ is an edge function with learnable parameters[6] Θ and $\square$ is a channel-wise symmetric aggregator (*e.g.* mean, sum, or max). Using the same Θ for all the edge functions and using a symmetric $\square$ makes EdgeConv permutation-invariant. In ParticleNet the edge function is implemented in the form

$$h_\Theta(x_i, x_{i_j}) = h'_\Theta(x_i, x_{i_j} - x_i) \tag{3.7}$$

so neighbor features enter through their relative offset to the central particle; h'_Θ, whose weights are the same for all the edge functions, is realized as an multilayer perceptron (MLP), a set of several layers of interconnected neurons that transform input data into an output prediction with at least one nonlinear activation function.

A key advantage of EdgeConv is its stackability: it maps a point cloud to another point cloud with the same number of vertices but updated per-point feature vectors, allowing multiple EdgeConv blocks to be used one after the other to build hierarchical feature representations. Furthermore, EdgeConv can be used with a dynamic graph update: the learned features produced by earlier EdgeConv layers can serve as new coordinates for computing nearest neighbors, allowing the network to focus on nodes that are most similar in the learned feature space. This dynamic graph strategy has been shown to improve performance over a static graph. More details on the PNet architecture are described in Section 5.1.

ParticleNet-MD (PNet-MD) is a variant of PNet employed in CMS analyses specifically designed for AK8 jets.

3.10.2 ParticleNet-MD

ParticleNet-MD is a mass-decorrelated variant of ParticleNet designed to tag AK8 jets avoiding the dependence on the AK8 jet mass while preserving classification power. A selection on a mass-dependent classifier can sculpt the background mass distribution and can produce fake bumps in the mass spectrum [93]. PNet-MD adopts the same EdgeConv / dynamic-graph backbone of PNet but changes the training strategy so that the classifier cannot exploit jet mass or p_T as a discriminating features: the model is trained on dedicated variable-mass signal samples ($X \to q\bar{q}$ spin-0 sample) and on QCD background that are reweighted to produce flat, identical mass and p_T distributions

[5] The graph is not fixed, but is recalculated at each layer of the network: the neighbours of each node are determined dynamically based on the updated features of the nodes themselves.

[6] Internal variables of a neural network that are adjusted through training to encode patterns of training inputs.

between signal and background, thereby reducing mass sculpting in background after selection. PNet-MD provides a set of probabilities per jet, one for each class it is trained to recognize:

- $P(Xbb)$: a resonance decaying into two b-quarks (*e.g.* $H \to b\bar{b}$);
- $P(Xcc)$: a resonance decaying into two c-quarks;
- $P(Xqq)$: a resonance decaying into two light quarks or into two gluons;
- P(QCD2b): a generic QCD jet with 2 or more b-hadrons;
- P(QCD2c): a generic QCD jet with 2 or more c-hadrons and no b-hadrons;
- P(QCDb): a generic QCD jet with 1 b-hadron;
- P(QCDc): a generic QCD jet with 1 c-hadron and no b-hadrons;
- P(QCDothers): a generic QCD jet with 0 heavy-flavour hadrons.

The raw probabilities are combine into discriminators tailored to a particular signal vs background separation. As an example, the PNet-MD bbvsQCD and PNet-MD ccvsQCD discriminators are defined as

$$\text{PNet-MD bbvsQCD} = \frac{P(Xbb)}{P(Xbb)+P(QCD2b)+P(QCD2c)+P(QCD1b)+P(QCD1c)+P(QCDothers)},$$

$$\text{PNet-MD ccvsQCD} = \frac{P(Xcc)}{P(Xcc)+P(QCD2b)+P(QCD2c)+P(QCD1b)+P(QCD1c)+P(QCDothers)}.$$

Figure 3.7 shows the background selection rate as a function of the efficiency of $X \to b\bar{b}$ ($c\bar{c}$) signal jets, in terms of the receiver operating characteristic (ROC) curves using simulated events in the 2018 data-taking conditions, comparing PNet-MD with the previous jet taggers. The ROC curve is obtained by varying the bbvsQCD (ccvsQCD) threshold applied to the output discriminant of the classifiers. By applying different threshold values, jets are classified as signal-like or background-like. For each choice of threshold, the corresponding signal efficiency and background efficiency are computed. Scanning the threshold over the full range of the score produces a set of points in the plane defined by these two quantities. The ROC curve is then constructed by plotting the background selection rate as a function of the signal efficiency. The double-b tagger [69] is a boosted decision tree algorithm trained to distinguish $H \to b\bar{b}$ jets from the QCD multijet background. The DeepAK8-MD algorithm [94] is a deep neural network (DNN)-based jet tagging algorithm developed for identifying resonance decays to $b\bar{b}$ or $c\bar{c}$. The DeepDoubleX tagging algorithm [95] is a DNN-based algorithm designed to identify $X \to b\bar{b}$ and $X \to c\bar{c}$ in the boosted topology. The PNet-MD model is better than the other taggers since it achieves a lower background selection rate at the same signal efficiency. In other words, for a given fraction of correctly identified signal jets, PNet-MD misidentifies fewer background jets as signal. For this reason PNet-MD has become the state-of-the-art tagger for AK8 jets, making it the leading AK8 jet classifier currently used in CMS analyses.
PNet-MD was re-trained for Run3 to deal with the new data-taking conditions and to accommodate additional signal topologies. Compared to the previous model, the most important differences are the following:

- **Training preselection**: a soft-drop mass preselection $m_{SD} > 25$ GeV was applied to the training sample, reducing sensitivity to very low-mass jets.

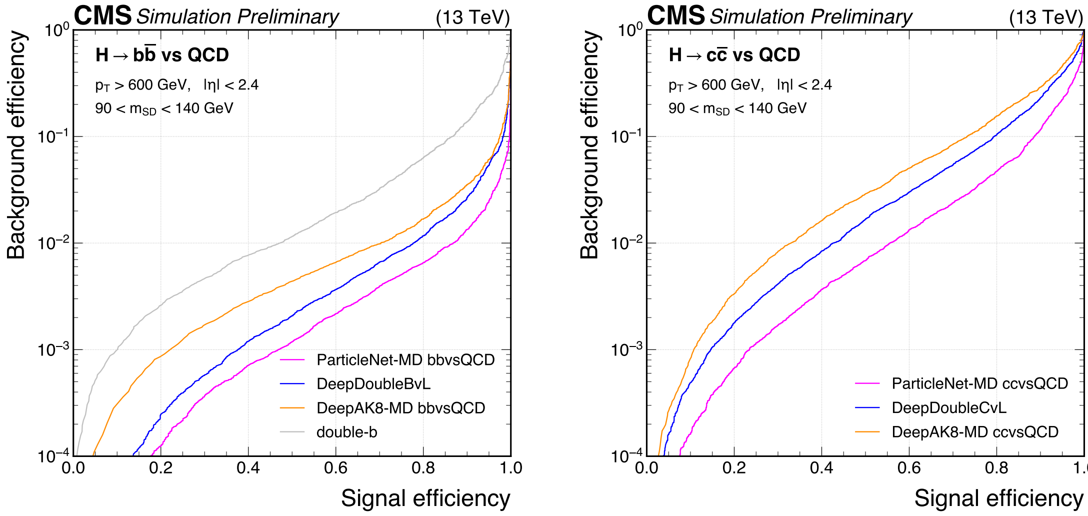

Figure 3.7: Comparison of the performance of the $X \to b\bar{b}$ (left) and $X \to c\bar{c}$ (right) tagging algorithms for $H \to b\bar{b}$ (left) and $H \to c\bar{c}$ (right) signal AK8 jets versus the inclusive QCD AK8 jets as background, using simulated events in the 2018 data-taking conditions. Performances are shown once applied the following AK8 selection $p_T > 600$ GeV, $|\eta| < 2.4$, $90 < m_{SD} < 140$ GeV [96].

- **Expanded signal composition**: the Run-3 training set included resonances decaying to $\tau\tau$-pairs (both fully hadronic and semileptonic configurations), so that the network sees and learns from hadronic and semileptonic $\tau\tau$ topologies.
- **Multi-task output**: the Run-3 network performs classification and a mass-regression simultaneously allowing correction to the AK8 jet mass, which improves the stability and the calibration of the predicted mass.
- **Improved heavy-flavour discrimination**: the output probabilities of the previous model P(QCDbb), P(QCDcc), P(QCDb), P(QCDc), were merged in the probability scores P(QCD2hf) and P(QCD1hf), which correspond respectively to the probability that the AK8 is a generic QCD jet with 2 or 1 heavy-flavour hadrons. At the same time, the training samples were tailored to enhance separation of QCD jets that contain two heavy-flavor particles, yielding better rejection of these background topologies.

It results in a PNet-MD model with the following output probabilities:

- $P(Xbb)$: a resonance decaying into two b-quarks;
- $P(Xcc)$: a resonance decaying into two c-quarks;
- $P(Xqq)$: a resonance decaying into two light quarks;
- $P(Xgg)$: a resonance decaying into two gluons;
- $P(X\tau\tau)$: a resonance decaying into two τ leptons, with τ both decaying hadronically;

- $P(X\tau\mu)$: a resonance decaying into two τ leptons, with one τ decaying hadronically and the other one leptonically into a muon ($\tau \to \mu\nu_\mu\nu_\tau$);
- $P(X\tau e)$: a resonance decaying into two τ leptons, with one τ decaying hadronically and the other one leptonically into an electron ($\tau \to e\nu_e\nu_\tau$);
- $P(\text{QCD2hf})$: a generic QCD jet with 2 or more b-hadrons or with 2 ore more c-hadrons and no b-hadrons;
- $P(\text{QCD1hf})$: a generic QCD jet with 1 b-hadron or with 1 c-hadron and no b-hadrons;
- $P(\text{QCD0hf})$: a generic QCD jet with 0 heavy-flavour hadrons.

With this new model, it is possible to separate resonances that decay into two b-quarks from QCD background using the assigned score:

$$\text{PNet-MD}_{\text{bbvsQCD}} = \frac{P(Xbb)}{P(Xbb) + P(QCD2hf) + P(QCD1hf) + P(QCD0hf)}.$$

It is essential to validate the new Run-3 ParticleNet-MD score before using it in physics analyses to ensure that the model response on real data is well modeled by simulation. I was responsible for validating PNet-MD on the initial Run-3 data, and the results are shown in the Chapter 4.

3.10.3 Boosted jet flavor tagging at HLT

PNet-MD was also the first mass-decorelated deep-learning boosted-jet tagger deployed in the CMS HLT menu at the start of Run-3 (2022), for AK8 bb and $\tau\tau$ tagging. However, given that many of the PNet-MD input variables either depend on expensive/late reconstruction steps (full muon reconstruction, full PF identification, detailed caloremeter clustering), or require track/vertex information that HLT either does not produce or produces with lower completeness/precision, the trigger version (or online version) of PNet-MD was trained on HLT samples with a restricted number of inputs.
Even if PNet-MD reduced the mass/p_T dependence, it didn't completely eliminate the dependency on p_T and mass. Improvements are still needed to reduce residual bias and to make taggers robust across wider phase space. I was responsible for training a new online version of PNet-MD as independent as possible from mass and p_T, and the results are shown in the Chapter 5, together with a description of the input features and the architecture of the online PNet-MD version.

3.10.4 ParticleTransformer

ParT is a deep neural network architecture based on the transformer paradigm that has been adapted for jet tagging in HEP. Instead of treating jets as ordered sequences or fixed grids, it takes as input sets of particle features and uses mechanisms that allow the model to weigh and combine information from different particles, capturing complex relationships among them. By incorporating physics-motivated pairwise interactions into its processing, ParT can learn subtle patterns in the jet structure that help it distinguish between different jet types, often outperforming previous models on jet classification benchmarks. Although ParT represents a more modern architecture, it was not considered in this thesis because, at the time this work began, it had been introduced only shortly before, and its performance had not yet been thoroughly studied.

Chapter 4

PNet-MD validation

Work presented in this chapter was performed to validate the Run-3 version of the PNet-MD model. In particular, its discrimination power was evaluated for the selection of $Z \to b\bar{b}$ + jets events in a boosted regime, requiring the presence of two AK8 jets with momentum higher than 450 and 200 GeV respectively. This phase-space is an optimal region for the test, having the Z boson production a lower statistics compared to the main background made of QCD multijet, the discrimination power of the algorithm can be therefore highlighted. The dominant QCD multijet background was estimated with a data-driven technique, extrapolating the contribution in a mass window around the Z-pole mass from the sideband regions. The signal model $Z \to b\bar{b}$, the other Z decay modes into quarks and the $W \to q\bar{q}'$ were instead taken from MC simulation. The analysis was performed considering 4 different PNet-MD$_{\mathrm{bbvsQCD}}$ score regions, defined to have the same amount of $Z \to b\bar{b}$ contribution, demonstrating how the score is able to enhance the $Z \to b\bar{b}$ contribution at high score probability, keeping a good agreement between data and MC. The CMS data collected during 2022 were considered for the validation. The work described in this chapter is part of the following Detector Performance note [97].

4.1 Data and simulated samples

4.1.1 Data

The data considered for the analysis belongs to the JetMET primary dataset, collecting events with high-p_T jets and/or significant MET, selected by jet- and MET-based HLT paths for jet and MET physics studies. For 2022, the dataset collected events corresponding to an integrated luminosity of 34.4 fb^{-1}. CMS calibrates its luminosity measurement using van der Meer scans of the colliding beams [98] and then integrates continuous rate measurements from dedicated detectors; the overall uncertainty on the 2022 integrated luminosity is 1.4% [99]. The data-taking was divided into separate eras, and Table 4.1 shows, for each era, the corresponding dataset.

4.1.2 SM processes

The SM expectation consists of the hadronic decays of the Z and W bosons and the QCD multijet contribution. The QCD multijet component is estimated with a data-driven technique, while the Z and W contributions are instead derived from MC simulation. Simulated pp collision events are generated at $\sqrt{s}$= 13.6 TeV. The V+jets (V = Z, W) processes are modeled at leading order (LO)

Data dataset name
/JetMET/Run2022C-19Dec2023-v1/MINIAOD
/JetMET/Run2022D-19Dec2023-v1/MINIAOD
/JetMET/Run2022E-19Dec2023-v1/MINIAOD
/JetMET/Run2022F-19Dec2023-v2/MINIAOD
/JetMET/Run2022G-19Dec2023-v1/MINIAOD

Table 4.1: Names of the JetMET datasets collected in 2022.

accuracy using the MadGraph5_aMC@NLO (MG) v2.6.5 generator [100]. Only the Z (W) boson hadronic decay into a quark-antiquark pair $q\bar{q}$ ($q = u, d, s, c, b$) at the matrix element (ME) level are considered. For all processes, the parton shower is simulated with PYTHIA v8.230 [101], using the CP5 underlying event tune [102] with the NNPDF3.1 next-next LO (NNLO) parton distribution function (PDF) set [103]. The matching of jets from ME calculations and those from parton showers is done with the MLM technique [104] for LO samples. Additional *pp* interactions (PU) are modeled by superimposing minimum bias collisions generated with PYTHIA. The events are then reweighted to reproduce the pileup distribution observed in data. The interactions between particles and the material of the CMS detectors are simulated using GEANT4 [105]. All samples are simulated slices of H_T, where H_T is the scalar p_T sum of all the jets in the event. The simulated datasets used are listed in Table 4.2 with the corresponding cross section times relevant branching ratio (BR).

Sample	Generator	Parton Shower	Cross section	PDF set	H_T slice [GeV]	$\sigma \cdot BR$ [pb]
Zto2Q-4jets	MG5_aMC@NLO (LO)	PYTHIA8 (CP5)	LO (MLM)	NNPDF3.1 NNLO	200–400	1084
Zto2Q-4jets	MG5_aMC@NLO (LO)	PYTHIA8 (CP5)	LO (MLM)	NNPDF3.1 NNLO	400–600	124.7
Zto2Q-4jets	MG5_aMC@NLO (LO)	PYTHIA8 (CP5)	LO (MLM)	NNPDF3.1 NNLO	600–800	27.98
Zto2Q-4jets	MG5_aMC@NLO (LO)	PYTHIA8 (CP5)	LO (MLM)	NNPDF3.1 NNLO	> 800	14.54
Wto2Q-3jets	MG5_aMC@NLO (LO)	PYTHIA8 (CP5)	LO (MLM)	NNPDF3.1 NNLO	200–400	2747
Wto2Q-3jets	MG5_aMC@NLO (LO)	PYTHIA8 (CP5)	LO (MLM)	NNPDF3.1 NNLO	400–600	301.5
Wto2Q-3jets	MG5_aMC@NLO (LO)	PYTHIA8 (CP5)	LO (MLM)	NNPDF3.1 NNLO	600–800	65.08
Wto2Q-3jets	MG5_aMC@NLO (LO)	PYTHIA8 (CP5)	LO (MLM)	NNPDF3.1 NNLO	> 800	32.27

Table 4.2: List of signal MC samples used in the validation with the corresponding generator, parton shower simulator, production cross section order, PDF set, H_T slice and production cross sections at $\sqrt{s} = 13.6$ TeV times relevant branching ratios.

4.2 Event selection

Only events satisfying at least one of the following HLT requirements are considered for the validation. The specific HLT triggers and the corresponding selection cuts are listed below:

- HLT_PFHT1050: H_T (evaluated using AK4 jets with $p_T > 30$ GeV and $|\eta| < 2.5$) > 1050 GeV;
- HLT_PFJet500: an AK4 jet with $p_T > 500$ GeV;
- HLT_AK8PFJet500: an AK8 jet with $p_T > 500$ GeV;
- HLT_AK8PFJet400_TrimMass30: an AK8 jet with $p_T > 400$ GeV and trimmed mass[1] > 30 GeV;

[1] AK8 jet mass removing the soft and wide angle radiation with the trimming algorithm [106].

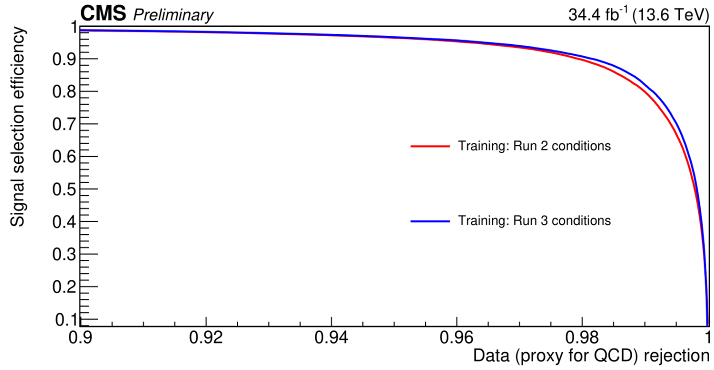

Figure 4.1: ROC curves obtained cutting on the Z boson candidate PNet-MD$_{\mathrm{bbvsQCD}}$ evaluated with the the PNet-MD model used at Run-2 (red curve) and the model trained for Run-3 (blue curve) [97].

- HLT_AK8PFHT800_TrimMass50: H_T (evaluated using AK8 jets with $p_T > 200$ GeV and $|\eta| < 2.5$) > 800 GeV and an AK8 jet with trimmed mass > 50 GeV.

The events were then selected applying the following request:

- leading-p_T AK8 jet: $p_T > 450$ GeV, $|\eta| < 2.4$ and $m_{SD} > 38$ GeV;
- subleading-p_T AK8: $p_T > 200$ GeV and $|\eta| < 2.4$;
- no electrons with $p_T > 20$ GeV, $|\eta| < 2.4$ satisfying the cut-based identification at the veto WP [74];
- no muons with $p_T > 20$ GeV, $|\eta| < 2.4$ satisfying the loose identification WP [76] and $I^{PF} < 0.4$ [48];
- no b-tagged AK4 jet (using the DeepJet medium working point[2]) with $p_T > 30$ GeV, $|\eta| < 2.4$ and a ΔR (with respect to the leading-p_T AK8 jet) < 0.8.

The leading-p_T AK8 corresponds to the Z boson candidate, while the subleading-p_T AK8 jet corresponds to the recoil-jet, usually produced in association with a boosted resonance. The vetos are used to suppress the $t\bar{t}$ background.
The Figure 4.1 shows the ROC curves obtained with PNet-MD$_{\mathrm{bbvsQCD}}$ comparing the Run-3 version with the one used during Run-2. The signal (MC $Z \to q\bar{q}$ +jets events where the Z boson candidate is compatible with $Z \to b\bar{b}$ at the generator level) efficiency and the data (used as QCD estimator) rejection have been estimated for events that pass the event selection and where the Z boson candidate m_{SD} is within the signal region. Thanks to the new training, a general improvement was observed in the tagging efficiency.

4.3 Analysis strategy

The region of interest (signal region) is defined around the Z boson mass peak $\approx 91.2\,$GeV, and m_{SD} is used to define this interval. Specifically, the signal region corresponds to $70 < m_{SD} < 126$

[2]Threshold on the P(b) + P(bb) + P(lepb) score in order to have a signal efficiency of 80%.

GeV, which fully contains the Z boson peak. The regions outside this interval, $38 < m_{SD} < 70$ GeV and $126 < m_{SD} < 198$ GeV, are treated as sideband regions. The m_{SD} distribution of the Z boson candidate for MC $Z \to q\bar{q}$ +jets events passing the selection is shown in Figure 4.2.

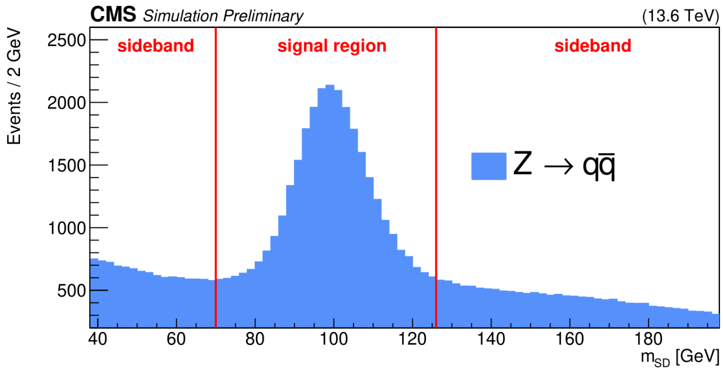

Figure 4.2: Z boson candidate m_{SD} distribution normalized to 34.4 fb^{-1} for the MC $Z \to q\bar{q}$ +jets events that pass the event selection. The red vertical bars separate the signal region from the sideband regions [97].

Events have been categorized into five PNet-MD$_{\mathrm{bbvsQCD}}$ regions. Each region has been chosen to contain the same amount (20%) of MC $Z \to b\bar{b}$ events. In particular, the $Z \to b\bar{b}$ events were selected if they fullfil the selection previously described and if the AK8 acting as Z boson candidate has a distance ΔR < 0.8 with respect to the two b-quarks, identified at generator level as produced by the Z boson decay. The lowest score region (PNet-MD$_{\mathrm{bbvsQCD}} < 0.641$) was not considered for the validation, since the data are dominated by QCD multijet events. The remaining regions, shown in Figure 4.3, are reported below:

- 4$^{\mathrm{th}}$ highest score region: $0.641 <$ PNet-MD$_{\mathrm{bbvsQCD}} \leq 0.875$
- 3$^{\mathrm{rd}}$ highest score region: $0.875 <$ PNet-MD$_{\mathrm{bbvsQCD}} \leq 0.957$
- 2$^{\mathrm{nd}}$ highest score region: $0.957 <$ PNet-MD$_{\mathrm{bbvsQCD}} \leq 0.988$
- highest score region: $0.988 <$ PNet-MD$_{\mathrm{bbvsQCD}} \leq 1$

4.4 QCD background estimation

The QCD multijet background is estimated, independently in each PNet-MD$_{\mathrm{bbvsQCD}}$ score region, extrapolating the QCD contribution in the signal region from a fit in the sideband regions of the m_{SD} distribution. The sideband regions were modeled using Chebyshev polynomials up to the 6$^{\mathrm{th}}$ order[3]. The binning scheme was optimized to ensure sufficient statistical precision while mitigating spurious fluctuations in the distributions and to maintain a number of bins exceeding the number of free fit parameters. In particular, the 4$^{\mathrm{th}}$ highest score region employs an m_{SD} bin width of 4 GeV, whereas all remaining score categories use a bin width of 8 GeV. As Chebyshev polynomials

[3] A 6$^{\mathrm{th}}$ order Chebyshev polynomial consists of seven terms, given by: $c_0 + c_1 \cdot x + c_2 \cdot (2x^2 - 1) + c_3 \cdot (4x^3 - 3x) + c_4 \cdot (8x^4 - 8x^2 + 1) + c_5 \cdot (16x^5 - 20x^3 + 5x) + c_6 \cdot (32x^6 - 48x^4 + 18x^2 - 1)$

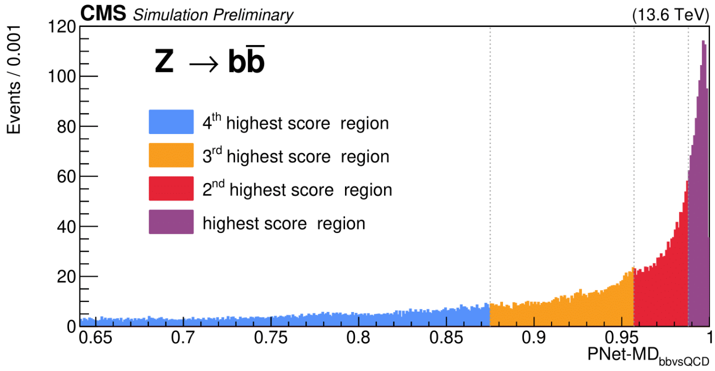

Figure 4.3: Z boson candidate PNet-MD$_{\text{bbvsQCD}}$ distribution, starting from 0.641, normalized to 34.4 fb^{-1} for the MC $Z \to q\bar{q}$ +jets events passing the event selection and compatible with $Z \to b\bar{b}$ at generator level. The different score regions are delimited by the dashed vertical bars and are indicated with different colors [97].

are defined on the interval [-1;1], the m_{SD} distributions were rescaled by 200 GeV ($x = \frac{m_{SD}}{200\ \text{GeV}}$) in order to map the observable onto the polynomial domain.
The optimal polynomial order was determined via an iterative procedure. Starting from the lowest-order Chebyshev polynomial, the fit was repeated with progressively higher orders until no statistically significant improvement in the description of the data was observed at the 5% significance level. The statistical preference between two polynomial models with number of parameters a and $b > a$ fitting N points was estimated using the Fisher F-test [107]. The test statistic

$$F = \frac{(\chi^2_a - \chi^2_b)/(b-a)}{\chi^2_b/(N-b)}$$

quantifies whether the observed decrease in χ^2 reflects a genuine improvement of the model description, or simply the increased flexibility associated with a higher polynomial order. Under the null hypothesis that the lower-order polynomial is sufficient, we compute the F-statistic with $(b-a)$ and $(N-b)$ degrees of freedom and define the p-value $p = P(F_{(b-a,N-b)} > F_{obs})$, *i.e.* the upper-tail probability from the F-distribution. This p-value represents the probability of observing a reduction in χ^2 at least as large as that measured if the additional parameters are not justified. The higher-order polynomial is selected only if $p < 0.05$; otherwise the lower-order model is retained. Figure 4.4 shows all the polynomials fitting the $m_{SD} \in [38, 70]\ \cup [126, 198]$GeV sideband in the four score region. The selected polynomial orders provide a good description of the sideband distributions in all regions while avoiding unnecessary model complexity. In fact, relative to these orders, higher-order polynomials yield nearly identical fits in the sideband.
To further improve the QCD background estimate, eight alternative definitions of the signal regions were considered, by enlarging or reducing the nominal signal region boundaries. In each alternative sideband, the Chebyshev polynomial order was chosen following the same iterative Fisher F-test procedure used in the nominal sideband region. The additional 8 regions considered are the following:

- $m_{SD} \in [38, 62]\ \cup [134, 198]$GeV
- $m_{SD} \in [38, 78]\ \cup [118, 198]$GeV

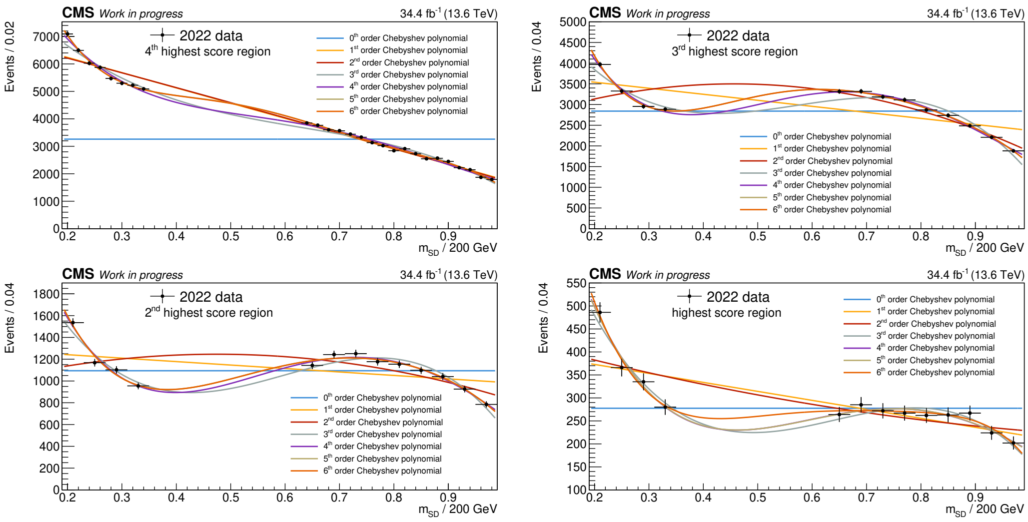

Figure 4.4: 2022 data for $m_{SD} \in [38, 70] \ \cup [126, 198]$GeV in the four score regions fitted by Chebyshev polynomials ranging from order 0 to order 6. The order of the polynomial that most closely approximates the data trend was chosen based on the Fisher F-test results at a 5% significance level. Based on the Fisher F-test results at a 5% significance level, the QCD in the 4$^{\text{th}}$ highest score region (top-left) and in 3$^{\text{rd}}$ highest score region (top-right) was modeled using a 5$^{\text{th}}$ order polynomial, the QCD in the 2$^{\text{nd}}$ highest score region (bottom-left) was modeled using a 4$^{\text{th}}$ order polynomial and the QCD in the highest score region (bottom-right) was modeled using a 3$^{\text{rd}}$ order polynomial.

- $m_{SD} \in [38, 54] \ \cup [142, 198]$GeV
- $m_{SD} \in [38, 86] \ \cup [110, 198]$GeV
- $m_{SD} \in [38, 62] \ \cup [126, 198]$GeV
- $m_{SD} \in [38, 78] \ \cup [126, 198]$GeV
- $m_{SD} \in [38, 70] \ \cup [134, 198]$GeV
- $m_{SD} \in [38, 70] \ \cup [118, 198]$GeV

In appendix A, all the fits are shown. For each score region, the nine analytical functions were then discretized within the nominal signal region using the same bin widths adopted for the corresponding sideband fits. Therefore, for a given fit f_i the predicted QCD yield in the bin b ($Y_{i,b}$) was obtained by integrating f_i over the b range. The final QCD estimate was computed as the bin per bin arithmetic mean of the nine discretized predictions $\overline{Y_{i,b}} = \frac{1}{9}\sum_{i=1}^{9} Y_{i,b}$. Figure 4.5 shows the data-driven estimate of the QCD in the four score regions, together with the nine analytic functions obtained from the fit in the m_{SD} sidebands.

4.5 Systematic uncertainties

The overall uncertainty on a measurement has both statistical and systematic contributions. The statistical component mainly depends on the raw amount of data available and gets smaller as more

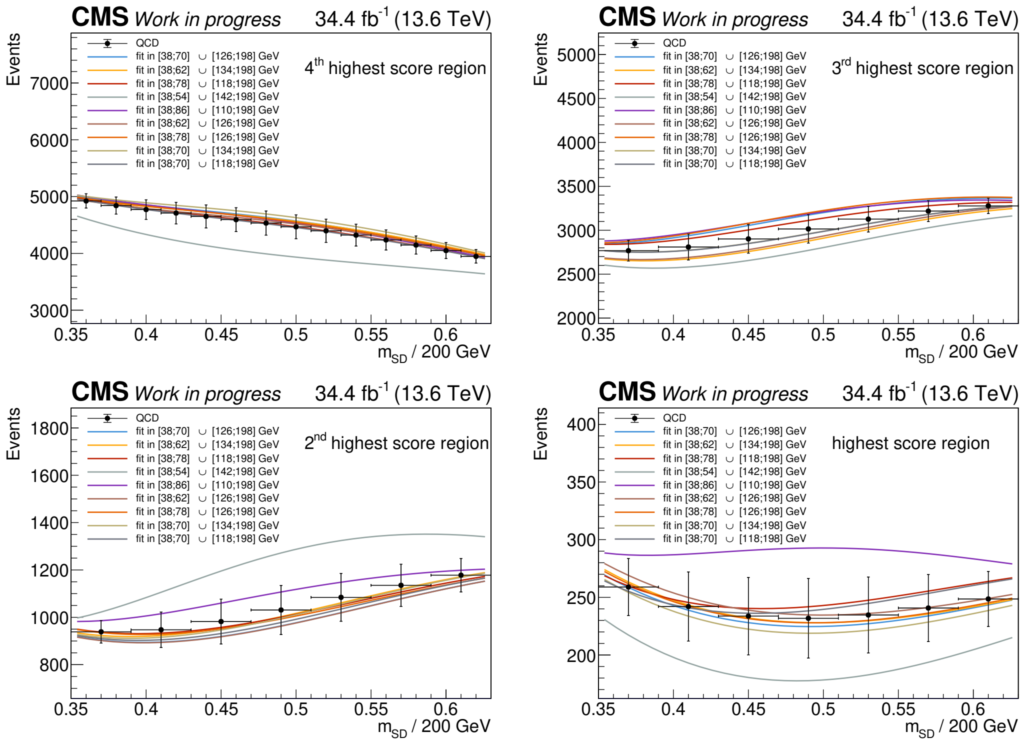

Figure 4.5: The fitted functions, obtained from the nine m_{SD} sidebands, and the final data-driven QCD estimate in the four score region.

data are collected. On the other hand, systematic uncertainties are due to imperfect knowledge of the measurement process, the experimental setup, or the theoretical modeling. The sources of the systematic uncertainties, which are applied exclusively to the $Z \to q\bar{q}$ and the $W \to q\bar{q}'$ samples, are reported below:

- **lumi**: the uncertainty on the integrated luminosity affects all the simulated samples and is 1.4%;
- **JES corrections**: The jet energy scale (JES) systematic uncertainty is decomposed into 27 independent sources [108], grouped into categories related to absolute (p_T-dependent) and relative (η-dependent) scale corrections (including statistical components), pileup-offset corrections, flavour- and sample-dependent response, fragmentation and single-pion calorimeter response, resolution-related effects, and residual variations of the jet energy response as a function of data-taking period, p_T, and η. For each source, a $\pm 1\sigma$ variation was propagated to the JES corrections of the two AK8 soft-drop subjets, and m_{SD} was recomputed as the invariant mass of the corresponding shifted subjets.

For QCD, the following two systematic uncertainties were assigned:

- **fit**: symmetric error in each x bin equal to the average of the nine fit function errors $\sigma_{i,b}$ associated to each bin;

- **window**: symmetric error in each x bin equal to the root mean square deviation of the $Y_{i,b}$ values obtained from the different fit functions.

All systematic uncertainties, with the exception of the luminosity, are treated as uncorrelated across the different score regions, and are therefore estimated independently in each region. The luminosity uncertainty is instead taken to be fully correlated among all score regions. Tables 4.3 reports the relative uncertainty with respect to the total expected SM yield.

	Score region			
	4^{th} highest	3^{rd} highest	2^{nd} highest	highest
Total SM expectation	65129	22503	8326	2546
Uncertainty				
window	3.7%	4.0%	6.6%	6.8%
fit	0.8%	1.3%	2.2%	4.5%
Statistical	0.24%	0.32%	0.6%	1.1%
lumi	0.05%	0.09%	0.17%	0.5%
RelativeSample	0.008%	0.011%	0.024%	0.04%
RelativeStatEC	0.002%	0.003%	0.018%	0.015%

Table 4.3: Main SM relative uncertainties for the different score regions. RelativeSample and RelativeStatEC are two of the 27 JES uncertainty sources, which respectively account for differences in the residual data-to-simulation corrections obtained with different MC samples and for the statistical uncertainty in the η-dependent calibration ($1.3 < |\eta| < 3$).

4.6 Results

To obtain a more accurate estimate of the processes used in the validation of the PNet-MD$_{\text{bbvsQCD}}$ score, a likelihood fit was performed on the m_{SD} distributions of all four score regions by mean of the Combine tool [109]. Two normalization factors were introduced for W+jets (r_W) and Z+jets (r_Z) samples. For Z+jets, although the sample was split by final state ($Z \to b\bar{b}$, $Z \to c\bar{c}$ and $Z \to u\bar{u}/d\bar{d}/s\bar{s}$), a single normalization factor was used across the different decay channels, as only the overall normalization was corrected and not the individual BRs. Since each score region probes a different phase space, separate normalization factors were introduced for W+jets and Z+jets in every score bin: the factors are defined as r_{Zc} and r_{Wc} for each score region, where $c = 1$ corresponds to the highest score region and $c = 4$ to the fourth-highest score region.
The expected yield in bin b of the score region c is given by the sum of the contributions of all samples s:

$$\mu_{b,c}(r_{Zc}, r_{Wc}, \theta) \;=\; r_{Zc} \sum_{s\in \text{Z+jets}} s_{s,b,c}(\theta) \;+\; r_{Wc}\, s_{\text{Wjets}b,c}(\theta) \;+\; s_{\text{QCD}b,c}(\theta). \tag{4.1}$$

where $s_{\text{QCD}b,c}$, $s_{\text{Wjets}b,c}$ and $s_{s,b,c}$ represent the expected yields of QCD, W+jets, and Z+jets processes in bin b and score region c, respectively, with the Z+jets contribution further categorized into bb, cc, and light-flavor jet components, and θ is a set of nuisance parameters parameterizing the uncertainties of the specific sample. θ parameters are introduced to take into account the effect of the uncertainties in the likelihood maximization process.
Given the observed number of events $n_{b,c}$ in bin b of score region c the full likelihood function can

be defined as

$$\mathcal{L}(\{r\},\theta) = \prod_c \prod_b \mathrm{Pois}(n_{b,c} \,|\, \mu_{b,c}(r_{Zc}, r_{Wc}, \theta)) \;\times\; \prod_k \pi_k(\theta_k), \tag{4.2}$$

where $\{r\}$ includes all 8 normalization factors r_{Zc} and r_{Wc}, $\mathrm{Pois}(n \mid \mu)$ denotes the Poisson probability of observing n events given an expectation value μ, and $\pi_k(\theta_k)$ (constraint term) corresponds to probability functions that model the nuisance parameter θ_k. The choice of constraint terms and the treatment of nuisance parameters in the likelihood are detailed in Appendix B.
The best-fit estimators $\hat{r}_{Zc}$ and $\hat{r}_{Wc}$ are obtained by maximizing the likelihood function with respect to all free parameters. For numerical stability, the fit is performed by minimizing $-2\ln\mathcal{L}$ instead of maximizing $\mathcal{L}$.
To quantify the uncertainty on the estimators $\hat{r}_{Zc}$ and $\hat{r}_{Wc}$, confidence intervals were constructed using the *profile likelihood*. For a given value of a parameter of interest (POI) r, which in this case are the normalization factors $\{r\}$, the profile likelihood is obtained by minimizing $-2\ln\mathcal{L}$ with respect to all nuisance parameters θ and the other free POIs r':

$$\mathcal{L}_{\mathrm{prof}}(r) \;=\; \mathcal{L}\big(r, \hat{\theta}(r), \hat{r}'_{\neq r}(r)\big), \qquad \text{where} \qquad \hat{\theta}(r), \hat{r}'_{\neq r}(r) = \arg\min_{\theta, r'_{\neq r}} [-2\ln\mathcal{L}(r,\theta,r'_{\neq r})].$$

The negative log-likelihood difference,

$$-2\Delta\ln\mathcal{L}(r) \;=\; -2\ln\frac{\mathcal{L}_{\mathrm{prof}}(r)}{\mathcal{L}_{\mathrm{prof}}(\hat{r})}, \tag{4.3}$$

measures how much less compatible the value r is with the observed data compared to the overall best-fit $\hat{r}$. Under the asymptotic approximation provided by Wilks' theorem [110], $-2\Delta\ln\mathcal{L}(r)$ follows a χ^2 distribution with one degree of freedom. Confidence intervals are then defined as the set of r values satisfying

$$-2\Delta\ln\mathcal{L}(r) < 1 \quad \text{for 68\% confidence level (CL)}, \qquad -2\Delta\ln\mathcal{L}(r) < 4 \quad \text{for 95.45\% CL}.$$

The maximum-likelihood fit simultaneously returns normalization factors and profils the nuisance parameters that encode statistical and systematic uncertainties. Figure 4.6 shows $-2\Delta\ln\mathcal{L}(r)$ as a function of the four r_Z normalization factors. All the r_Z values are compatible with unity within the uncertainties, this implies that the classifier score is sufficiently well modeled to describe the data, with only small residual adjustments necessary.
The estimated values of $\hat{r}_W$ are reported in Table 4.4, together with their total and statistical uncertainties. In the 4^{th} highest score region, $\hat{r}_W$ is not compatible with unity, while it is consistent with unity within uncertainties in the 3^{rd} highest score region. This indicates that a more accurate modeling of the W+jets contribution is required in the former region. In the 2^{nd} highest score region and in the highest score region, $\hat{r}_W$ is compatible with both zero and unity within uncertainties, reflecting the limited sensitivity to the W+jets process. This is due to the fact that, in these regions, the W+jets contribution is subdominant with respect to QCD and Z+jets production.

4.6.1 Impacts and pulls of nuisance parameters

The influence of nuisance parameters on the normalization factors was quantified by computing the "impacts". To the quality and consistency of the fit was assessed by measuring the "pulls".
The *impact* of a nuisance parameter θ is defined as the shift Δr in the POI r obtained by fixing θ at its best-fit value $\hat{\theta}$ and then at its $\pm 1\sigma$ best-fit values. For each fixed value of θ, all other parameters are profiled to minimize the profiled likelihood. In the Gaussian limit the resulting Δr is a measure

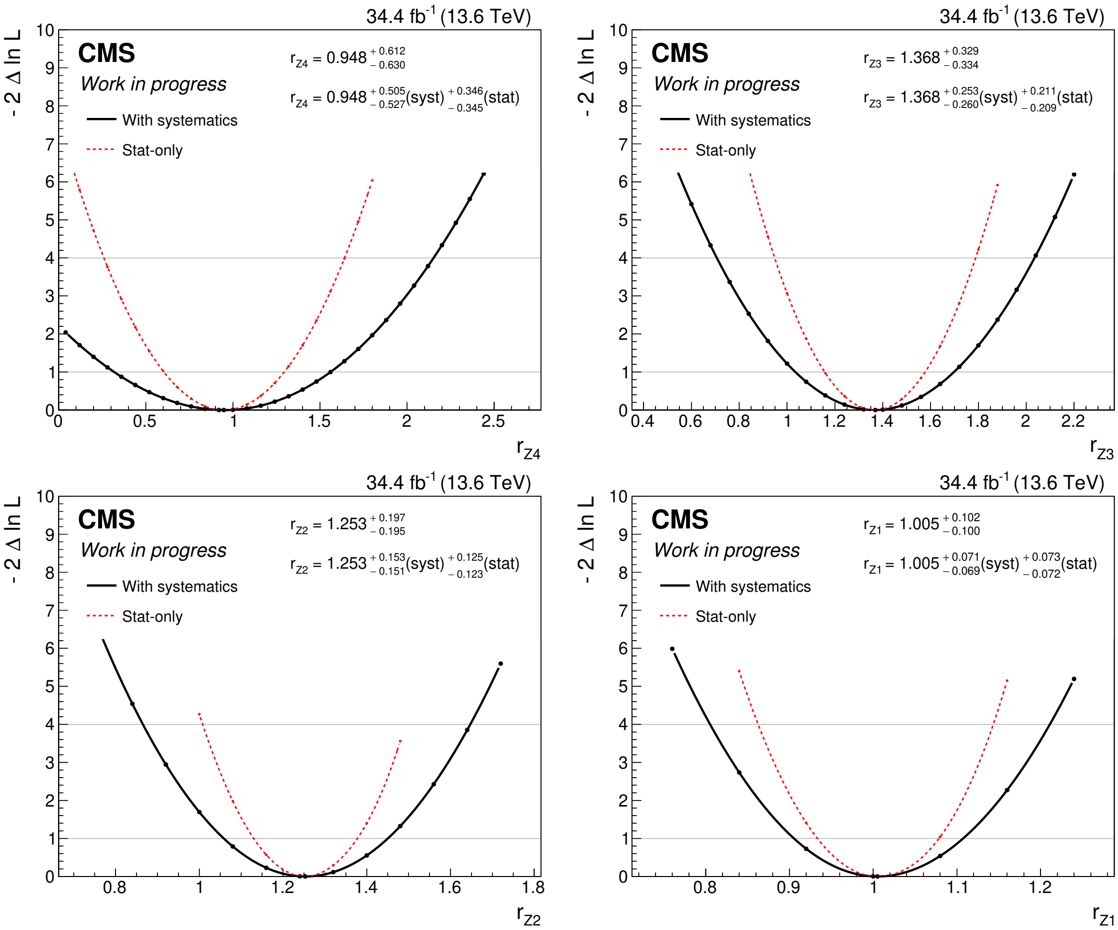

Figure 4.6: $-2\Delta\ln\mathcal{L}$ as a function of r_Z of the 4$^{\text{th}}$ highest score region (top-left), of the 3$^{\text{rd}}$ highest score region (top-right), of the 2$^{\text{nd}}$ highest score region (bottom-left) and of the highest score region (bottom-right). Each figure shows the value that minimizes $-2\Delta\ln\mathcal{L}$. The red dashed curves are $-2\Delta\ln\mathcal{L}$ when only statistical uncertainties are considered, while the black solid curves include also the systematic uncertainties. The written uncertainties correspond to the 68% CL.

	$\hat{r}_W$	Total uncertainty	Statistical uncertainty
4$^{\text{th}}$ highest score region	1.74	-0.30/+0.31	-0.25/+0.25
3$^{\text{rd}}$ highest score region	1.09	-0.51/+0.55	-0.36/+0.36
2$^{\text{nd}}$ highest score region	0.54	-0.54/+0.70	-0.54/+0.59
highest score region	1.3	-1.3/+2.2	-1.3/+1.7

Table 4.4: Best-fit values of the W+jets normalization factor $\hat{r}_W$ in the four score regions, together with their total and statistical uncertainties. Uncertainties are quoted at 68% confidence level.

of the correlation between the POI r and the nuisance parameter θ. The impacts are therefore useful to identify which nuisance parameters have the largest effect on the uncertainty of POI.
The *pull* of a nuisance parameter is defined as the deviation of the post-fit estimate $\hat{\theta}$ (obtained from the $-2\Delta \ln \mathcal{L}$ minimization on data) from its pre-fit (initial) value θ_I, normalized by the quadrature sum of the pre-fit uncertainty σ_I and the post-fit uncertainty σ:

$$\text{pull}(\theta) \;=\; \frac{\hat{\theta} - \theta_I}{\sqrt{\sigma_I^2 + \sigma^2}}\,.$$

Figures 4.7, 4.8, 4.9, 4.10 show the impacts and pulls of the leading 30 systematic uncertainties on the Z normalization factors. The suffix _s*c* appended to uncertainty names, such as window_s0, denotes the score region in which the uncertainty is evaluated, here $c = 0$ corresponds to the 4$^{\text{th}}$ highest score region, $c = 1$ to the 3$^{\text{rd}}$ highest, $c = 2$ to the 2$^{\text{nd}}$ highest and $c = 3$ to the highest. prop_binScore_*c*_bin*b* is the MC statistical uncertainty in the score region c and distribution bin b. The fact that nuisance parameters are pulled but remain within the [-1,1] range indicates that the fit adjusts them only within their nominal 1σ uncertainties. This behavior shows that the data are generally consistent with the expected templates, and that no nuisance needs to be shifted beyond its pre-defined uncertainty to accommodate the observations. It is important to note that even small pulls do not necessarily imply a negligible impact on the parameter of interest; the actual influence depends on the correlation between each nuisance and the POI, which is reflected in the corresponding impact plots. Among all systematic uncertainties, the one related to *window* exhibits both the largest pull and the largest impact on the fit result in all score regions.

4.6.2 Post-fit plots

The resulting normalization factors and the nuisance shifts were applied on the m_{SD} distributions to produce the post-fit histograms. Figure 4.11 compares the data m_{SD} distribution with the the post-fit predicted ones in the four score regions. The agreement between data distribution and the expectation (within uncertainties) supports the reliability of the QCD data-driven estimate. In addition, Figure 4.11 shows that the m_{SD} peak associated with $Z \to b\bar{b}$ becomes progressively more pronounced when moving from the lowest to the highest score region, indicating that the PNet-MD$_{\text{bbvsQCD}}$ discriminator provides robust separation between signal and background.
Figure 4.12 shows the PNet-MD$_{\text{bbvsQCD}}$ data distribution after QCD subtraction compared to the stacked post-fit $Z \to q\bar{q}$ and $W \to q\bar{q}'$ templates. The amount of $W \to q\bar{q}'$ and mis-identified $Z \to q\bar{q}$ ($q = u, d, s, c$) events reduces as the PNet-MD$_{\text{bbvsQCD}}$ score increases, as expected. As evidence of this, Table 4.5 shows the percentage on the total number of post-fit $Z \to q\bar{q}$ events depending on the quark flavors. In particular, in the highest score region the $Z \to b\bar{b}$ percentage among all the $Z \to q\bar{q}$ events is 96.6%, demonstrating how the PNet-MD is able to identify the $Z \to b\bar{b}$ contribution.

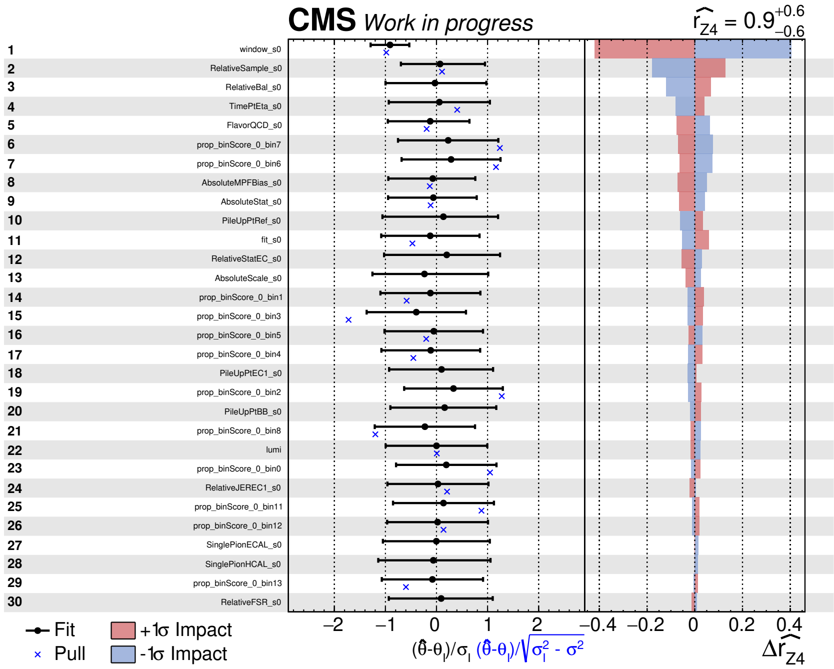

Figure 4.7: List of the 30 nuisances parameters with the highest impact on the measurement of r_{Z4}. The left row shows the pull of the nuisance parameters, while the right one shows the impact of the nuisances. The most important source of systematic uncertainty that affects the measurement is the QCD systematics "window" estimated in the 4$^{\text{th}}$ highest score region. Since the "window_s0" nuisance has a negative pull, the QCD yield in the 4$^{\text{th}}$ highest score region was overestimated in the pre-fit model.

In light of the results presented and the observed performance of the method, it appears feasible that, with improved modeling of the QCD contribution and a more accurate determination of the W+jets normalization, this approach could be employed to extract scale factors for any bbvsQCD score.

	$Z \to b\bar{b}$	$Z \to c\bar{c}$	$Z \to u\bar{u}/d\bar{d}/s\bar{s}$
4$^{\text{th}}$ highest score region	48.8%	35.5%	15.7%
3$^{\text{rd}}$ highest score region	73.9%	20.5%	5.6%
2$^{\text{nd}}$ highest score region	87.7%	10.9%	1.4%
highest score region	96.6%	3.18%	0.26%

Table 4.5: Composition of the Z-candidate post-fit sample in the four score regions.

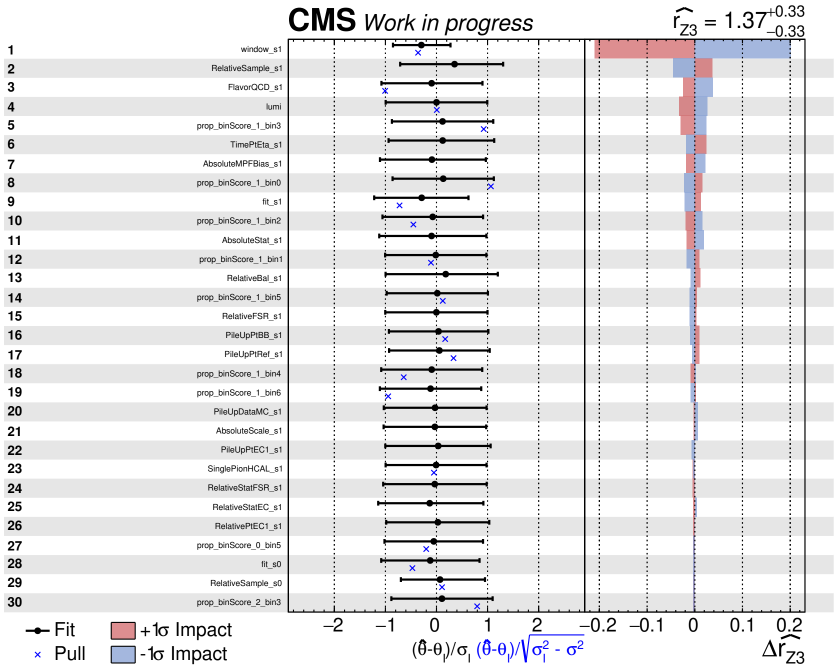

Figure 4.8: List of the 30 nuisances parameters with the highest impact on the measurement of r_{Z3}. The left row shows the pull of the nuisance parameters, while the right one shows the impact of the nuisances. The most important source of systematic uncertainty that affects the measurement is the QCD systematics "window" estimated in the 3rd highest score region.

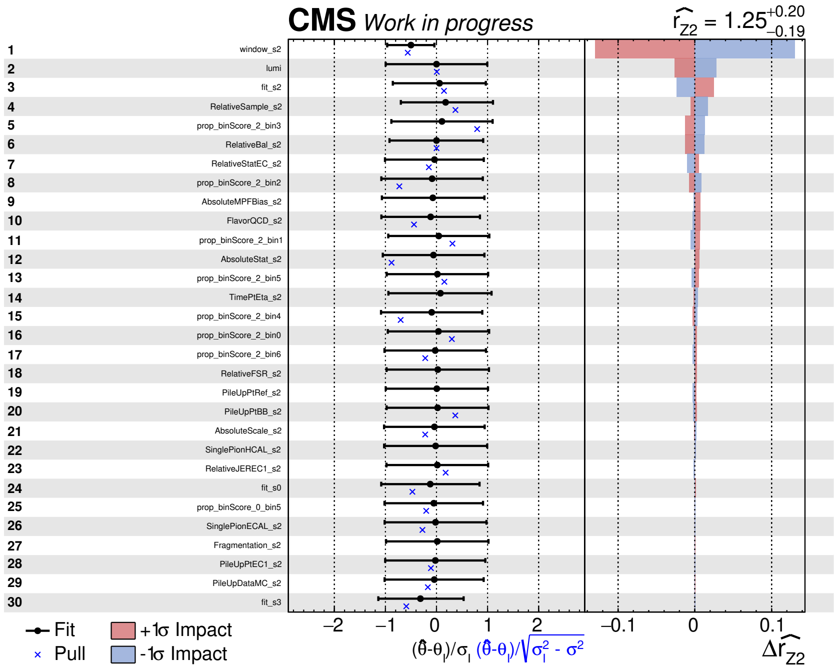

Figure 4.9: List of the 30 nuisances parameters with the highest impact on the measurement of r_{Z2}. The left row shows the pull of the nuisance parameters, while the right one shows the impact of the nuisances. The most important source of systematic uncertainty that affects the measurement is the QCD systematics "window" estimated in the 2$^{\text{nd}}$ highest score region.

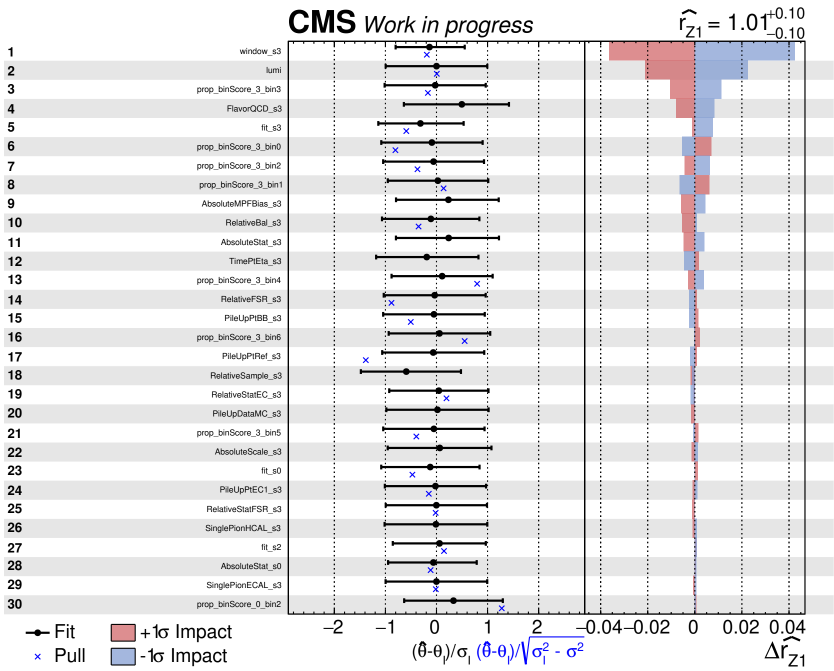

Figure 4.10: List of the 30 nuisances parameters with the highest impact on the measurement of r_{Z1}. The left row shows the pull of the nuisance parameters, while the right one shows the impact of the nuisances. The most important source of systematic uncertainty that affects the measurement is the QCD systematics “window” estimated in the highest score region.

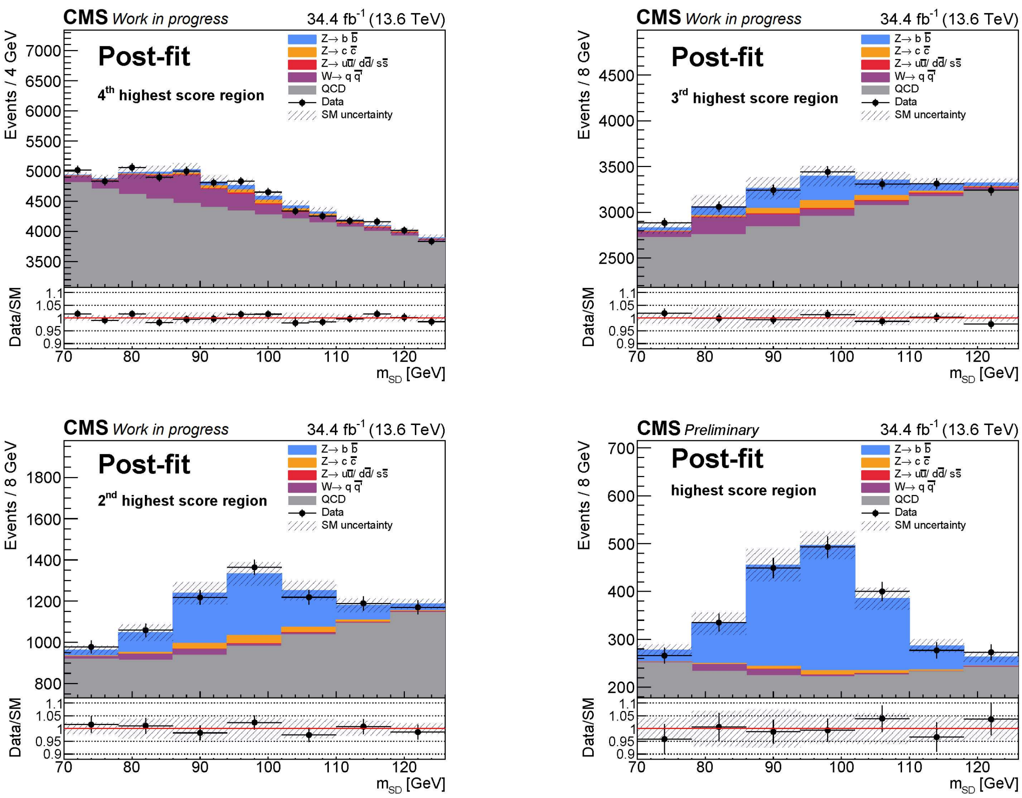

Figure 4.11: Post-fit data (black markers) to prediction (stack of QCD, $W \to q\bar{q}'$ and $Z \to q\bar{q}$ with the latter split on the basis of its decay products) comparison in the four score regions. The upper plot shows the m_{SD} distributions while the lower plot shows the ratio between data and prediction.

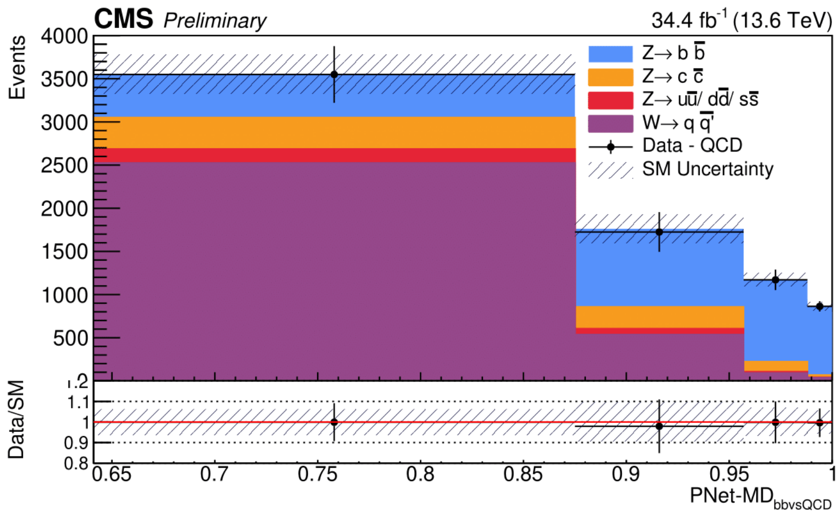

Figure 4.12: Post-fit data, after subtracting the QCD contribution, (black markers) compared in the highest score region with stack of $W \to q\bar{q}'$ and $Z \to q\bar{q}$ with the latter split on the basis of its decay products. The upper plot shows the PNet-MD$_{\rm bbvsQCD}$ distributions while the lower plot shows the ratio [97].

Chapter 5

PNet-MD training and implementation at HLT

PNet-MD was the first mass-decorrelated, deep-learning boosted-jet tagger deployed in the CMS HLT menu [111] (the list of all HLT paths). It is used for tagging resonances (X) decaying into $b\bar{b}$ and $\tau\bar{\tau}$, which are reconstructed as AK8 jet. The model uses online inputs supplied by the HLT reconstruction chain rather than the full offline object collection: several of the features that PNet-MD relies on offline are either not available in the HLT, or are available only with reduced completeness and resolution. To accommodate this, the HLT version of PNet-MD was trained with the HLT reconstruction and on a restricted set of inputs. Although PNet-MD has small dependence from p_T and mass, the HLT model shows a sculpting effect at low p_T and m_{SD}. The observed sculpting at low p_T and m_{SD} suggests that the model has not learned a fully decorrelated representation with respect to these variables. This residual correlation causes the score threshold to act as an implicit kinematic selection: jets in the low-p_T and low-m_{SD} region tend to receive systematically lower scores and are therefore more likely to be rejected. As a result, many events containing such jets are removed, reducing the efficiency in that phase space and introducing a distortion of the underlying kinematic distributions. The goal of the activity consisted in the training a new PNet-MD model at HLT which reduces the dependence on p_T and m_{SD}, without compromising performance in terms of signal identification, in order to add the model in the 2026 CMS pp collision HLT menu. To ensure a model with a uniform response, the model was trained using a graviton-like resonance (X) decaying into two Higgs bosons, with both masses of X and H variable, as signal and QCD as background. This chapter presents the selection of the training samples, the training workflow within the adopted model development framework, the evaluation of the model performance, and dedicated studies to ensure reduced dependence on p_T and m_{SD}, while satisfying the HLT latency and integration requirements.

5.1 PNet-MD model

The new PNet-MD model receives two distinct types of inputs:

- **PF candidates**: PF objects, with $p_T > 0.1$ GeV, that constitute the AK8 jet. From those, only the first 120 ranked by their p_T were selected.
- **SVs**: SVs reconstructed with the IVF algorithm (see Section 3.2) and within a distance $\Delta R <$

0.8 with respect to the AK8 jet axis. Among these, only the first 10 SVs, ranked by their transverse impact parameter significance with respect to the primary vertex—defined as the ratio of the transverse displacement Δxy to its uncertainty $\sigma_{\Delta xy}$ —were retained.

If the jet contains fewer than 120 PF objects or 10 SVs, the remaining points are filled with dummy (padding) points, with all features set to zero. Padding point are later masked during training. In the training phase, it is usually not convenient to use all the available jets at once. Instead, the dataset is divided into smaller subsets called batches. Each batch contains a limited number of jets that are loaded into memory and processed together in a single training step. The parameters of the model are then updated and this procedure is repeated batch by batch until the entire dataset has been processed. The PF candidate and SV features are initially processed in two separate streams. In each stream, the features of each point are mapped into a 32-dimensional space, producing a learned embedding that represents each point in a way suitable for subsequent processing. The projection is implemented using the following steps:

- BatchNorm1D($\mathrm{C_{in}}$),
- Conv1D($\mathrm{C_{in}} \to \mathrm{C_{out}}$),
- BatchNorm1D($\mathrm{C_{out}}$),
- ReLU

where $\mathrm{C_{in}}$ is the number of features (channels) associated to PF candidates or SVs and $\mathrm{C_{out}}$ is the number of output features. These operations allow to standardize the dimensionality of features and reduce dependence on the original scales of physical variables. The number of output channels was set to 32.

BatchNorm1D($\mathbf{C_{in}}$) normalizes each feature independently so that it has zero mean and unit variance across M measurements. Given a set of measurements x_i, with $i = 0, \ldots, M-1$, the mean (μ_c) and the variance (σ_c^2) for feature index c are evaluated as follows:

$$\mu_c = \frac{1}{M} \sum_{i=0}^{M-1} x_{i,c} \text{ and } \sigma_c^2 = \frac{1}{M} \sum_{i=0}^{M-1} \left(x_{i,c} - \mu_c\right)^2.$$

In this specific case, $M = N \times P$ where N is the number of jets per batch and P is the number of PF candidates or SVs per jet. The input features are then standardized for each x_i as follows:

$$\hat{x}_{i,c} = \frac{x_{i,c} - \mu_c}{\sqrt{\sigma_c^2 + \varepsilon}}$$

where $\varepsilon = 10^{-5}$. However, the final outputs provided by BatchNorm1D($\mathrm{C_{in}}$) are the following quantities:

$$y_{i,c} = \gamma_c \, \hat{x}_{i,c} + \beta_c$$

where γ and β are learnable parameter vectors of size $\mathrm{C_{in}}$. The elements of γ are set to start at 1, while the elements of β start at 0.

Conv1D($\mathbf{C_1 \to C_2}$) transforms tensors with input size (N, $\mathrm{C_1}$, P) through a 1×1 convolution into tensors with output size (N, $\mathrm{C_2}$, P), where N is the number of jets per batch, P is the number of PF candidates or SVs per jet, $\mathrm{C_1}$ and $\mathrm{C_2}$ are the number of input and output features. Given the batch index n, the input feature index i, the output feature index j and the PF candidates/SVs index p, the output features after the convolution are given by

$$y_{n,j,p} = \sum_{i=0}^{C_1-1} W_{j,i,0}\, x_{n,i,p}$$

where $x_{n,i,p}$ are the elements of the input tensor and the elements of $W \in \mathbb{R}^{C_2 \times C_1}$ are learnable weights. The weights are set to start from random values in the range $\left[-\sqrt{\frac{1}{C_1}}\,;\,\sqrt{\frac{1}{C_1}}\right]$.

ReLU is applied to the output $y_{n,j,p}$ of the second Batch Normalization block. Formally, the ReLU activation function is defined as

$$\mathrm{ReLU}(y_{n,j,p}) = \begin{cases} y_{n,j,p} & \text{if } y_{n,j,p} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Thus, all negative $y_{n,j,p}$ are set to zero, while positive values are left unchanged.

The model uses 3 EdgeConv blocks (see Section 3.10.1), each one using $K = 16$ neighbors to build a local graph for each point. The output of every block is used as the input to the subsequent one. Neighbor search is performed only among real PF candidates or SVs, while the dummy points introduced for padding are excluded through masking. In the first EdgeConv block, for each point the 16 nearest neighbors are identified by minimizing the Euclidean distance computed from the geometric coordinates $(\Delta\eta, \Delta\phi)$ with respect to the jet axis. Once identified, the EdgeConv block constructs the corresponding edge features for each input tensor $x \in \mathbb{R}^{N \times C \times P}$, where N is the batch size, C is the number of features per point (32 in this case) after the BatchNorm1D→Conv1D→BatchNorm1D→ReLU sequence and P is the number of PF objects plus SVs considered per jet (130). For each central point cloud features x_i, the relative feature differences with respect to its K neighbors x_j are then computed as $x_j - x_i$. These relative vectors are concatenated with the original point features x_i along the feature dimension, resulting in an edge feature tensor of shape $N \times 2C \times P \times K$. This construction allows the network to simultaneously exploit the absolute feature representation of each point cloud and the relational information between each point cloud and its local neighborhood, which is a key component of the EdgeConv operation. In each EdgeConv block the edge feature tensor is subsequently processed by a sequence of three layers:

- Conv2D($\mathrm{C_{in}} \rightarrow \mathrm{C_{out}}$),
- BatchNorm2D($\mathrm{C_{out}}$),
- ReLU.

The output of each layer is fed as input to the next one, enabling a progressive refinement of the local feature representations.

Conv2D is equivalent to Conv1D, but applied to a tensor with an additional dimension corresponding to the neighbor index k. The operation performs the same linear transformation across the feature dimension, independently for each point p and each neighbor k.

BatchNorm2D(C), where C denotes the number of features, normalizes each feature independently so that it has zero mean and unit variance across M measurements as BatchNorm1D(C).

The only difference with respect to BatchNorm1D is that relative feature differences with the K nearest neighbors are also included

Following the three layers of the Conv2D→BatchNorm2D→ReLU sequence, a neighbor aggregation step is applied to combine the information from the K nearest neighbors. Specifically, the feature responses associated with each central point are averaged over the neighbor dimension, resulting in an aggregated feature tensor. For a given batch index n, feature channel index c, and point index p, the aggregated feature is defined as

$$fts_{n,c,p} = \frac{1}{K}\sum_{k=0}^{K-1} x_{n,c,p,k}$$

This operation produces a single feature vector per node by summarizing the contributions of its local neighborhood, thereby yielding a fixed-size representation that can be propagated to subsequent network layers.
After the neighbor aggregation step, the original point features provided as input of the EdgeConv block are passed through a sequence of Conv1D(32→$\mathrm{C_{out}}$)→BatchNorm1D($\mathrm{C_{out}}$), where $\mathrm{C_{out}}$ is the number of output channels of the last Conv2D($\mathrm{C_{in}}$→$\mathrm{C_{out}}$)→BatchNorm2D($\mathrm{C_{out}}$)→ReLU sequence in the EdgeConv block, producing a transformed feature tensor denoted as $sc_{n,c,p}$. The final output of the EdgeConv block is then obtained by adding $sc_{n,c,p}$ to the aggregated edge features $fts_{n,c,p}$ and applying a ReLU activation function to the resulting sum. This operation implements a residual connection, which facilitates gradient propagation and stabilizes the training process. Through this mechanism, the network can preserve the original point-level information while simultaneously incorporating the newly learned relational features, effectively allowing each block to refine rather than overwrite the input representation.
From the second block onward, the neighbor search is therefore performed in the learned feature space: the Euclidean distance is computed between the C-dimensional feature vectors output by the previous block. In this way, the first iteration connects spatially nearby points, while subsequent iterations connect points that are close in the feature representation.
For the three EdgeConv blocks, the numbers of output features are chosen as (64, 64, 64), (128, 128, 128) and (256, 256, 256), respectively, meaning that each block consists of three layers of the Conv2D→BatchNorm2D→ReLU sequence with identical output feature dimensionality.
The tensors produced by the three EdgeConv blocks are concatenated along the channel dimension for a total of 448 (64+128+256) features and then transformed using a fusion block consisting of Conv1D(448→384)→BatchNorm1D→ReLU. To obtain an overall representation of the jet, the per-point features x are aggregated by means of a sum over the valid points P' (the points that really exist and not the padding ones added artificially just to make the network work), normalized by the actual number of elements present in the jet. For a given batch index n, feature channel index c, and point index p, the aggregated feature $Fts_{n,c}$ is defined as

$$Fts_{n,c} = \frac{1}{P'}\sum_{p=0}^{P'-1} x_{n,c,p}$$

The overall representation of the jet is finally processed by a fully-connected network consisting of the following algorithms, ordered exactly as listed below:

- Linear(384→256)
- ReLU

- Dropout(0.1)
- Linear(256→N_{classes})

Linear($\mathbf{C_1 \rightarrow C_2}$) applies a liner transformation to tensors with input size (N, C_1) into tensors with output size (N, C_2), where N is the number of elements per batch, C_1 and C_2 are the number of channels. Given the batch index n, the input channel index i and the output channel index j, the input features after the linear transformation $y_{n,j}$ are given by

$$y_{n,j} = \sum_{i=0}^{C_1-1} W_{j,i}\, x_{n,i} + b_j$$

where $x_{n,i}$ are the elements of the input tensor, the elements of $W \in \mathbb{R}^{C_2 \times C_1}$ and the biases $b_j \in \mathbb{R}^{C_2}$ are the learnable weights. b_j are added to each element of the batch. The weights and the biases are initialized at random values in the range $\left[-\sqrt{\frac{1}{C_1}}\ ;\ \sqrt{\frac{1}{C_1}}\right]$.

Dropout(p) sets some features of the nodes to zero with probability p during training. All the features that are not set to zero are scaled by the factor $\frac{1}{1-p}$. The purpose of dropout is to prevent the network from becoming overly dependent on specific features and thus to reduce overfitting.

The final raw output values, obtained from the last linear transformation, are called logits, *i.e.* the unnormalized network scores for each class identified as targets. The model is designed to classify events into 10 output classes ($N_{\text{classes}} = 10$):

- **Hbb**: a resonance decaying into a $b\bar{b}$ pair;
- **Hcc**: a resonance decaying into a $c\bar{c}$ pair;
- **Hqq**: a resonance decaying into a light quark pair $q\bar{q}$ with $q = u, d, s$;
- **Hgg**: a resonance decaying into a gluon pair;
- **H**$\tau\tau$: a resonance decaying into a $\tau\bar{\tau}$ pair, with both τ decaying hadronically;
- **H**$\tau\mu$: a resonance decaying into a $\tau\bar{\tau}$ pair, with one τ decaying hadronically and the other leptonically into a muon ($\tau \rightarrow \mu\nu_\mu\nu_\tau$);
- **H**τe: a resonance decaying into a $\tau\bar{\tau}$ pair, with one τ decaying hadronically and the other leptonically into an electron ($\tau \rightarrow e\nu_e\nu_\tau$);
- **QCD2hf**: a QCD jet with 2 or more b-hadrons or with 2 ore more c-hadrons and no b-hadrons;
- **QCD1hf**: a QCD jet with 1 b-hadron or with 1 c-hadron and no b-hadrons;
- **QCD0hf**: QCD jet with zero heavy-flavor hadrons.

The logits are transformed into the output probabilities, *e.g.* prob(Hbb), by a softmax function. Let $z_{i,c}$ be the logit for a jet i and class c, the softmax for jet i giving the probability associated to the class c ($p_{i,c}$) is

$$p_{i,c} = \text{softmax}_c(z_i) = \frac{e^{z_{i,c}}}{\sum_{j=1}^{N_{classes}} e^{z_{i,j}}}, \qquad \text{with} \qquad \sum_{c=1}^{N_{classes}} p_{i,c} = 1.$$

The here described resulting network has these $\sim$ 580k learning parameters. During training, the network processes one batch of jets at a time. For each batch it computes the loss function L_{batch}, a scalar-valued function that quantifies the discrepancy between the network predictions and the true labels. The gradients of the loss function with respect to the network parameters θ, $\nabla_\theta L_{\text{batch}}(\theta)$, are then computed. The parameters are updated according to a gradient descent step:

$$\theta_{t+1} \;=\; \theta_t - lr\, \nabla_\theta L_{\text{batch}}(\theta_t), \tag{5.1}$$

where θ_t are the parameters at step t and lr is the learning rate. This procedure is repeated over batches to update the network parameters and minimize the expected loss.
The model trained in this thesis is built upon the Weaver framework [112]. Weaver is an open-source, ML R&D framework developed for high-energy physics (HEP) that streamlines the full training workflow. Weaver reads common HEP formats (*e.g.* ROOT [113] files and awkward [114] arrays) and performs on-the-fly pre-processing (selections, derived variables, standardization) driven by a YAML [115] data configuration file. Weaver loads the model and the loss by importing a user-provided Python model file. At runtime Weaver composes the dataset and then runs a PyTorch [116] training loop with check-pointing[1], logging[2], optimizer configuration[3] and profiting of multi Graphics Processing Units (GPUs) support. Trained PyTorch models can be exported to ONNX [117] for fast inference, which simplifies deploying models in the HLT environment.

5.2 Simulated samples

All MC samples considered were generated at $\sqrt{s} = 13.6$ TeV and processed using the corresponding 2025 detector and run conditions. Parton showering and hadronization were modeled with PYTHIA v8.230, employing the CP5 tune to describe the underlying event using the NNPDF3.1 PDFs at NNLO accuracy. PU effects were incorporated by superimposing minimum-bias interactions, the resulting PU distribution is reweighted to reproduce that observed in data. Event reconstruction is performed up to the HLT level, without applying any additional offline reconstruction. The CMS detector response is simulated with GEANT4.

5.2.1 Signal samples

To ensure that the network output probabilities are independent of the mass and transverse momentum of the target resonance, the training signal is modeled using a hypothetical $X \to HH$ process (Xto2H), where a graviton-like resonance (X) decays into two Higgs bosons. The masses of both X and H are varied in the simulation, with m_X spanning the range 500–6000 GeV and m_H ranging from 20 to 250 GeV. To guarantee adequate statistics in each decay channel, the Higgs boson is forced to decay with equal probability (1/7) into $b\bar{b}$, $c\bar{c}$, gg, $\tau\bar{\tau}$, $e\bar{e}$, and $\mu\bar{\mu}$ final states, while the remaining 1/7 branching fraction is uniformly distributed among the $u\bar{u}$, $d\bar{d}$, and $s\bar{s}$ decays. Events were generated at LO using PYTHIA8 (CP5 tune).
The performance of the model were also observed for $H \to b\bar{b}$ as a dedicated case study because of its relevance and its data-like kinematics. It is ideal for probing the analysis performance: it permits validation of selection and reconstruction procedures under conditions close to those observed in data, and is particularly useful for studying signal-specific effects. $H \to b\bar{b}$ produced via ggF (GluGluHTo2B) is simulated using the HJMINLO event generator [118] with the Higgs boson mass

[1] Saving the current state of the model periodically so that training can be resumed later without loss of progress.
[2] Recording metrics, losses, and other relevant information during training for monitoring.
[3] The choice and setup of the algorithm that updates the model parameters based on the computed gradients.

set to 125 GeV including finite top quark mass effects [119]. The simulated datasets used are listed in Table 5.1.

Process	Sample	Generator	M_H slice [GeV]	M_X slice [GeV]	Cross section	$\sigma \cdot BR$ [pb]
$X \to HH$	XTo2H	PYTHIA8 (LO)	20–250	500–1000	LO	-
$X \to HH$	XTo2H	PYTHIA8 (LO)	20–250	1000–4000	LO	-
$X \to HH$	XTo2H	PYTHIA8 (LO)	20–250	4000–6000	LO	-
ggF $H \to b\bar{b}$	GluGluHTo2B	POWHEG HJMINLO (NLO)	125	-	NLO	0.306

Table 5.1: List of signal MC samples used in the PNet-MD training and testing with the corresponding generator, M_H slice, M_X slice, production cross section order and production cross sections at $\sqrt{s} = 13.6\,\text{TeV}$ times relevant branching ratios. The cross sections of the XToHH samples are not shown because they were not used.

Figure 5.1 shows the p_T and m_{SD} distributions for AK8 jets matching a generator level $H \to b\bar{b}$ for XTo2H and GluGluHTo2B datasets. While the p_T and m_{SD} distributions derived from the XtoHH dataset are fairly flat, the GluGluHTo2B p_T distribution shows an exponentially decreasing shape and the GluGluHTo2B m_{SD} distribution shows a peak around 125 GeV and an exponential decrease at higher values. The sofGluGluHTo2B m_{SD} distribution exhibits a shoulder below m_{SD} ~100 GeV, arising from low-p_T, non-boosted jets. In these jets, the soft-drop algorithm removes a significant fraction of low-momentum constituents at large angles, leading to a measured mass that is lower than the true jet mass.

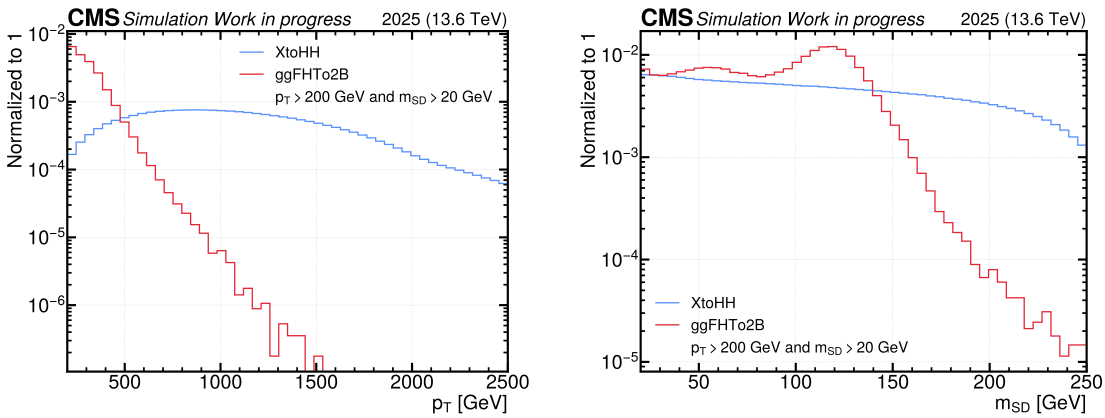

Figure 5.1: p_T (left) and m_{SD} (right) distributions for AK8 jets matching a generator level $H \to b\bar{b}$ of XTo2H (light blue histograms) or GluGluHTo2B (red histograms) datasets, with $p_T > 200$ GeV and $m_{SD} > 20$ GeV. All distributions are normalized to 1.

5.2.2 Background samples

The QCD multijet background is described by hard QCD scattering processes producing final states with four energetic partons at the ME level, providing an accurate representation of high-multiplicity multijet topology. Events are generated at LO using MadGraph5_aMC@NLO. The matching between jets produced at the ME level and those arising from the parton shower is performed using the MLM matching scheme, in order to avoid double counting of radiative emissions. The simulated datasets used are listed in Table 5.2.

Sample	Generator	H_T slice [GeV]	Cross Section	$\sigma \cdot BR$ [pb]
QCD-4Jets	MG5_aMC@NLO (LO)	40–70	LO (MLM)	3.12×10^8
QCD-4Jets	MG5_aMC@NLO (LO)	70–100	LO (MLM)	5.87×10^7
QCD-4Jets	MG5_aMC@NLO (LO)	100–200	LO (MLM)	2.52×10^7
QCD-4Jets	MG5_aMC@NLO (LO)	200–400	LO (MLM)	1.96×10^6
QCD-4Jets	MG5_aMC@NLO (LO)	400–600	LO (MLM)	9.70×10^4
QCD-4Jets	MG5_aMC@NLO (LO)	600–800	LO (MLM)	1.32×10^4
QCD-4Jets	MG5_aMC@NLO (LO)	800–1000	LO (MLM)	3.05×10^3
QCD-4Jets	MG5_aMC@NLO (LO)	1000–1200	LO (MLM)	8.93×10^2
QCD-4Jets	MG5_aMC@NLO (LO)	1200–1500	LO (MLM)	3.86×10^2
QCD-4Jets	MG5_aMC@NLO (LO)	1500–2000	LO (MLM)	1.26×10^2
QCD-4Jets	MG5_aMC@NLO (LO)	> 2000	LO (MLM)	26.6

Table 5.2: List of QCD multijet MC samples used in the PNet-MD training and testing, with the corresponding generator, H_T slice, production cross section order and production cross sections at $\sqrt{s} = 13.6$ TeV times relevant branching ratios. H_T is the scalar p_T sum of all the jets in the event.

5.3 Training

QCD and XtoHH AK8 jets used for training and testing are required to satisfy the following selection cuts:

- jet $|\eta| < 2.4$ to be well within the tracker acceptance and away from its edge;
- jet $p_T > 200$ GeV, to look at a medium and highly boosted regime;
- The reconstructed jet m_{SD} and the invariant mass of the nearest generator-level jet (with $\Delta R < 0.4$ from the reconstructed jet) are required to be greater than 20 GeV.

These requirements ensure good tracking coverage, focus the training on the kinematic region with medium, high boosted jet and exclude soft/very-low-mass jets.
Signal and background jets were classified according to their generator level information and associated to the 10 PNet-MD classes, described in section 5.1. All jets from QCD datasets were labeled as QCD and sub-classified according to the number of generator-level heavy-flavor hadrons among the constituents of the jet.
Regarding the signal datasets, a jet was considered as signal jet if matched to a generator-level Higgs H_{Gen} which implies that:

- $\Delta R(\text{jet}, H_{Gen}) < 0.4$;
- the ΔR distanc between jet and both H_{Gen} daughters is 0.5.

Matches were considered valid only when the jet and the H_{Gen} were mutual nearest neighbors in ΔR. For decay channels that do not involve a Higgs decaying to τ-pairs (specifically $H \to b\bar{b}$, $H \to c\bar{c}$, $H \to q\bar{q}$, and $H \to gg$), the label assigned to the jet was the H_{Gen} decay mode (*e.g.* a jet matched to a Higgs that decays to $b\bar{b}$ is labeled Hbb, etc.). While, to identify jets compatible with $H \to \tau\tau$, the τ decay modes had to be first identified. A τ decaying into a muon (τ_μ) or an electron (τ_e) is defined as a gen-level τ which have three daughters and one of the daughters is a muon or an electron and neutrinos; its four-vector is the sum of the three daughters four-vectors. Instead, a hadronic τ (τ_h) is formed from the gen-level τ decay products, using only the daughter particles that are not neutrinos, electrons or muons, and summing their four-momenta if the total p_T is greater than 20 GeV. A jet matched to a generator-level Higgs decaying into taus is labeled H$\tau\tau$, H$\tau\mu$, or Hτe if it shows a distance ΔR lower than 0.5 with respect to two τ_hs, one τ_h and one τ_μ, or one τ_h and one τ_e respectively.

5.3.1 Model inputs

A feature vector of 23 (10) elements is associated to the PF candidates (SVs), using the same input features as in the current model in the HLT menu (previous model). Here the first 15 PF candidates features are described:

- **jet_pfcand_pt_log**: $\ln p_T$ of the PF candidate ($\ln p_T$ is better than p_T because the p_T distribution covers several order of magnitude);
- **jet_pfcand_energy_log** $\ln E$ of the PF candidate ($\ln E$ is better than E because the E distribution covers several order of magnitude);
- **jet_pfcand_deta**: $\Delta\eta$ distance between the PF candidate and the AK8 jet axis;
- **jet_pfcand_dphi**: $\Delta\phi$ distance between the PF candidate and the AK8 jet axis;
- **jet_pfcand_eta**: η of the PF candidate;
- **jet_pfcand_charge**: charge associated to the PF candidate;
- **jet_pfcand_frompv**: integer value between 0 and 3, indicating how likely the PF candidate is originated from the PV;
- **jet_pfcand_nlostinnerhits**: integer value between 0 and 2 indicating if the PF candidate has 0, 1 or more than 1 missing hits (obtained from the track extrapolation) in the inner layers of the pixel tracker before the first reconstructed hit;
- **jet_pfcand_dz**: Δz distance between the PF candidate and the PV;
- **jet_pfcand_dxy**: Δxy distance between the PF candidate and the PV;
- **jet_pfcand_etarel**: PF candidate rapidity on the AK8 jet direction;
- **jet_pfcand_pperp_ratio**: the momentum component of the PF candidate perpendicular to the AK8 jet direction;
- **jet_pfcand_ppara_ratio**: the momentum component of the PF candidate parallel to the AK8 jet direction;
- **jet_pfcand_npixhits**: number of hits of the PF candidate in the pixel tracker;
- **jet_pfcand_nstriphits**: number of hits of the PF candidate in the silicon strip tracker.

The remaining 8 PF candidate features are evaluated using the track associated with the PF candidate:

- **jet_pfcand_track_chi2**: normalized χ^2 for the reconstruction of the track;
- **jet_pfcand_track_qual**: the reconstruction quality of the track, represented by an integer ranging from 0 to a maximum value of 5;
- **jet_pfcand_dzsig**: significance of the Δz distance between the PF candidate and the PV ($\Delta z/\sigma_{\Delta z}$), $\sigma_{\Delta z}$ is defined only if a track is associated with the PF candidate;
- **jet_pfcand_dxysig**: significance of the Δxy distance between the PF candidate and the PV ($\Delta xy/\sigma_{\Delta xy}$), $\sigma_{\Delta xy}$ is defined only if a track is associated with the PF candidate;
- **jet_pfcand_trackjet_d3d**: the signed 3D impact parameter of the track with respect to the PV and the jet direction (I_{3D});
- **jet_pfcand_trackjet_d3dsig**: significance of I_{3D} ($|I_{3D}|/\sigma_{I_{3D}}$);
- **jet_pfcand_trackjet_dist**: minimum distance between the track and the direction of the AK8 jet (assuming that it starts from PV);
- **jet_pfcand_trackjet_decayL**: distance along the jet direction between the track and the PV.

If no associated track is available, all 8 track features are set to zero.
For the SVs, the following observables are considered:

- **jet_sv_pt_log**: $\ln p_T$ of the SV;
- **jet_sv_mass**: the SV mass is estimated as the invariant mass of the four-momenta of all associated charged tracks;
- **jet_sv_deta**: $\Delta\eta$ distance between the SV and the AK8 jet axis;
- **jet_sv_dphi**: $\Delta\phi$ distance between the SV and the AK8 jet axis;
- **jet_sv_ntrack**: number of charged particles used to reconstruct the SV;
- **jet_sv_chi2**: normalized χ^2 for the SV reconstruction;
- **jet_sv_dxy**: the signed Δxy distance between the SV and the PV along the jet direction;
- **jet_sv_dxysig**: significance of the signed Δxy distance between the SV and the PV along the jet direction ($|\Delta xy|/\sigma_{\Delta xy}$);
- **jet_sv_d3d**: the signed 3D distance between the SV and the PV along the jet direction;
- **jet_sv_d3dsig**: significance of the signed 3D distance between the SV and the PV along the jet direction.

As an example, Figure 5.2 show the unweighted jet_pfcand_pt_log, jet_pfcand_trackjet_d3dsig, jet_sv_dxysig and jet_sv_deta distributions for XtoHH AK8 jets labeled as $H \to b\bar{b}$ and QCD AK8 jets. The distributions show a clear difference in shape between the two samples. In appendix C, all the remaining input features are shown.

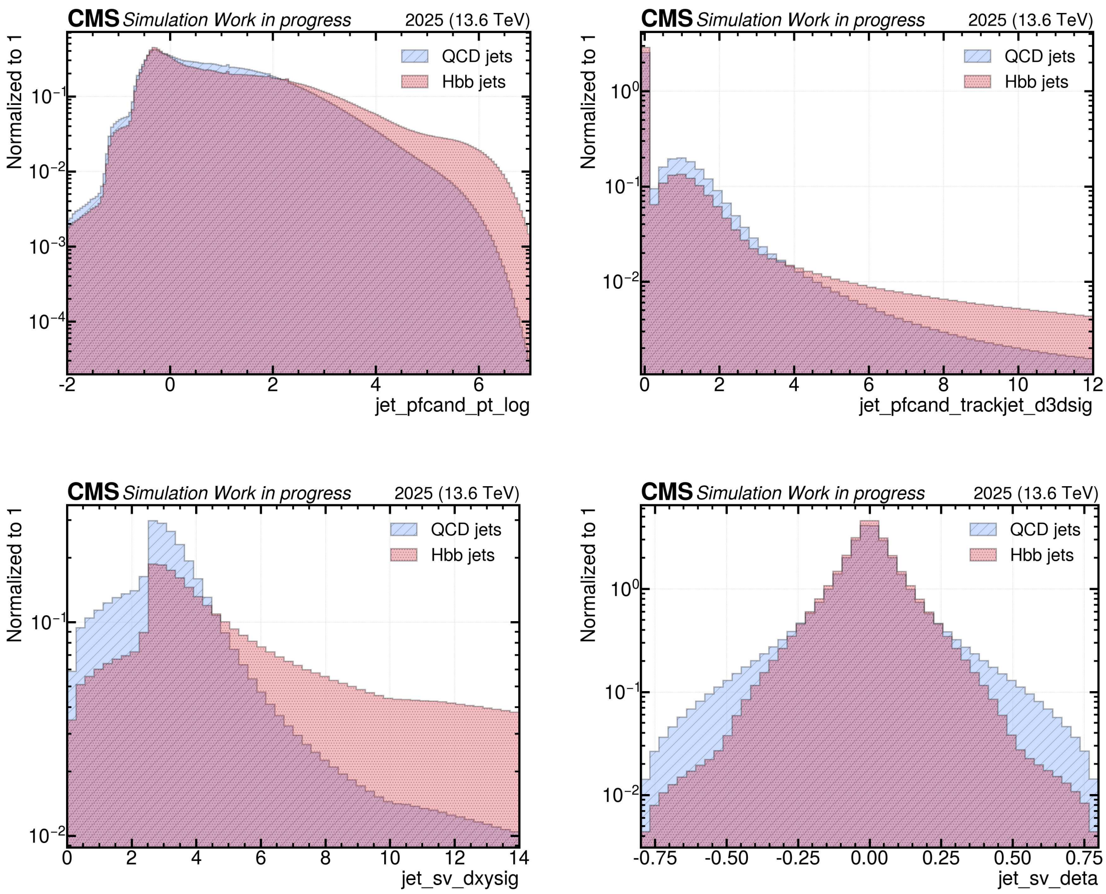

Figure 5.2: Normalized to 1 unweighted jet__pfcand__pt__log (top-left), jet__pfcand__trackjet__d3dsig (top-right), jet__sv__dxysig (bottom-left) and jet__sv__deta (bottom-right) distributions for XtoHH AK8 jets labeled as $H \to b\bar{b}$ (red histograms) and QCD AK8 jets (light blue histograms). All jets satisfy the training kinematic requirements.

5.3.2 Samples optimization

Although the signal sample used for training was generated to provide a uniform coverage of the two-dimensional (p_T, m_{SD}) phase space for AK8 jets, both signal and background samples didn't show a fully flat distribution in momentum and mass. To mitigate this effect and to ensure the independence of the PNet-MD output with respect to p_T and m_{SD}, it was decided to allow Weaver to apply per-label jet weights to flatten a two-dimensional (p_T, m_{SD}) distribution with the following binning:

- $p_T = [200, 250, 315, 396, 499, 628, 791, 995, 1253, 1577, 1987, 2500]$ GeV
- $m_{SD} = [20, 45, 65, 85, 105, 125, 145, 165, 185, 205, 225, 250]$ GeV

chosen to guarantee enough statistics in each bin and for the various classes (Hbb, Hcc, Hqq, H$\tau\tau$, H$\tau\mu$, Hτe), QCD2hf, QCD1hf, QCD0hf).
Let $n_{l,ij}$ be the raw (unweighted) count in bin (i, j) for class l. From the distribution of non-zero bins a small threshold τ_l was defined as

$$\tau_l = 0.01 \cdot \text{median}\big(\{n_{l,ij} \mid n_{l,ij} > 0\}\big),$$

and bins with $n_{l,ij} \leq \tau_l$ were ignored when estimating the reference population (r_l) for the reweighting of the specific class l. Among the remaining bins, the reference value r_l was defined as the 15$^{\text{th}}$ percentile of $\{n_{l,ij} > \tau_l\}$; it represents a low but robust level of bin population, avoiding sensitivity to extreme values. The per-bin preliminary weight was

$$w_{l,ij} = \begin{cases} 0, & n_{l,ij} = 0, \\ \min\left(1, \dfrac{r_l}{n_{l,ij}}\right), & n_{l,ij} > 0. \end{cases}$$

The class effective weighted count is

$$E_l \;=\; \frac{1}{C_l} \sum_{i,j} n_{l,ij}\, w_{l,ij}$$

where C_l is a class normalization factor, used to enhance or reduce a given class population. A class-level equalization was enforced so that each label contributes approximately the same effective number of events to the sampler. Let the common conservative target be

$$T \;=\; 0.9 \cdot \min_l \, E_l$$

the class scaling factor is

$$s_l \;=\; \frac{T}{E_l}$$

which is applied uniformly to the per-bin. The final weighted count in bin (i, j) is therefore

$$n_{l,ij}^{\text{final}} \;=\; n_{l,ij}\, s_l\, w_{l,ij}.$$

This procedure therefore performs conservative downsampling of abundant classes, reducing overfitting risk (the situation in which a model fits data too closely, capturing noise rather than general features), while leaving low-population bins at most unchanged.
The class normalization factor chosen for the training were:

- H$\tau\tau$, H$\tau\mu$, H$\tau e \to 1.0$: τ-containing Higgs decay channels are relatively rare but have distinctive signatures. They were kept at unit weight to avoid overfitting while letting the per-bin reweighting handle the (p_T, m_{SD}) flattening.

- Hbb, Hcc, Hqq $\to 1.5$: hadronic Higgs decays to quarks are prime physics targets, in particular $H \to b\bar{b}$ and $H \to c\bar{c}$. A moderate upweight (1.5) allow the network to allocate capacity to distinguish them from QCD. This helps improve sensitivity to important signal modes while remaining conservative.

- Hgg $\to 1.0$: gluon final states are often kinematically similar to QCD light jets and less of a primary signal target here, so they are left at baseline weight. This avoids wasting capacity on a class that is both ambiguous and less critical for the physics goals.

- QCD2hf, QCD1hf, QCD0hf $\to 2.0$: the QCD classes represent the dominant background and must be modeled robustly across the kinematic plane. Increasing their global weight helps the network learn fine-grained background structure and reduces the chance that reweighting to flatten kinematics excessively suppresses background influence. A factor of 2 is a deliberate, moderate upweight to prioritize background discrimination while preventing the loss from being completely dominated by it.

Of all the training statistics available, 6 out of 7 jets were used for training, while the rest were used for testing. The raw distributions of QCD and signal jets are shown in Figure 5.3, together with the (p_T, m_{SD}) weights distributions for jets labeled as QCD2hf and Hbb. As can be seen, at low p_T there are considerably more QCD jets than signal ones, while at high p_T the opposite is observed. Consequently, weaver assigns weights that differ by few orders of magnitude to flatten the distributions, with the result that the produced model does not behave uniformly across the entire phase space.
To reduce the difference in orders of magnitude between the weights, and to balance the raw statistics between QCD and signal, the QCD sample statistics were downsampled for $p_T < 791$ GeV so that, in each p_T bin, the QCD and signal samples had similar statistics. For the same reason, the signal statistics for $p_T >$1987 GeV were downsampled to balance the QCD statistics.
Figure 5.4 shows the weighted (p_T, m_{SD}) distribution for QCD2hf and Hbb jets. In most of the (p_T, m_{SD}) bins, the number of weighted jets is the same, indicating that the balancing procedure is effective across a wide region of the phase space.
Although this binning scheme produces apparently flat two-dimensional distributions, a finer binning reveals residual structures that would have biased the algorithm development. Figure 5.5 shows the weighted m_{SD} distributions for QCD2hf and Hbb jets, presented separately for transverse momenta above and below 315 GeV. For $p_T > 315$ GeV, the QCD2hf and Hbb distributions are compatible, whereas for $p_T < 315$ GeV GeV the QCD sample exhibits a relative deficit of low-m_{SD} events compared to Hbb. These shape differences induce a residual dependence of the model on p_T, in contrast with the original motivation of the training strategy. Consequently, the training was restricted to jets with $p_T > 315$ GeV, while jets with $p_T > 200$ GeV were retained for testing.

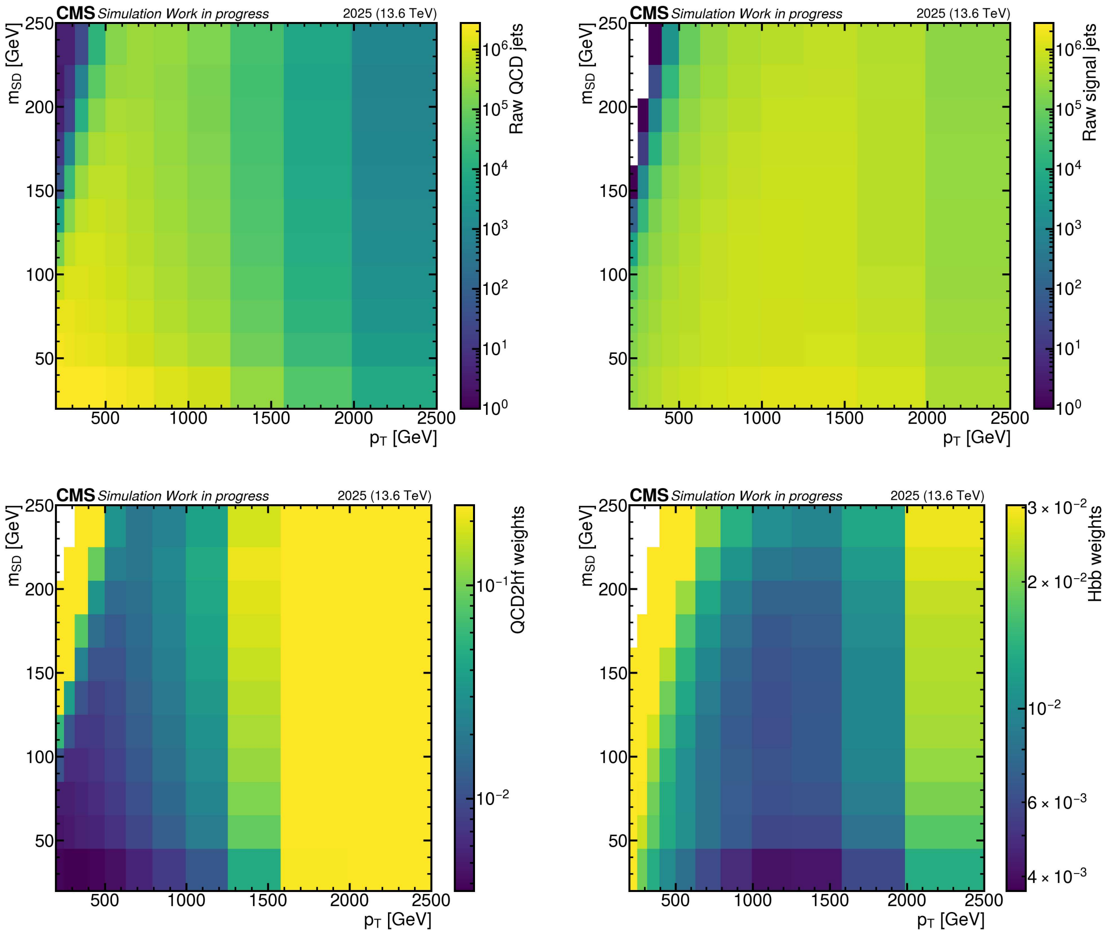

Figure 5.3: Unweighted (p_T, m_{SD}) distributions for QCD (top-left) and XToHH (top-right) jets. (p_T, m_{SD}) weights for QCD2hf (bottom-left) and Hbb (bottom-right) jets. The histogram binning is the same used for the reweighting procedure.

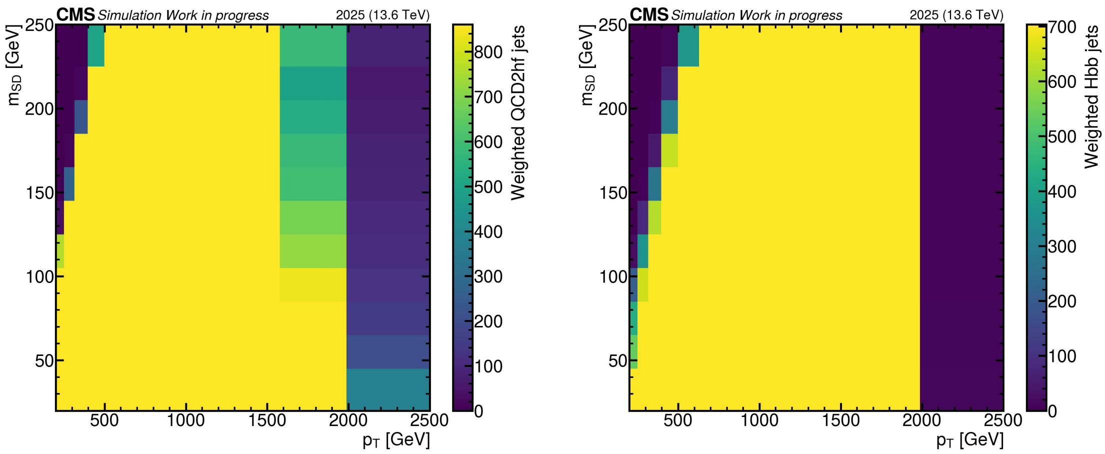

Figure 5.4: Weighted (p_T, m_{SD}) distributions for QCD2hf (left) and Hbb (right) jets. The binning is the same used in the training.

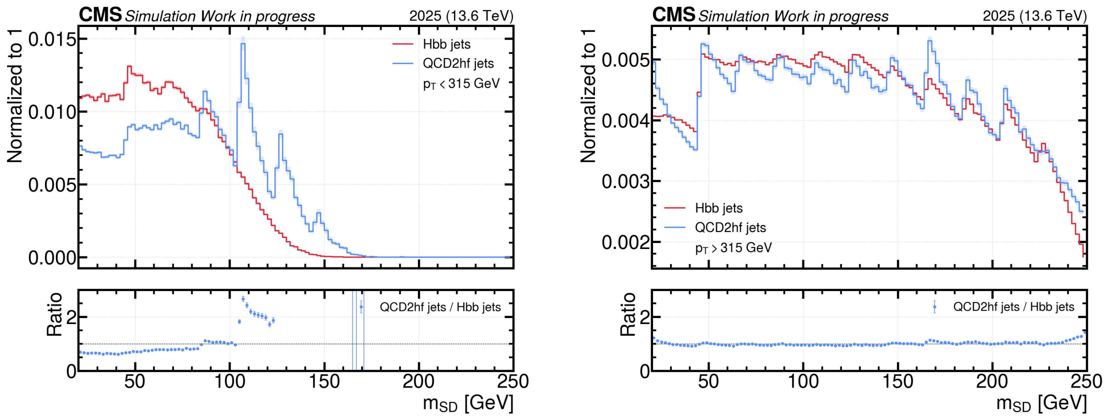

Figure 5.5: Weighted m_{SD} distributions for QCD2hf (light blue histograms) and Hbb (red histograms) jets with $p_T < 315$ GeV (left) and $p_T > 315$ GeV (right). All distributions are normalized to 1. The lower plots show the ratio between the QCD2hf distributions and the Hbb ones.

5.3.3 Training development

The model was trained using four NVIDIA A100 GPUs provided by the ReCaS-Bari data center [120]. To avoid overfitting, from the set of jets employed in the training procedure, 80% constituted the real training sample, whereas the remaining 20% formed the validation sample[4] used to check the model accuracy (number of correctly identified jets over the total number of jets). The training was performed for multiple epochs, each corresponding to one complete pass over the full training set. During each training epoch Weaver updates the model weights after each batch in order to minimize the loss function. Therefore multiple epochs are required to converge to a stable solution. For the first 14 epochs, the learning rate remained constant, maintaining the initial value lr_0. Starting from epoch number 14, the learning rate was multiplied by a factor $\gamma = 0.01^{\frac{1}{6}} \approx 0.681$ according to

$$lr_{\text{epoch}+1} = \gamma \cdot lr_{\text{epoch}}.$$

This gradual reduction of the learning rate allows the optimizer to take smaller and more precise steps as training progresses, which helps in stabilizing the training convergence. The reaming training parameters are the following:

- **num_workers** = 6: number of threads used to load an process the dataset;
- **num_epochs** = 20: number of training epochs;
- **batch_size** = 1024: number of jets contained in each batch during training;
- **start_lr** = $5 \cdot 10^{-3}$: starting value for the learning rate.

The PyTorch CrossEntropyLoss function [121] was used as loss function to measure the model performance error. The cross-entropy averaged over N jets is

$$\mathrm{L} = -\frac{1}{N}\sum_{i=1}^{N} \log p_{i,c} = -\frac{1}{N}\sum_{i=1}^{N}\left(z_{i,c} - \log \sum_{j=1}^{N_{classes}} e^{z_{i,j}}\right).$$

where $z_{i,c}$ is the logit for jet i and true class c and $p_{i,c}$ is the softmax for jet i giving the true class c. For each jet i, the predicted class $y_i \in \{Hbb, \ldots, QCD2hf\}$ corresponds to the logit z_{i,y_i} with the highest value. The classification accuracy was estimated by comparing, for each jet, the predicted class to the true label, counting a jet as correct if the predicted class matched the label. The accuracy for an epoch was computed as the total number of correctly classified jets divided by the total number of jets processed. Figure 5.6 shows the average loss function and the accuracy evaluated on the training sample and the accuracy evaluated on the validating sample as function of the epochs. The model used was the one produced after the epoch number 14, since this corresponds to the highest validation accuracy (∼0.58). This epoch corresponds to an average loss and a training accuracy of ∼0.97 and ∼0.61 respectively.

[4]The validation sample is a subset of the training data that is held out and not used to update model parameters; it is used to monitor generalization performance, guide hyperparameter tuning, and detect overfitting by providing an unbiased estimate of how well the model is likely to perform on unseen data.

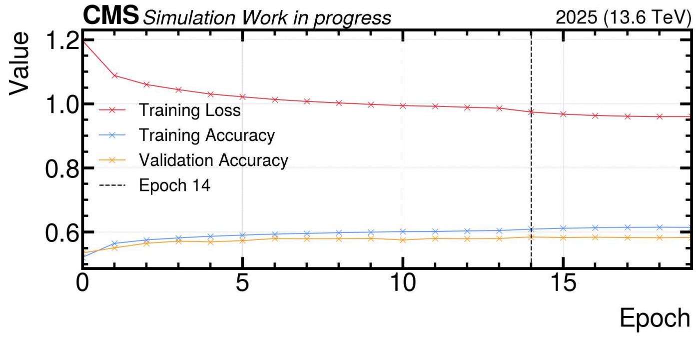

Figure 5.6: The average cross entropy loss (red line) and the classification accuracy (light blue line) of the training sample and the classification accuracy of the validating sample (yellow line) after each training epoch. Epoch 14 is the one with the highest validation accuracy (∼0.58), corresponding to an average loss and a training accuracy of ∼0.97 and ∼0.61 respectively.

5.3.4 Testing

The trained model (new model) was compared with the previous model on the remaining 1/7 jets constituting the testing sample. Jets were weighted as in the training stage. The jets used for testing are selected with the same criteria as the training sample, except that the p_T requirement is lowered to $p_T > 200$ GeV. Figure 5.7 shows the PNet-MD probability outputs prob(Hbb), prob(Hcc), prob(H$\tau\tau$) and prob(QCD), the sum of prob(QCD2hf), prob(QCD1hf) and prob(QCD0hf), for the jets with the corresponding true label. The new model more accurately indicate the probability associated with the corresponding label, given that its distributions are less populated at 0 and more populated near 1 than those obtained with the previous model.
Figure 5.8 shows the average value of the PNet-MD$_{\text{bbvsQCD}}$ score (see Section 3.10.2) for the new and the previous model in each (p_T, m_{SD}) bin for QCD and Hbb jets. The new model improved the independence of p_T and m_{SD} with respect to the previous one, since the score average value is much more uniform across the (p_T, m_{SD}) plane.
Figure 5.9 shows the ROC curves obtained for both PNet-MD models, considering Hbb, Hcc, and H$\tau\tau$ jets as signal and QCD jets as background, for AK8 jets with $|\eta| < 2.4$, $p_T > 230$ GeV and $m_{SD} > 40$ GeV. For each ROC curve, corresponding to a given Hxx signal category, the discriminating score is defined as

$$\frac{prob(Hxx)}{prob(Hxx) + prob(QCD2hf) + prob(QCD1hf) + prob(QCD0hf)}.$$

The plots show that, at a given signal efficiency, the new model provides a stronger rejection of QCD background once tested on flat (p_T, m_{SD}) samples. For instance, with the new PNet-MD$_{\text{bbvsQCD}}$ score, the threshold corresponding to a $H \to b\bar{b}$ signal efficiency of 94% yields an approximately 40% reduction in QCD efficiency relative to the previous model. Given the improvements achieved with the new model, a path with the new model was proposed and later included into the CMS HLT trigger. Prior to inclusion of the new path in the *pp* data-taking menu, feasibility studies were required to estimate the path selection rate in data and to quantify the expected contributions from signal and background.

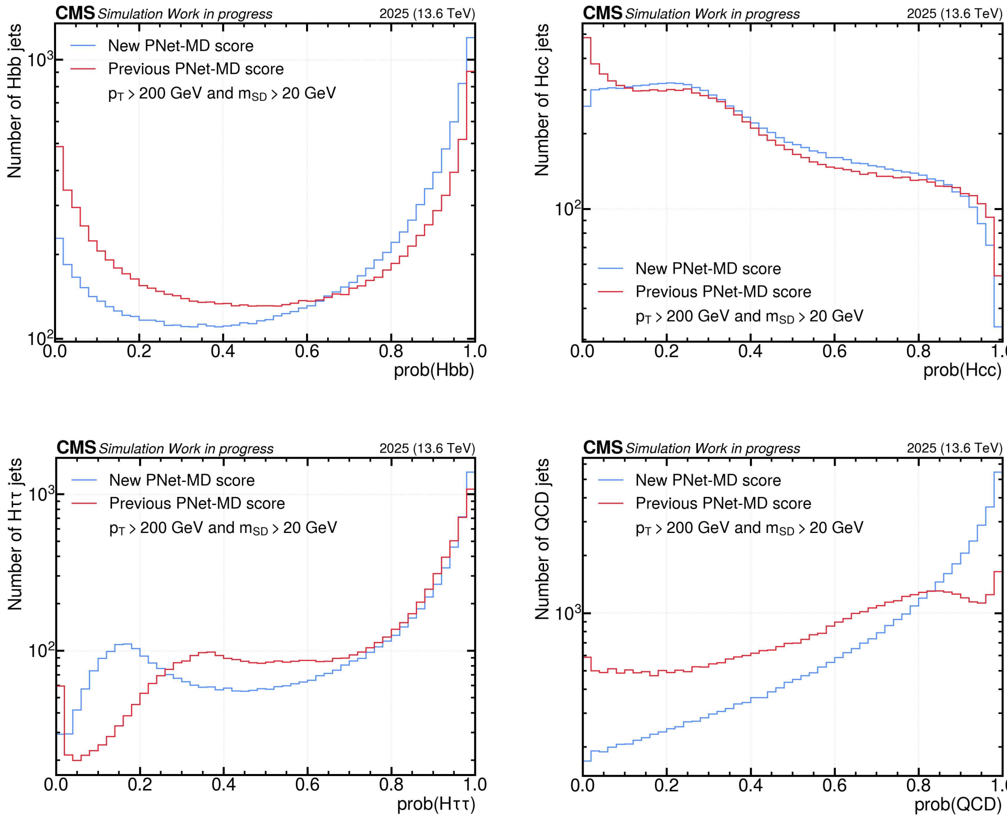

Figure 5.7: Distributions of the predicted probabilities for a jet to belong to the classes Hbb (top-left), Hcc (top-right), H$\tau\tau$ (bottom-left), and QCD (bottom-right). Each panel shows the probability assigned to the corresponding class, obtained with the new PNet-MD model (light blue histogram) and the previous PNet-MD model (red histogram). Jets are weighted as in the training.

5.4 Feasibility studies

The new model was tested on 2025 data corresponding to 0.23 fb^{-1} to evaluate the dependence from p_T and m_{SD} and to estimate the impact in pp data-taking at trigger level. Figure 5.10 illustrates the impact of applying a PNet-MD score selection that ensures the same signal efficiency (94%) for both discriminators on the p_T and m_{SD} distributions. The application of the previous score results in a visible sculpting of the distributions, particularly in m_{SD}. In contrast, the new trained algorithm does not induce any significant distortion of the distribution shapes. This behavior demonstrates that the new model is effectively independent of the AK8 p_T and m_{SD}, as intended. It was therefore considered for inclusion in the HLT menu as a new AK8 tagging algorithm for $X \to b\bar{b}$ resonances. However, at the HLT stage, computing and input/output resources are highly constrained: the HLT must process the full collision rate in real time within fixed CPU, memory, and bandwidth budgets. Any added HLT path that increases per-event CPU or the overall menu rate can therefore exceed available resources. For this reason, quantifying any new path impact on timing and trigger

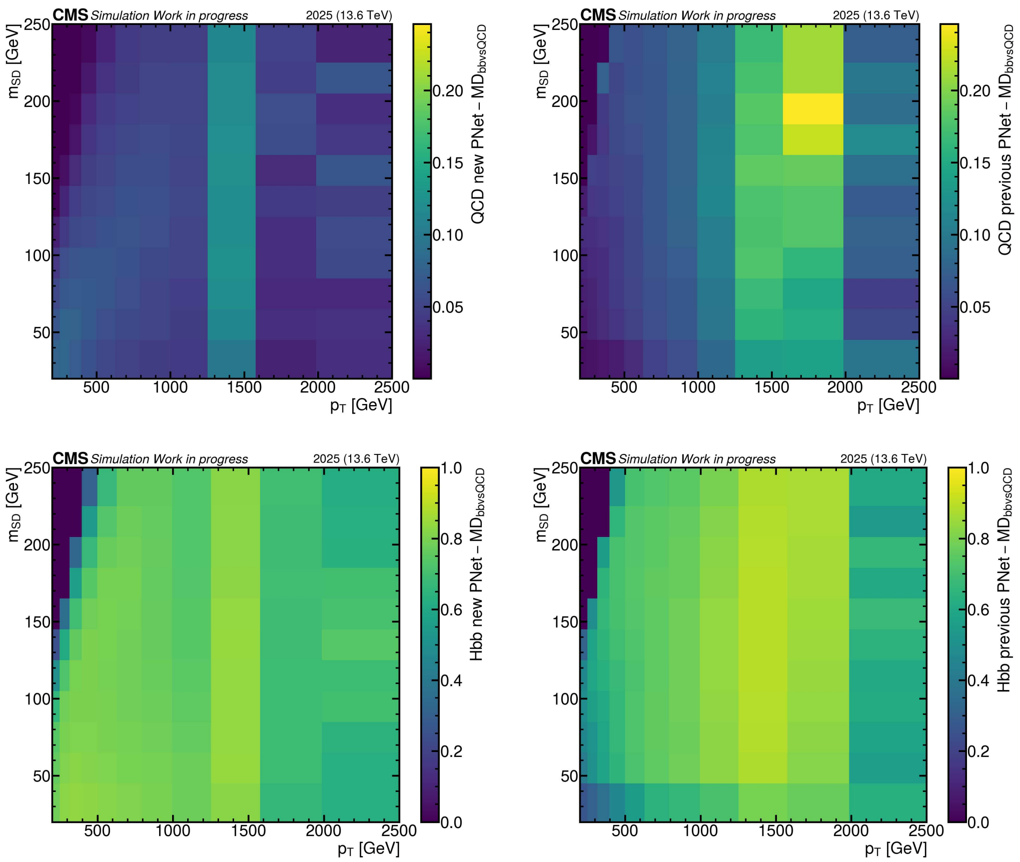

Figure 5.8: Average of the PNet-$MD_{bbvsQCD}$ score evaluated with the new model (left) and with the previous model (right) for QCD (above) and Hbb (below) jets in the (p_T, m_{SD}) distribution.

rates is essential before integration. The HLT_AK8PFJet230_SoftDropMass40_PNetV02BB0p14[5] path was added to a copy of the latest official HLT menu used for 2025 *pp* collisions. The updated menu was re-run on one of the last 2025 run, corresponding to a short high-rate data-taking period at close to the maximum PU. These data consist of events passing the L1 trigger (*i.e.* the input to the HLT) and correspond to an integrated luminosity of 0.53 fb^{-1}. Timing was measured as overall throughput (events processed per second). Rate was measured as the trigger firing frequency observed during the run; rates were normalized to a reference instantaneous luminosity ($2.1 \cdot 10^{34}\text{cm}^{-2}\text{s}^{-1}$). For comparison, the baseline measurements taken with the unchanged menu were used to report the effect of the new path. The throughput of the updated menu is 478.8 ± 0.9 evt/s while the baseline is 477.8 ± 1.1 evt/s. The pure rate (rate of events that pass only the HLT_AK8PFJet230_SoftDropMass40_PNetV02BB0p14 path) was of ≈6.3 Hz, out of a total path rate of ≈30 Hz. The fact that the measured throughputs are compatible within uncertainties, together with the low pure rate introduced by the path, indicates that the resources required to add

[5] As suggested by its name, the new path select events with an AK8 jet with p_T at least 230 GeV, $|\eta| < 2.5$, a soft drop mass of more than 40 GeV and a selection on the new score PNet-$MD_{bbvsQCD}$ greater than 0.14.

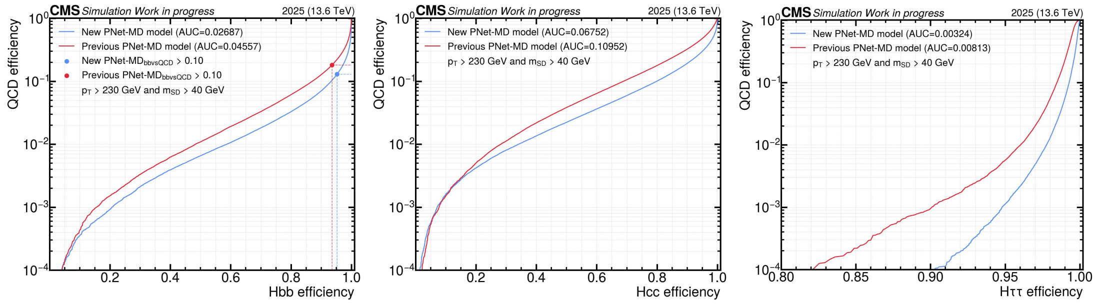

Figure 5.9: ROC curves obtained cutting on the PNet-MD discriminating scores evaluated with the previous model (red curve) and the new model (light blue curve) considering Hbb (left), Hcc (middle), and H$\tau\tau$ (right) jets as signal and QCD jets as background, for AK8 jets with $|\eta| < 2.4$, $p_T > 230$ GeV and $m_{SD} > 40$ GeV. The Area Under the Curve (AUC) is written for each model. Jets are weigthed as in the training. In the left plot, the points highlight the efficiency values for a PNet-MD$_{\mathrm{bbvsQCD}}$ above 0.10.

the path to the HLT menu do not impose a significant load on the HLT system.
The last check done was to evaluate the behavior of the new model in a specific physics case: the Higgs boson produced via gluon gluon fusion decaying into b-quarks. QCD was considered as background. The processes were analyzed in three kinematic regions:

- region 1: $p_T > 230$ GeV and $m_{SD} > 40$ GeV;
- region 2: $p_T > 450$ GeV and $m_{SD} > 40$ GeV;
- region 3: $p_T > 450$ GeV and $m_{SD} > 40$ GeV and $-6.0 < \rho < -2.1$, where $\rho = 2\ln(m_{SD}/p_T)$.

The first region is the kinematic region selected by the HLT_AK8PFJet230_SoftDropMass40_PNetV02BB0p14 path, the second one is the kinematic region where the Higgs bosons are high-boosted, the third one is a kinematic region used for single Higgs boson search documented in [122]. Table 5.3 lists the number of signal events (S), the number of background events (B), and the resulting significance calculated as $S/\sqrt{B}$. For signal events the entry counts only those where the AK8 jet matched to the Higgs boson passes the kinematic selection and score thresholds; for background events B counts occurrences in which at least one AK8 jet satisfies the same requirements. Event yields are normalized to an integrated luminosity of 30 fb^{-1}, the expected luminosity for 2026.

	region 1			region 2			region 3		
	B	S	S/$\sqrt{\mathrm{B}}$	B	S	S/$\sqrt{\mathrm{B}}$	B	S	S/$\sqrt{\mathrm{B}}$
Previous PNet-MD$_{\mathrm{bbvsQCD}} > 0.1$	$1.7{\cdot}10^7$	1115	0.27	$1.8{\cdot}10^6$	161	0.12	$1.5{\cdot}10^6$	155	0.13
New PNet-MD$_{\mathrm{bbvsQCD}} > 0.14$	$3.8{\cdot}10^7$	1073	0.17	$1.6{\cdot}10^6$	151	0.12	$1.5{\cdot}10^6$	147	0.12

Table 5.3: S and B yields and significance $S/\sqrt{\mathrm{B}}$ after selections. Signal entries require the AK8 jet matched to the Higgs to pass the kinematic and score thresholds; background entries require at least one AK8 jet to pass them. Yields are scaled to 30 fb^{-1} (expected luminosity for 2026).

The number of events in the three different kinematic regions show that, despite the new model include more QCD events in the first region, the significance and number of events obtained are

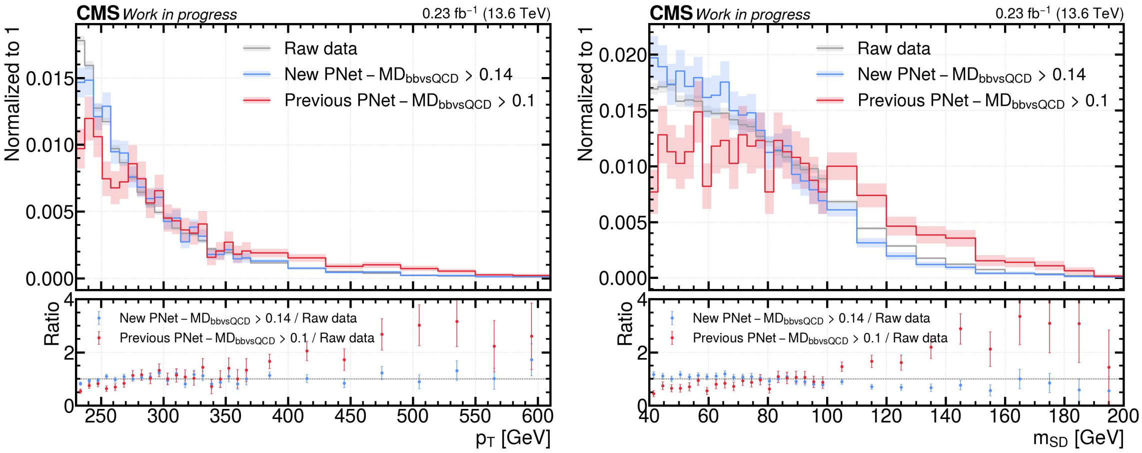

Figure 5.10: 2025 data p_T (left) and m_{SD} (right) distributions normalized to 1 for jets with $|\eta| < 2.4$, $p_T > 230$ GeV, $m_{SD} > 40$ GeV (raw data) without any score selection (grey lines), with a selection on the new PNet-MD$_{\mathrm{bbvsQCD}}$ greater than 0.14 (light blue lines) and on previous above 0.10 (red lines). The reported discriminator values were chosen to guarantee a signal efficiency on $H \to b\bar{b}$ of 94% for both models. Colored bars correspond to the statistical uncertainties. The lower plots show the ratio between the distributions, which have been selected with PNet-MD$_{\mathrm{bbvsQCD}}$, and the raw ones.

compatible with those of the current model in the typical regions of the analyses (assuming compatibility selection on HLT and offline values). The key advantage of the new model lies in its robustness and independence from the kinematic variables p_T and m_{SD}, which minimizes sculpting effects and reduces the risk of introducing biases in measurements. Further supporting this, the new score selects a larger number of jets at high transverse momentum ($p_T > 650$ GeV) or low m_{SD} ($m_{SD} < 100$ GeV) compared to the previous score, as illustrated in Figure 5.11, and demonstrates as the new model is able to target a broader phase-space.

5.5 2026 HLT menu

All tests returned positive results, and the HLT_AK8PFJet230_SoftDropMass40_PNetV02BB0p14 path was therefore added to the 2026 *pp* HLT menu. In addition to the nominal path, a backup strategy was required in case the CMS rate becomes too high and the path needs to be turned off. The decision was to add another path, hereinafter referred to as backup path, with the same kinematic thresholds but a higher score threshold. This increase of the threshold has been studied for the testing jets described in the section 5.3.4. By increasing the PNet-MD$_{\mathrm{bbvsQCD}}$ threshold to 0.26, the HLT_AK8PFJet230_SoftDropMass40_PNetV02BB0p26 path is expected to collect $\sim$34% fewer QCD jets and only $\sim$4% fewer Hbb jets. The total path rate is $\approx$18.7 Hz, of which the pure contribution from this path is $\approx$3 Hz, *i.e.* $\sim$50% lower than the nominal path. Based on these results, both paths were added to the 2026 *pp* HLT menu to collect events.
Figure 5.12 shows the rate of the two paths on the first 2026 data corresponding to an integrated luminosity of 0.53 fb^{-1} collected up to March 11$^{\text{th}}$, together with the corresponding pileup. The rate values are compatible with the estimated ones.

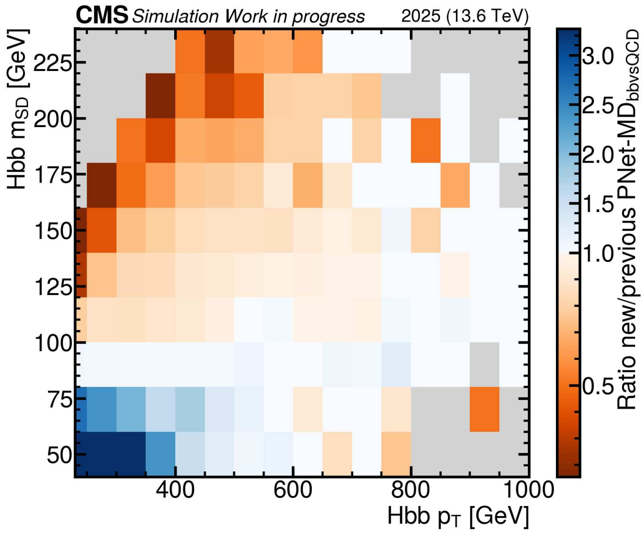

Figure 5.11: Ratio plot showing the yields variation of the AK8 $H \to b\bar{b}$ jet selected with the new trained model with respect to what is selected with the previous model. Both selections were applied to have the same signal efficiency as defined in section 5.3.4.

The performance of the proposed algorithm indicates that the adopted training strategy is intrinsically robust with respect to p_T and m_{SD} modeling. These results indicate that the adopted training approach can be established as a reproducible, well-documented benchmark and used as a development template for future algorithms targeted at the High-Luminosity LHC.

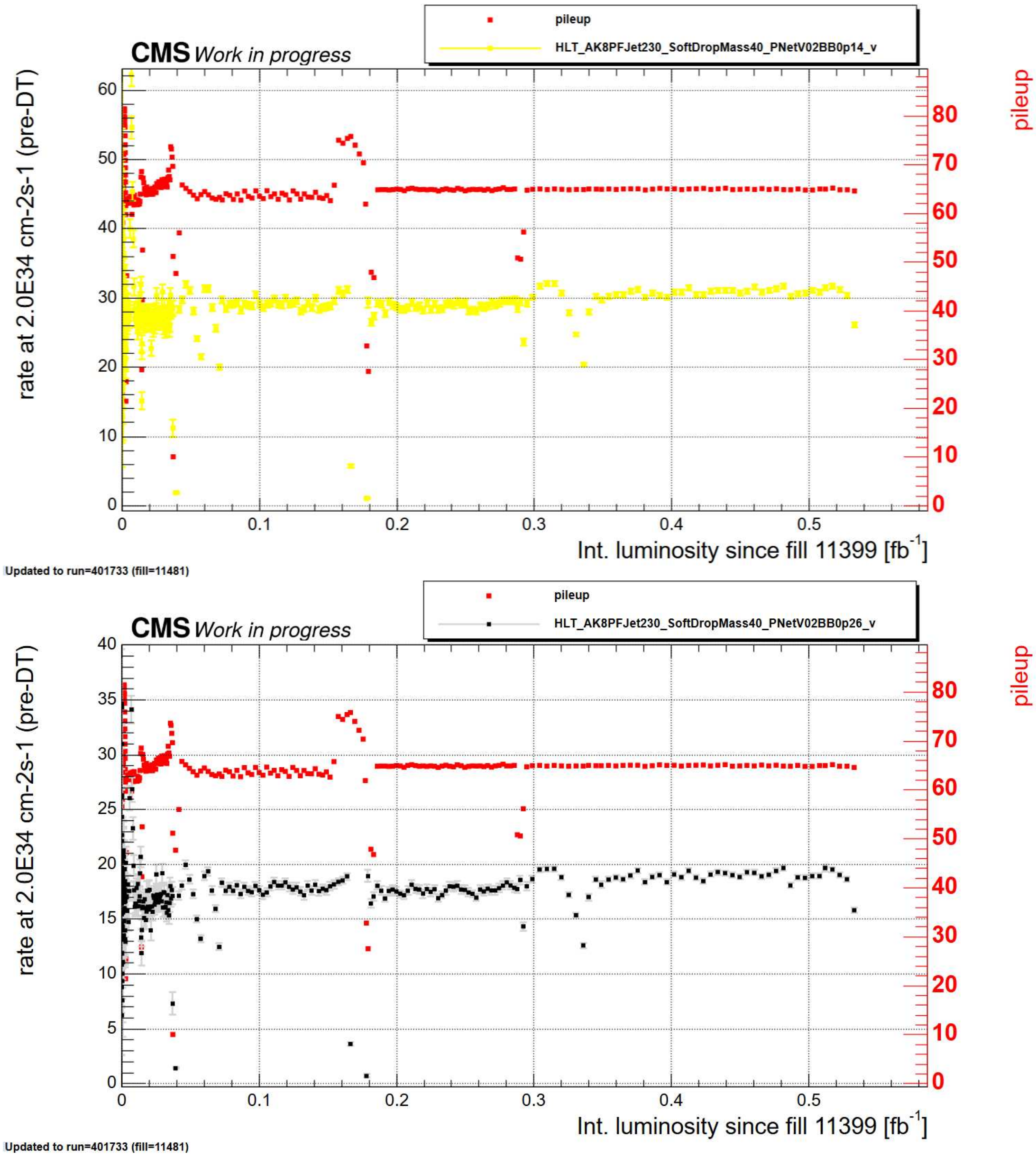

Figure 5.12: HLT_AK8PFJet230_SoftDropMass40_PNetV02BB0p14 (top) and HLT_AK8PFJet230_SoftDropMass40_PNetV02BB0p26 (bottom) path rate normalized to a reference instantaneous luminosity ($2 \cdot 10^{34}$cm^{-2-1}) and pileup (red points) as function of the 2026 integrated luminosity.

Conclusion

This thesis documents the ParticleNet algorithm, used by the CMS collaboration since 2019 for jet-flavor tagging, and its adaptation for AK8 jets, ParticleNet-MD (PNet-MD), which was developed to be largely independent of jet transverse momentum (p_T) and mass. Thanks to its graph-based architecture, PNet-MD has demonstrated superior performance relative to earlier algorithms and it has been continuously optimized by the collaboration for both *offline* (after the full even reconstruction reconstruction chain) and *online* (trigger) scenarios. Within this work I validated the most recent offline PNet-MD release using the first Run 3 data and trained a new online PNet-MD variant now incorporated into the 2026 data-taking.
The 2022 proton–proton collision data were used to ensure that the classifier response is well modeled by the Standard Model (SM) prediction and and that higher discriminator values enhance the boosted boson resonance. A boosted Z-boson enriched region was chosen for validation. Events with a leading-p_T AK8 jet (Z candidate) of $p_T > 450$ GeV and with a subleading-p_T AK8 jet (recoil jet) of $p_T > 200$ GeV were collected. A veto on electron, muon and AK4 b-tagged jet was applied to reduce $t\bar{t}$ contamination. With these selections, the SM processes involved are $Z \to q\bar{q}$, $W \to q\bar{q}'$ and QCD processes. The Z candidate soft-drop mass (m_{SD}) was required to lie in a m_{SD} window centered around Z boson mass ($m_{\mathrm{SD}} \in [76, 126]$ GeV). The Z-candidate PNet-MD score was used to partition events into five score regions, guaranteeing the same amount of $Z \to b\bar{b}$ jets in each score region, by means of the PNet-MD$_{\mathrm{bbvsQCD}}$ discriminant

$$\text{PNet-MD}_{\mathrm{bbvsQCD}} = \frac{P(Xbb)}{P(Xbb) + P(QCD2hf) + P(QCD1hf) + P(QCD0hf)}.$$

The lowest score region (PNet-MD$_{\mathrm{bbvsQCD}}$ <0.641) was not involved in the validation since data in this region were overpopulated by QCD contribution. The dominant QCD background was estimated from data by fitting independently in each score regions the sideband $m_{SD} \in [38, 70] \cup [126, 198]$GeV with Chebyshev polynomials up to the 6$^{\mathrm{th}}$ order, determining the optimal order using the Fisher F-test at 5% significance level. To account for statistical fluctuations and fit-range dependence, other eight alternative sidebands were fitted and the QCD contribution in the signal region was taken as the average of the nine fits. A binned likelihood fit as a function of m_{SD} was performed assigning a different normalization factor for $Z \to q\bar{q}$ and $W \to q\bar{q}'$ for each of score regions, to obtain a more accurate estimate of the involved processes. Nuisance parameters for MC statistical uncertainties, jet energy scale variations, the integrated luminosity uncertainty and uncertainties on the QCD estimate were considered; all uncertainties except luminosity were treated as uncorrelated across score regions. In each score region the $Z \to q\bar{q}$ normalization factors were found to be consistent with unity within uncertainties. The post-fit m_{SD} distributions show a progressively more pronounced Z boson peak as the discriminator value increases, illustrating the ability of PNet-MD to separate hadronic resonances from QCD jets. Moreover, the post-fit PNet-MD$_{\mathrm{bbvsQCD}}$ distribution shows good agreement between data and SM expected. In conclusion, the validation demonstrates that

PNet-MD provides a robust and well-modelled discriminant for separating hadronically decaying boosted resonances from QCD background in data. The achieved data-to-prediction agreement supports the use of PNet-MD in future CMS searches and measurements that exploit boosted hadronic final states. This successful validation therefore lays the way for measurements of hadronically decaying Z bosons using PNet-MD. The validation study is reported in CMS Detector Performance Note [97].

The online PNet-MD model (default model) showed a sculpting effect on the data distribution at low p_T and m_{SD}. For this reason, a new version of the model was trained in order to reduce the dependence on p_T and m_{SD} while maintaining the same tagging performance. The new model was trained using graviton-like resonance (X) decaying to two Higgs bosons, with $m_X \in [500, 6000]$ GeV and $m_H \in [20, 250]$ GeV in order to cover a very large phase space. The Higgs boson was simulated with the same probability of decaying into $b\bar{b}$, $c\bar{c}$, light quarks, gg, $e\bar{e}$, $\mu\bar{\mu}$ and $\tau\bar{\tau}$ to ensure good statistics for all the network outputs considered. QCD was assumed as background. During the testing and training phases, the jets were weighted to ensure that the (p_T, m_{SD}) distribution, corresponding to each of the network output classes, was as uniform as possible. Due to the unbalance in statistics, QCD jets for $p_T < 791$ GeV and signal jets for $p_T >$1987 GeV were randomly discarded to have the same number of QCD jets and signal jets. Furthermore, only jets with $p_T >$315 GeV were used during training given the different m_{SD} distribution at low p_T. The default model and the new model were compared for AK8s with $p_T > 230$ GeV, pseudorapidity $|\eta| < 2.4$ and $m_{SD} > 40$ GeV. The new model PNet-MD$_{\mathrm{bbvsQCD}}$ shows lower QCD efficiency with respect to the default one, once measured at the same $H \to b\bar{b}$ efficiency. In particular, the threshold of the new PNet-MD$_{\mathrm{bbvsQCD}}$ discriminator, corresponding to a $H \to b\bar{b}$ efficiency of 94%, reduces the QCD efficiency by approximately 40%. Given the improved performance, a new trigger selection that requires an AK8 jet with $p_T > 230$ GeV, $m_{SD} > 40$ GeV, $|\eta| < 2.5$ and PNet-MD$_{\mathrm{bbvsQCD}} > 0.14$ was proposed to be included into the 2026 CMS data-taking selections list (menu). The menu performance with the new selection was validated on one of the last 2025 run. Compared to the baseline menu, the additional requirements of the new menu in terms of timing (number of events processed per second) and rate (number of events saved per second) were negligible. Based on these results, the new selection has been included in the 2026 data-taking menu. A backup selection has been also proposed for the menu, in case the CMS rate of the proposed selection becomes too high. The backup selection has the same kinematic thresholds but with a higher score threshold (PNet-MD$_{\mathrm{bbvsQCD}} > 0.26$), which allows to reduce by half the expected rate. Under high-rate conditions, only the backup selection will be used. Preliminary results on the first 2026 data show the expected rate performance for both selections. Given the achieved level of independence, this training strategy can be considered a reference framework for future models designed to be robust with respect to p_T and m_{SD}.

Appendix A

QCD fits

As part of the validation of PNet-MD$_{\mathrm{bbvsQCD}}$, described in Chapter 4, the QCD contribution in each PNet-MD$_{\mathrm{bbvsQCD}}$ score region is obtained by fitting the m_{SD} sidebands with analytic functions (Chebyshev polynomials up to 6^{th} order) and extrapolating these fits into the nominal signal region ($m_{\mathrm{SD}} \in [76, 126]$ GeV). The nine alternative sideband regions considered are the ones listed in the main text (reproduced here for convenience):

- $m_{\mathrm{SD}} \in [38, 70] \ \cup \ [126, 198]$ GeV
- $m_{\mathrm{SD}} \in [38, 62] \ \cup \ [134, 198]$ GeV
- $m_{\mathrm{SD}} \in [38, 78] \ \cup \ [118, 198]$ GeV
- $m_{\mathrm{SD}} \in [38, 54] \ \cup \ [142, 198]$ GeV
- $m_{\mathrm{SD}} \in [38, 86] \ \cup \ [110, 198]$ GeV
- $m_{\mathrm{SD}} \in [38, 62] \ \cup \ [126, 198]$ GeV
- $m_{\mathrm{SD}} \in [38, 78] \ \cup \ [126, 198]$ GeV
- $m_{\mathrm{SD}} \in [38, 70] \ \cup \ [134, 198]$ GeV
- $m_{\mathrm{SD}} \in [38, 70] \ \cup \ [118, 198]$ GeV

m_{SD} was rescaled as

$$x = \frac{m_{\mathrm{SD}}}{200 \text{ GeV}},$$

so that the Chebyshev polynomials are evaluated on the usual domain $[-1; 1]$. For each score region:

1. the sidebands are fitted with Chebyshev polynomials ranging from zeroth to 6^{th} order.
2. Nested models (order a vs order $b > a$) are compared with the Fisher F-test at the 5% significance level to select the simplest model that adequately describes the data.
3. The nine selected analytic functions are used to produce nine discrete predictions in the nominal signal region. The final QCD prediction is the per-bin arithmetic mean of these nine predictions.

This appendix collects plots corresponding to all the sideband fits not reported in Chapter 4.

A.1 Plots

Figure A.1 shows all the polynomials fitting the $m_{SD} \in [38, 62] \cup [134, 198]$GeV sideband in the four score region. Based on the Fisher F-test results at a 5% significance level, the QCD in the 4^{th} highest score region (top-left) was modeled using a 5^{th} order polynomial, the QCD in 3^{rd} highest score region (top-right) and in the 2^{nd} highest score region (bottom-left) was modeled using a 4^{th} order polynomial and the QCD in the highest score region (bottom-right) was modeled using a 3^{rd} order polynomial.

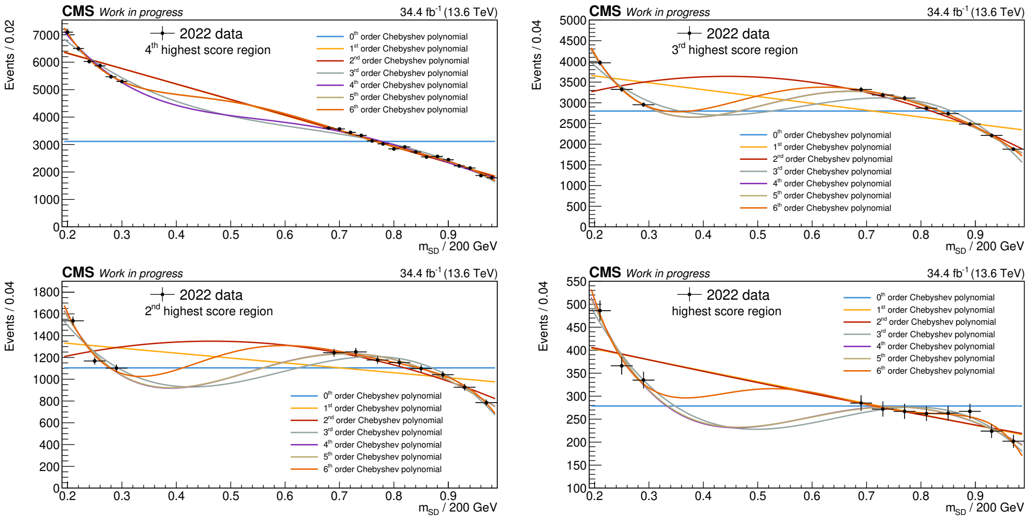

Figure A.1: 2022 data for $m_{SD} \in [38, 62] \cup [134, 198]$GeV in the four score regions fitted by Chebyshev polynomials ranging from order 0 to order 6. The order of the polynomial that most closely approximates the data trend was chosen based on the Fisher F-test results at a 5% significance level.

Figure A.2 shows all the polynomials fitting the $m_{SD} \in [38, 78] \cup [118, 198]$GeV sideband in the four score region. Based on the Fisher F-test results at a 5% significance level, the QCD in the 4$^{\text{th}}$ highest score region (top-left) and in 3$^{\text{rd}}$ highest score region (top-right) was modeled using a 5$^{\text{th}}$ order polynomial, the QCD in the 2$^{\text{nd}}$ highest score region (bottom-left) and in the highest score region (bottom-right) was modeled using a 4$^{\text{th}}$ order polynomial.

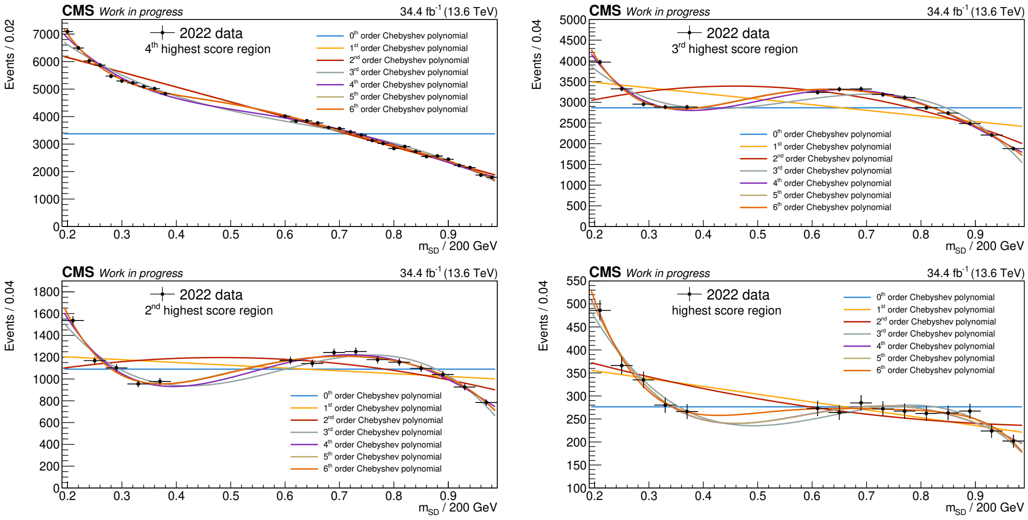

Figure A.2: 2022 data for $m_{SD} \in [38, 78] \cup [118, 198]$GeV in the four score regions fitted by Chebyshev polynomials ranging from order 0 to order 6. The order of the polynomial that most closely approximates the data trend was chosen based on the Fisher F-test results at a 5% significance level.

Figure A.3 shows all the polynomials fitting the $m_{SD} \in [38, 54] \cup [142, 198]$GeV sideband in the four score region. Based on the Fisher F-test results at a 5% significance level, the QCD in the 4^{th} highest score region (top-left) and in 3^{rd} highest score region (top-right) was modeled using a 4^{th} order polynomial, the QCD in the 2^{nd} highest score region (bottom-left) was modeled using a 5^{th} order polynomial and the QCD in the highest score region (bottom-right) was modeled using a 3^{rd} order polynomial.

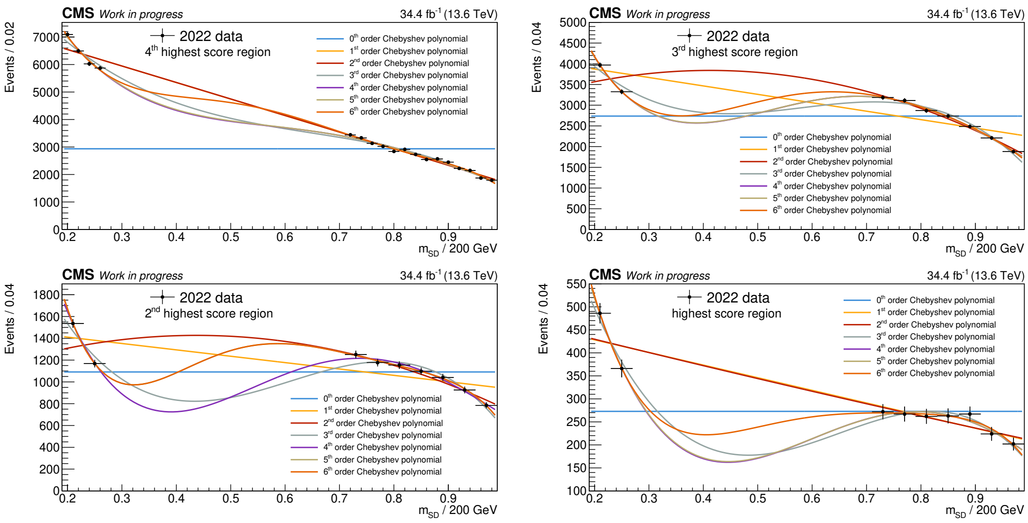

Figure A.3: 2022 data for $m_{SD} \in [38, 54] \cup [142, 198]$GeV in the four score regions fitted by Chebyshev polynomials ranging from order 0 to order 6. The order of the polynomial that most closely approximates the data trend was chosen based on the Fisher F-test results at a 5% significance level.

Figure A.4 shows all the polynomials fitting the $m_{SD} \in [38, 86] \cup [110, 198]$GeV sideband in the four score region. Based on the Fisher F-test results at a 5% significance level, the QCD in all score regions was modeled using a 5^{th} order polynomial.

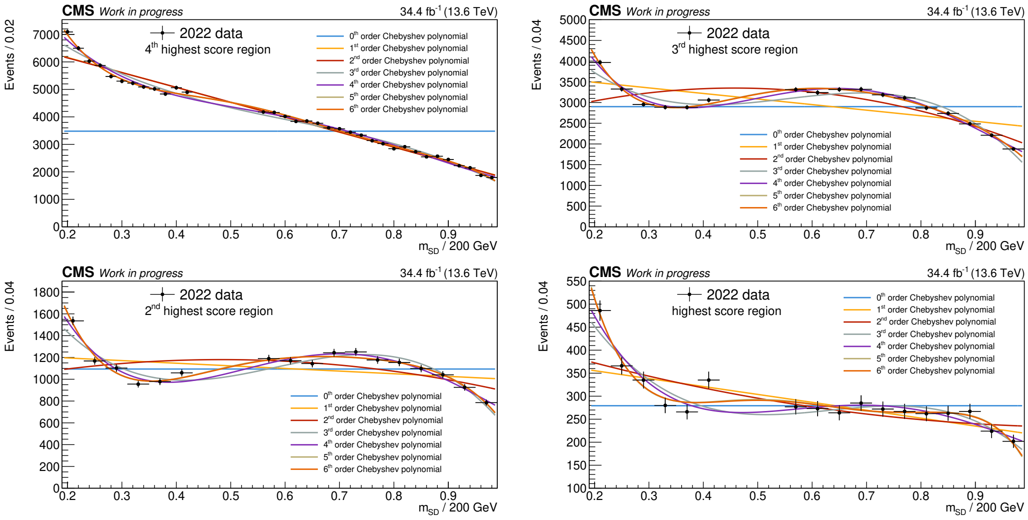

Figure A.4: 2022 data for $m_{SD} \in [38, 86] \cup [110, 198]$GeV in the four score regions fitted by Chebyshev polynomials ranging from order 0 to order 6. The order of the polynomial that most closely approximates the data trend was chosen based on the Fisher F-test results at a 5% significance level.

Figure A.5 shows all the polynomials fitting the $m_{SD} \in [38, 62] \cup [126, 198]$GeV sideband in the four score region. Based on the Fisher F-test results at a 5% significance level, the QCD in the 4$^{\text{th}}$ highest score region (top-left) was modeled using a 5$^{\text{th}}$ order polynomial, the QCD in 3$^{\text{rd}}$ highest score region (top-right) and in the 2$^{\text{nd}}$ highest score region (bottom-left) was modeled using a 4$^{\text{th}}$ order polynomial and the QCD in the highest score region (bottom-right) was modeled using a 3$^{\text{rd}}$ order polynomial.

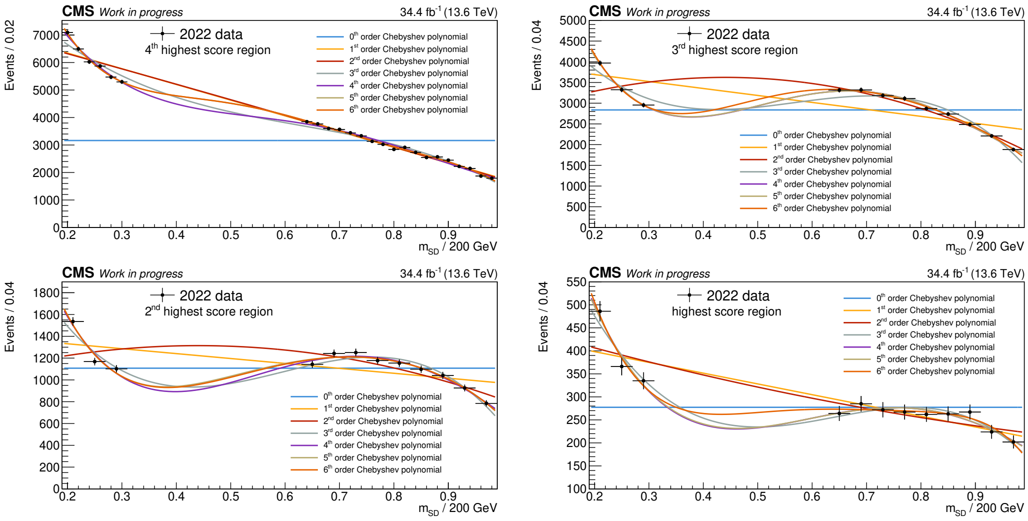

Figure A.5: 2022 data for $m_{SD} \in [38, 62] \cup [126, 198]$GeV in the four score regions fitted by Chebyshev polynomials ranging from order 0 to order 6. The order of the polynomial that most closely approximates the data trend was chosen based on the Fisher F-test results at a 5% significance level.

Figure A.6 shows all the polynomials fitting the $m_{SD} \in [38, 78] \cup [126, 198]$GeV sideband in the four score region. Based on the Fisher F-test results at a 5% significance level, the QCD in the 4^{th} highest score region (top-left) and in 3^{rd} highest score region (top-right) was modeled using a 5^{th} order polynomial, the QCD in the 2^{nd} highest score region (bottom-left) was modeled using a 4^{th} order polynomial and the QCD in the highest score region (bottom-right) was modeled using a 3^{rd} order polynomial.

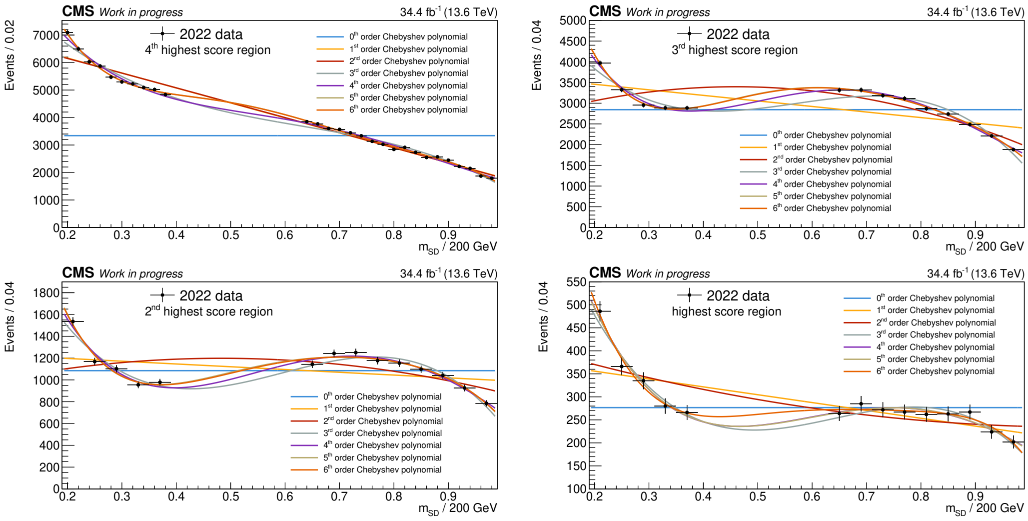

Figure A.6: 2022 data for $m_{SD} \in [38, 78] \cup [126, 198]$GeV in the four score regions fitted by Chebyshev polynomials ranging from order 0 to order 6. The order of the polynomial that most closely approximates the data trend was chosen based on the Fisher F-test results at a 5% significance level.

Figure A.7 shows all the polynomials fitting the $m_{SD} \in [38, 70] \cup [134, 198]$GeV sideband in the four score region. Based on the Fisher F-test results at a 5% significance level, the QCD in the 4$^{\text{th}}$ highest score region (top-left) was modeled using a 5$^{\text{th}}$ order polynomial, the QCD in 3$^{\text{rd}}$ highest score region (top-right) and in the 2$^{\text{nd}}$ highest score region (bottom-left) was modeled using a 4$^{\text{th}}$ order polynomial and the QCD in the highest score region (bottom-right) was modeled using a 3$^{\text{rd}}$ order polynomial.

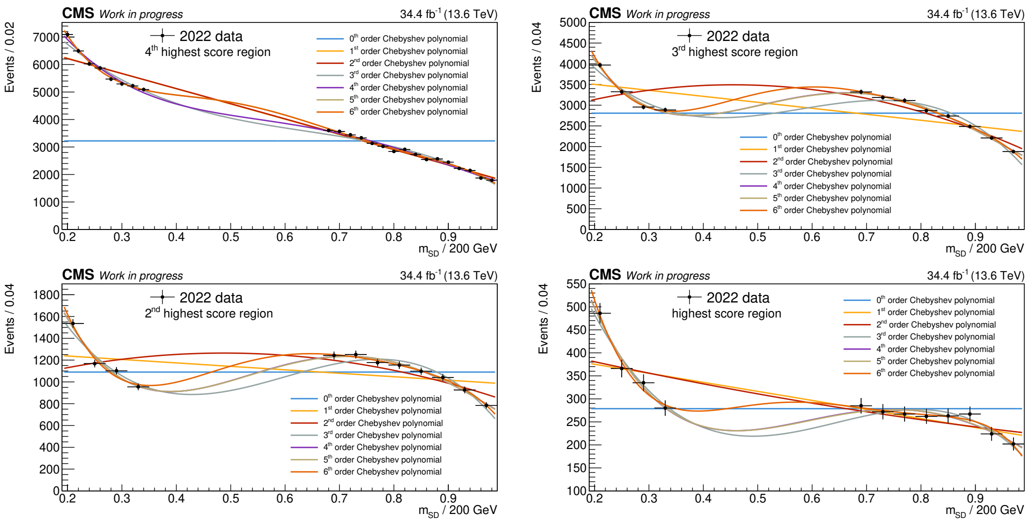

Figure A.7: 2022 data for $m_{SD} \in [38, 70] \cup [134, 198]$GeV in the four score regions fitted by Chebyshev polynomials ranging from order 0 to order 6. The order of the polynomial that most closely approximates the data trend was chosen based on the Fisher F-test results at a 5% significance level.

Figure A.8 shows all the polynomials fitting the $m_{SD} \in [38, 70] \cup [118, 198]$GeV sideband in the four score region. Based on the Fisher F-test results at a 5% significance level, the QCD in the 4$^{\mathrm{th}}$ highest score region (top-left) was modeled using a 5$^{\mathrm{th}}$ order polynomial, the QCD in 3$^{\mathrm{rd}}$ highest score region (top-right), in the 2$^{\mathrm{nd}}$ highest score region (bottom-left) and in the highest score region (bottom-right) was modeled using a 4$^{\mathrm{th}}$ order polynomial.

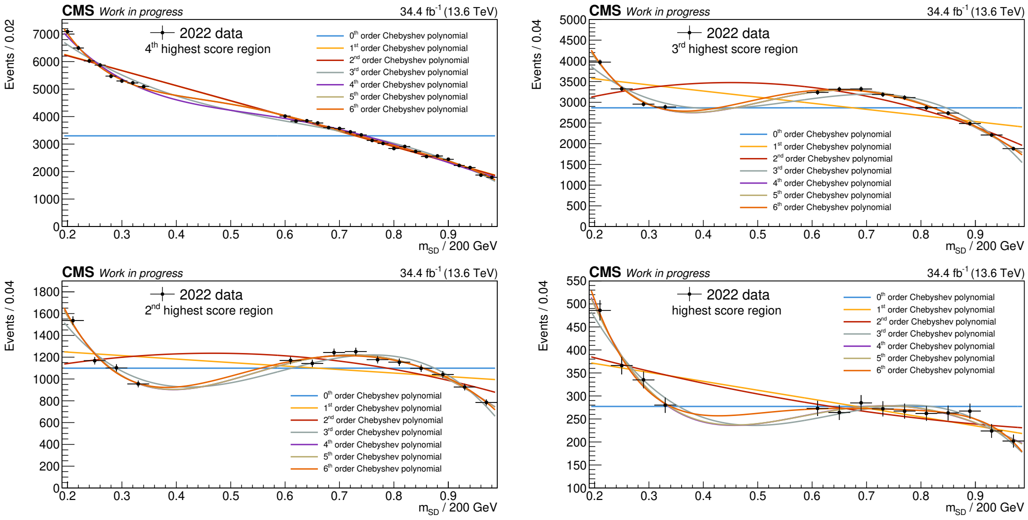

Figure A.8: 2022 data for $m_{SD} \in [38, 70] \cup [118, 198]$GeV in the four score regions fitted by Chebyshev polynomials ranging from order 0 to order 6. The order of the polynomial that most closely approximates the data trend was chosen based on the Fisher F-test results at a 5% significance level.

Appendix B

Modeling of Uncertainties in the Profile Likelihood

As part of the PNet-MD$_{\mathrm{bbvsQCD}}$ validation (see Chapter 4), a binned likelihood fit to m_{SD} was performed to better estimate the processes involved: the MC simulated $Z \to q\bar{q}$, $W \to q\bar{q}'$ and the data-driven QCD. The fit model includes nuisance parameters (θ) that account for both statistical and systematic uncertainties. The nuisance parameters used in the fit are summarized below.

- MC statistical uncertainty: per-bin statistical uncertainty on simulated templates (not the data-driven QCD);
- lumi: integrated luminosity normalization uncertainty applied to all simulated samples;
- JES uncertainties: 27 independent jet energy scale sources propagated to the two AK8 soft-drop subjets of $Z \to q\bar{q}$ and $W \to q\bar{q}'$ samples;
- QCD uncertainties: two uncertainties on QCD estimate due to the m_{SD} distribution fit functions and to the use of the nine mass windows.

The "lumi" systematic is implemented in the likelihood fit by introducing a nuisance parameter θ_k that is constrained by a unit Gaussian prior,

$$\theta_k \sim \mathcal{N}(0,1), \qquad \pi_k(\theta_k) = \frac{1}{\sqrt{2\pi}}\, e^{-\theta_k^2/2},$$

where $\mathcal{N}(0,1)$ is a gaussian with mean 0 and standard deviation 1, with $\pi_k(\theta_k)$ entering in the likelihood. The effect of this nuisance on the expected yield in score region c and distribution bin b ($\mu_{b,c}$) is multiplicative:

$$\mu_{b,c} \longrightarrow \mu_{b,c} \times f_k(\theta_k) \qquad \text{with} \qquad f_k(\theta_k) = e^{\sigma_k \theta_k},$$

where

$$\sigma_k = \ln(1.014).$$

For QCD and JES uncertainties two shifted histograms ("Up" and "Down") were provided, for each affected process and channel, in addition to the nominal histogram. Denote the nominal bin contents by $h_b^{(0)}$ and the shifted ones by $h_b^{(+1)}$ and $h_b^{(-1)}$. Define the corresponding bin fractions

$$f_b^{(0)} = \frac{h_b^{(0)}}{\sum_{b'} h_{b'}^{(0)}}, \qquad f_b^{(\pm 1)} = \frac{h_b^{(\pm 1)}}{\sum_{b'} h_{b'}^{(\pm 1)}}.$$

Set $\delta_b^{\pm} = f_b^{(\pm 1)} - f_b^{(0)}$. For the nuisance ν associated to this uncertainty (with $\nu \sim \mathcal{N}(0,1)$), the interpolated bin fraction for $|\nu| \leq 1$ is

$$f_b(\nu) = f_b^{(0)} + \frac{1}{2}\Big[(\delta_b^{+} - \delta_b^{-})\,\nu + \frac{1}{8}(\delta_b^{+} + \delta_b^{-})\big(3\nu^6 - 10\nu^4 + 15\nu^2\big)\Big],$$

with linear extrapolation for $|\nu| > 1$.

$$f_b(\nu) = \begin{cases} f_b^{(+1)} + (\,\nu - 1\,)\,\dfrac{df_b}{d\nu}\Big|_{\nu=+1} = f_b^{(0)} + \nu\,\delta_b^{+}, & \nu > 1, \\ f_b^{(-1)} + (\,\nu + 1\,)\,\dfrac{df_b}{d\nu}\Big|_{\nu=-1} = f_b^{(0)} + \nu\,\big(-\delta_b^{-}\big), & \nu < -1. \end{cases}$$

The total normalization is interpolated multiplicatively so that the summed yield at $\nu = \pm 1$ matches the provided shifted histograms. Let the total nominal normalization be $N^{(0)} = \sum_b h_b^{(0)}$ and the shifted normalizations be $N^{(\pm 1)} = \sum_b h_b^{(\pm 1)}$. Define

$$\kappa^{+} = \frac{N^{(+1)}}{N^{(0)}}, \qquad \kappa^{-} = \frac{N^{(-1)}}{N^{(0)}}.$$

The interpolated normalization factor for the nuisance ν is defined piecewise for $|\nu| \leq 1$ as follows:

$$N(\nu) = N^{(0)} \begin{cases} \exp(\nu \ln \kappa^{+}), & 0 \leq \nu \leq 1, \\ \exp(-\nu \ln \kappa^{-}), & -1 \leq \nu < 0. \end{cases}$$

For $|\nu|>1$, the interpolation is:

$$N(\nu) = \begin{cases} N^{(0)}(\kappa^{+})^{\nu}, & \nu > 1, \\ N^{(0)}(\kappa^{-})^{-\nu}, & \nu < -1. \end{cases}$$

The expected bin content is then

$$\mu_b(\nu) = N(\nu)\, f_b(\nu),$$

and $\mu_b(\nu)$ replaces the corresponding quantity in the Poisson term of the likelihood.
The nuisance parameters denoted as `prop_binScore_c_binb`, where c labels the score region and b the bin index within that region, are introduced to account for the statistical uncertainty of MC templates in individual bins. Only MC-based processes contribute to these uncertainties; data-driven backgrounds, such as QCD in this validation, do not generate corresponding nuisances. The expected yield in bin b of region c is modified as

$$\mu_{b,c}(r_{Zc}, r_{Wc}, , \theta_{\neq \nu_{b,c}}, \nu_{b,c}) = \sum_{s \in \mathrm{MC}} r_s \cdot \left[s_{s,b,c}(\theta_{\neq \nu_{b,c}}) + \nu_{b,c}\, \sigma_{b,c}^{\mathrm{MC}}\right] + s_{\mathrm{QCD}b,c}(\theta_{\neq \nu_{b,c}})\,, \tag{B.1}$$

where $\nu =$ `prop_binScore_c_binb` $\sim \mathcal{N}(0,1)$ is a standard Gaussian-constrained nuisance, and

$$\sigma_{b,c}^{\mathrm{MC}} = \sqrt{\sum_{s \in \mathrm{MC}} (\sigma_{s,b,c})^2}$$

represents the combined statistical uncertainty of all MC processes in that bin. The Poisson term for the bin is evaluated using $\mu_{b,c}(\nu)$.

Appendix C

PNet-MD inputs

In this appendix, the distributions of the PNet-MD input variables used by the HLT model that were not displayed in Subsection 5.3.1 are presented. These plots complement those shown in the previous chapter and are included here for completeness. The distributions shown in this appendix were obtained using the full available statistics of the QCD and XtoHH samples, selecting AK8 jets that pass the following requirements:

- jet $|\eta| < 2.4$;
- jet $p_T > 200$ GeV;
- The reconstructed jet m_{SD} and the invariant mass of the nearest generator-level jet (with $\Delta R < 0.4$ from the reconstructed jet) are required to be greater than 20 GeV.

For the QCD sample, all jets are considered, whereas for the XtoHH sample only jets matched to an $H \rightarrow b\bar{b}$ decay are retained.

C.1 PF candidates plots

In this section the features of the PF objects, with $p_T > 0.1$ GeV, that constitute the AK8 jet, are shown. Their definition can be found in Subsection 5.3.1. Figure C.1 shows the unweighted jet_pfcand_energy_log, jet_pfcand_dphi and jet_pfcand_deta distributions. Figure C.2 shows the unweighted jet_pfcand_eta, jet_pfcand_charge and jet_pfcand_frompv distributions. Figure C.3 shows the unweighted jet_pfcand_nlostinnerhits, jet_pfcand_dz and jet_pfcand_dxy distributions. Figure C.4 shows the unweighted jet_pfcand_etarel, jet_pfcand_pperp_ratio and jet_pfcand_ppara_ratio distributions. Figure C.5 shows the unweighted jet_pfcand_npixhits, jet_pfcand_nstriphits and jet_pfcand_track_chi2 distributions. Figure C.6 shows the unweighted jet_pfcand_track_qual, jet_pfcand_dzsig and jet_pfcand_dxysig distributions. Figure C.7 shows the unweighted jet_pfcand_trackjet_d3d, jet_pfcand_trackjet_dist and jet_pfcand_trackjet_decayL distributions. jet_pfcand_energy_log, jet_pfcand_charge, jet_pfcand_nlostinnerhits, jet_pfcand_dxy, jet_pfcand_track_qual, jet_pfcand_dxysig and jet_pfcand_trackjet_decayL are the PNet-MD input variables with the with greater differences in shape between $H \rightarrow b\bar{b}$ and QCD.

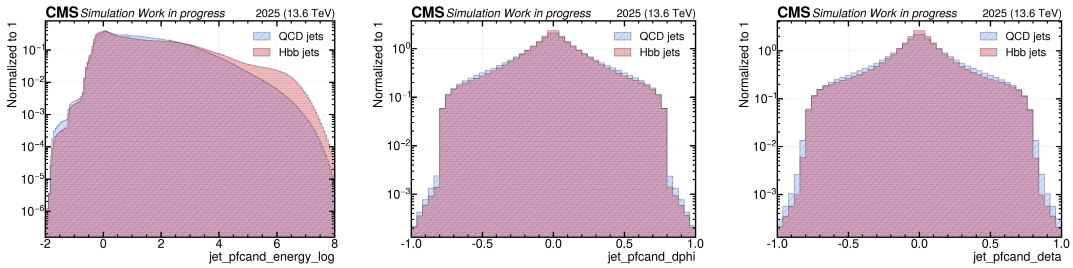

Figure C.1: Normalized to 1 unweighted jet__pfcand__energy__log (left), jet__pfcand__dphi (middle) and jet__pfcand__deta (right) distributions for XtoHH AK8 jets labeled as $H \to b\bar{b}$ (red histograms) and QCD AK8 jets (light blue histograms).

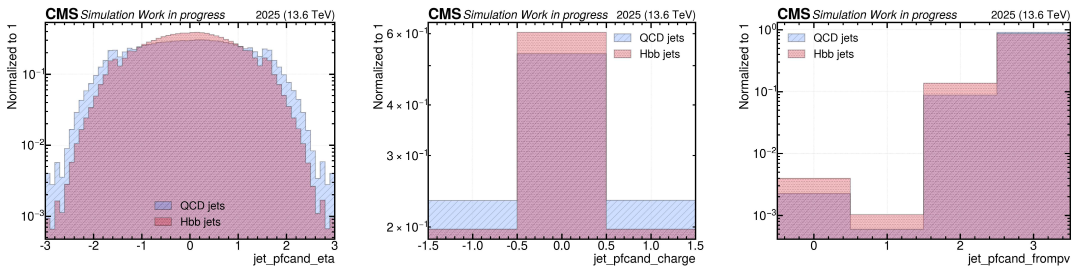

Figure C.2: Normalized to 1 unweighted jet__pfcand__eta (left), jet__pfcand__charge (middle) and jet__pfcand__frompv (right) distributions for XtoHH AK8 jets labeled as $H \to b\bar{b}$ (red histograms) and QCD AK8 jets (light blue histograms).

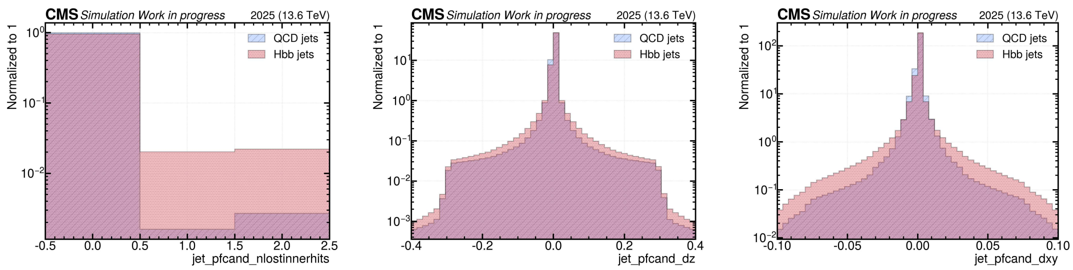

Figure C.3: Normalized to 1 unweighted jet__pfcand__nlostinnerhits (left), jet__pfcand__dz (middle) and jet__pfcand__dxy (right) distributions for XtoHH AK8 jets labeled as $H \to b\bar{b}$ (red histograms) and QCD AK8 jets (light blue histograms).

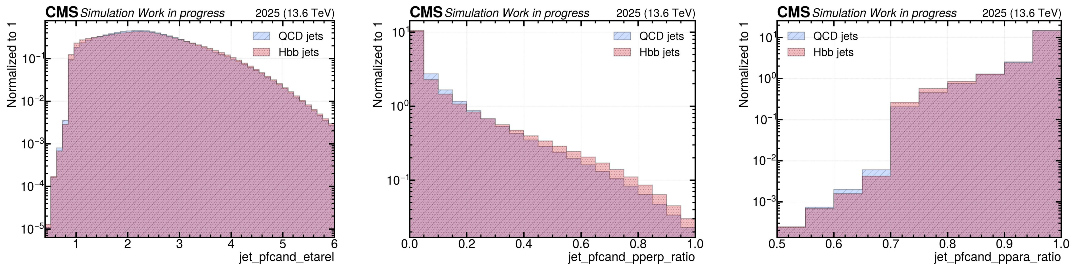

Figure C.4: Normalized to 1 unweighted jet__pfcand__etarel (left), jet__pfcand__pperp__ratio (middle) and jet__pfcand__ppara__ratio (right) distributions for XtoHH AK8 jets labeled as $H \to b\bar{b}$ (red histograms) and QCD AK8 jets (light blue histograms).

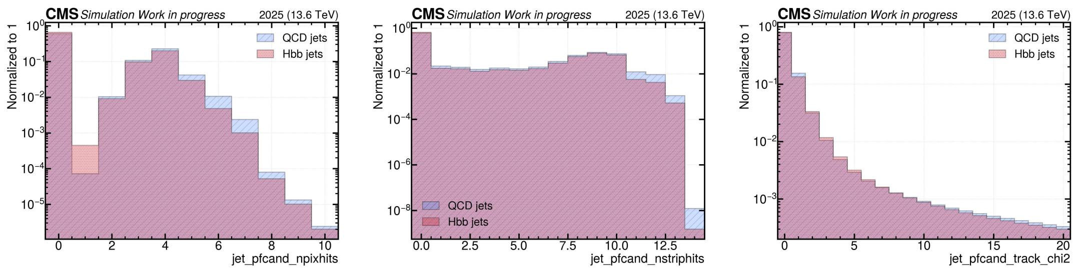

Figure C.5: Normalized to 1 unweighted jet__pfcand__npixhits (left), jet__pfcand__nstriphits (middle) and jet__pfcand__track__chi2 (right) distributions for XtoHH AK8 jets labeled as $H \to b\bar{b}$ (red histograms) and QCD AK8 jets (light blue histograms).

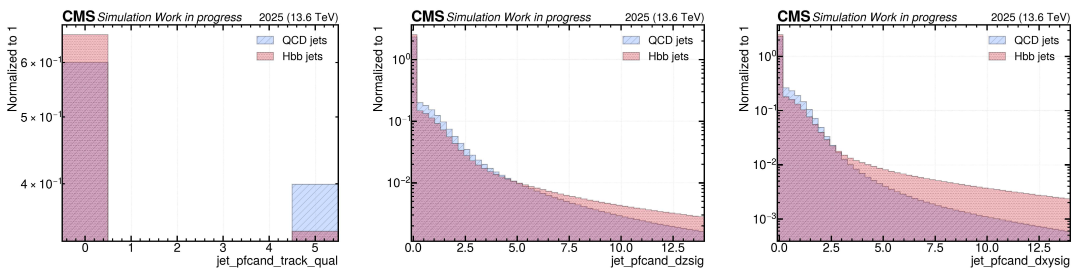

Figure C.6: Normalized to 1 unweighted jet__pfcand__track__qual (left), jet__pfcand__dzsig (middle) and jet__pfcand__dxysig (right) distributions for XtoHH AK8 jets labeled as $H \to b\bar{b}$ (red histograms) and QCD AK8 jets (light blue histograms).

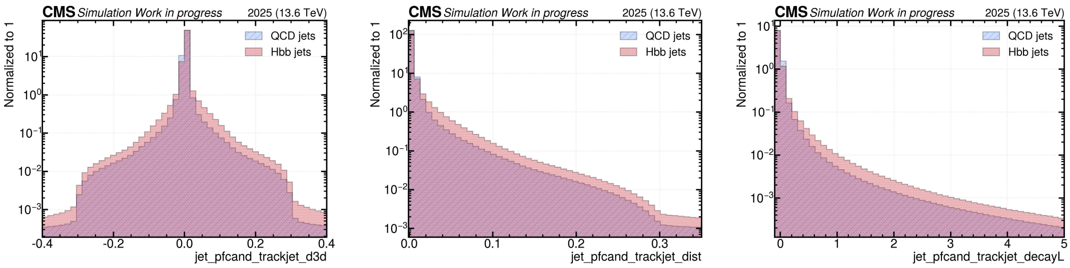

Figure C.7: Normalized to 1 unweighted jet__pfcand__trackjet__d3d (left), jet__pfcand__trackjet__dist (middle) and jet__pfcand__trackjet__decayL (right) distributions for XtoHH AK8 jets labeled as $H \to b\bar{b}$ (red histograms) and QCD AK8 jets (light blue histograms).

C.2 SVs plots

In this section the features of the SVs reconstructed with the IVF algorithm and within a distance $\Delta R < 0.8$ with respect to the AK8 jet axis, are shown. Their definition can be found in Subsection 5.3.1. Figure C.8 shows the unweighted jet_sv_pt_log, jet_sv_mass, jet_sv_dphi and jet_sv_ntrack distributions. Figure C.9 shows the unweighted jet_sv_chi2, jet_sv_dxy, jet_sv_d3d and jet_sv_d3dsig distributions. In all distributions, there is a difference in shape between $H \to b\bar{b}$ and QCD.

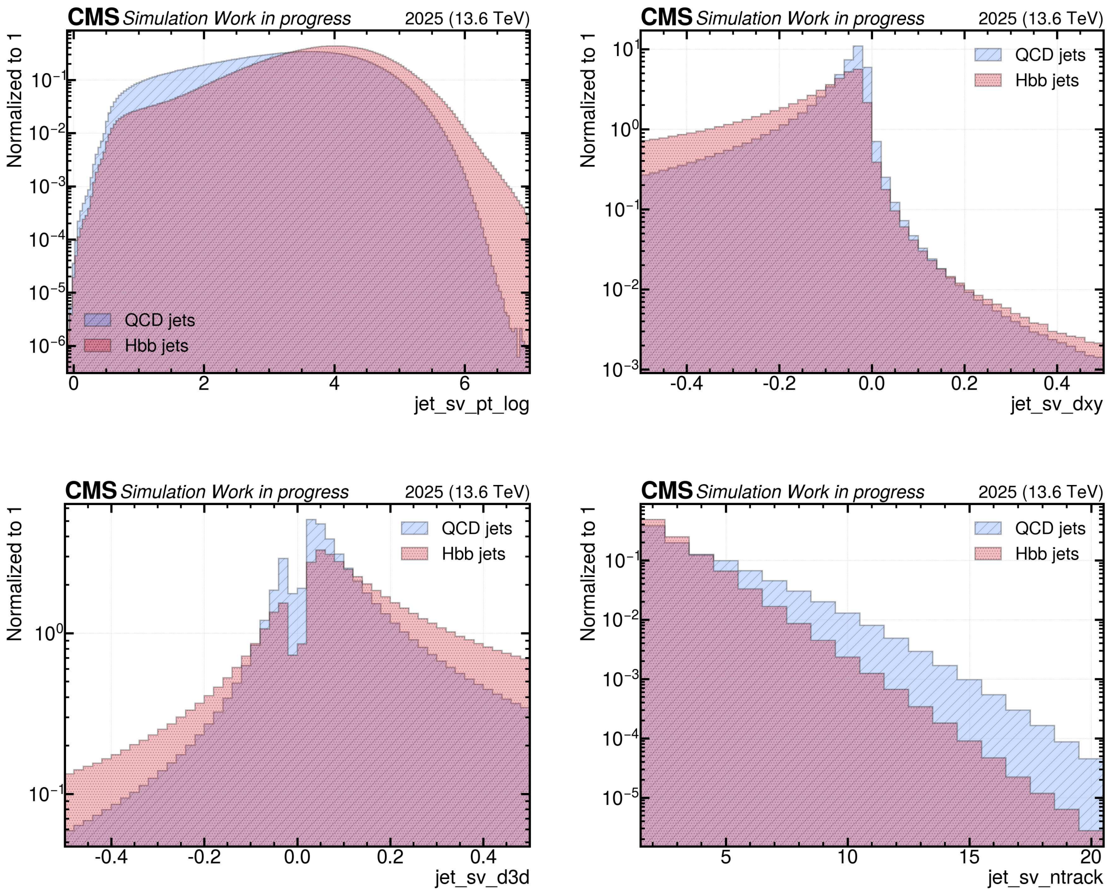

Figure C.8: Normalized to 1 unweighted jet_sv_pt_log (top-left), jet_sv_dxy (top-right), jet_sv_d3d (bottom-left) and jet_sv_d3dsig distributions (bottom-right) for XtoHH AK8 jets labeled as $H \to b\bar{b}$ (red histograms) and QCD AK8 jets (light blue histograms).

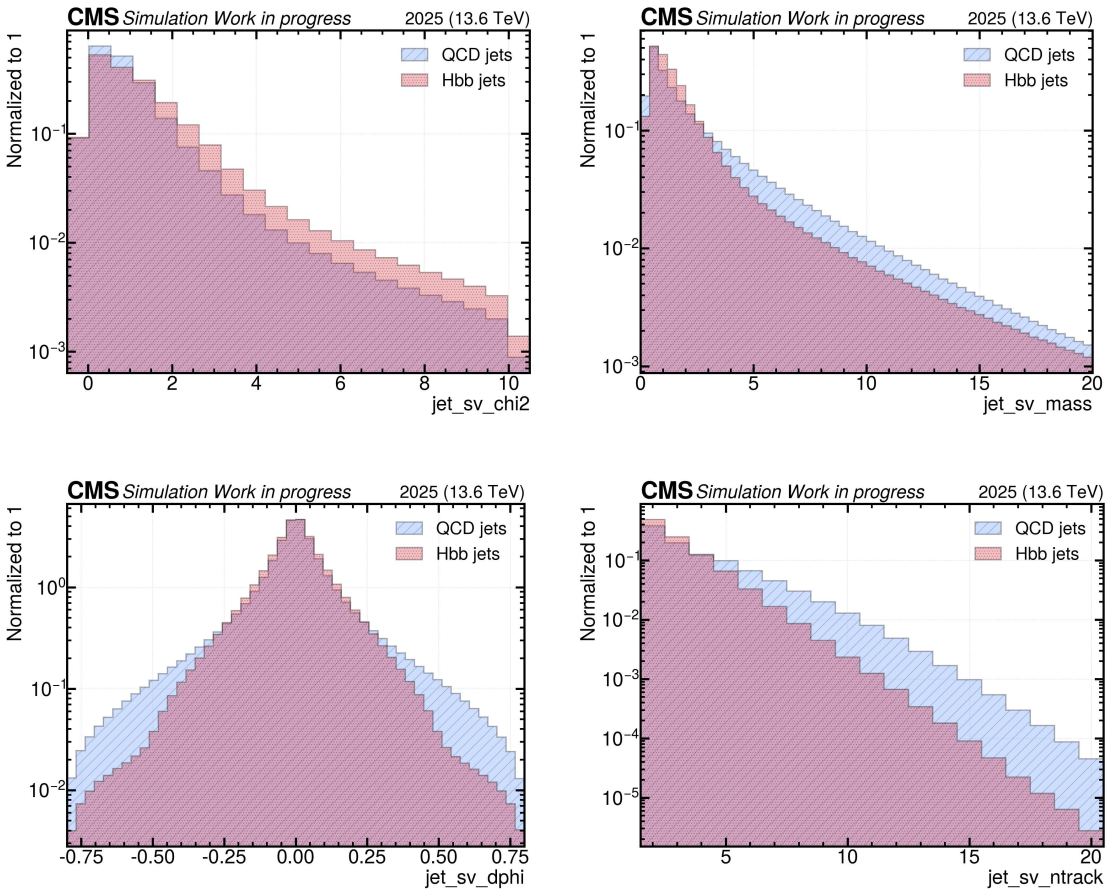

Figure C.9: Normalized to 1 unweighted jet__sv__chi2 (top-left), jet__sv__mass (top-right), jet__sv__dphi (bottom-left) and jet__sv__ntrack distributions (bottom-right) for XtoHH AK8 jets labeled as $H \to b\bar{b}$ (red histograms) and QCD AK8 jets (light blue histograms).

Bibliography

[1] O. Nachtmann. *Elementary particle physics: concepts and phenomena*. Springer, 1990.

[2] 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*, 716(1):1–29, sep 2012.

[3] CMS Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, 716(1):30–61, 2012.

[4] The CMS Collaboration. The CMS experiment at the CERN LHC. *Journal of Instrumentation*, 3(08):S08004, aug 2008.

[5] The ATLAS Collaboration. The ATLAS Experiment at the CERN Large Hadron Collider. *Journal of Instrumentation*, 3(08):S08003, aug 2008.

[6] Particle Data Group and S. Navas et al. Review of Particle Physics. *Phys. Rev. D*, 110(3):030001, 2024.

[7] The ATLAS Collaboration. Observation of a cross-section enhancement near the $t\bar{t}$ production threshold in $\sqrt{s} = 13$ TeV pp collisions with the ATLAS detector, ATLAS-CONF-2025-008, 2025. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2025-008.

[8] The CMS Collaboration. Observation of a pseudoscalar excess at the top quark pair production threshold, CMS-TOP-24-007. *Rep. Prog. Phys.*, 88:087801, 2025.

[9] John C. Collins, Davison E. Soper, and George Sterman. Heavy particle production in high-energy hadron collisions. *Nuclear Physics B*, 263(1):37–60, 1986.

[10] Gunnar S. Bali. Qcd forces and heavy quark bound states. *Physics Reports*, 343(1):1–136, 2001.

[11] Peter Arnold, Omar Elgedawy, and Shahin Iqbal. Are Parton Showers Inside a Quark-Gluon Plasma Strongly Coupled? A Theorist's Test, 2024.

[12] B.R. Webber. Fragmentation and Hadronization. Technical report, CERN, 2000.

[13] T. Kinoshita T. Aoyama, M. Hayakawa and M. Nio. Tenth-order electron anomalous magnetic moment: Contribution of diagrams without closed lepton loops. *Physical Review D*, 91(3), feb 2015.

[14] UA1 Collaboration. Experimental observation of isolated large transverse energy electrons with associated missing energy at s=540 GeV. *Physics Letters B*, 122(1):103–116, 1983.

[15] UA1 Collaboration. Experimental observation of lepton pairs of invariant mass around $95\,\mathrm{GeV}/c^2$ at the cern sps collider. *Physics Letters B*, 126(5):398–410, 1983.

[16] F. Englert and R. Brout. Broken symmetry and the mass of gauge vector mesons. *Phys. Rev. Lett.*, 13:321–323, Aug 1964.

[17] Peter W. Higgs. Broken symmetries and the masses of gauge bosons. *Phys. Rev. Lett.*, 13:508–509, Oct 1964.

[18] U Dore and D Orestano. Experimental results on neutrino oscillations. *Reports on Progress in Physics*, 71(10):106201, sep 2008.

[19] G. Dattoli and K. V. Zhukovsky. Neutrino mixing and the exponential form of the Pontecorvo-Maki-Nakagawa-Sakata matrix. *Eur. Phys. J. C*, 55:547–552, 2008.

[20] The ATLAS Collaboration. Combined Measurement of the Higgs Boson Mass from the $H \to \gamma\gamma$ and $H \to ZZ^* \to 4\ell$ decay channels with the ATLAS detector using $\sqrt{s}$= 7, $\sqrt{s}$=8 and $\sqrt{s}$=13 TeV pp collision data. *Physical Review Letters*, 131(25), December 2023.

[21] S. Myers and E. Picasso. The Large Electron-Positron Collider. *Scientific American*, 263(1):54–61, 1990.

[22] A. Barbiellini *et. al.* Precision electroweak measurements on the z resonance. *Physics Reports*, 427(5):257–454, 2006.

[23] Brian Moser. *Higgs Boson Physics*, pages 29–44. Springer Nature Switzerland, Cham, 2023.

[24] The CMS Collaboration. Measurements of the inclusive W and Z boson production cross sections and their ratios in proton-proton collisions at $\sqrt{s} = 13.6$TeV. *Journal of High Energy Physics*, 2026:47, 2026.

[25] D. de Florian and Others. Handbook of LHC Higgs cross sections: 4. Deciphering the nature of the Higgs sector, 2017.

[26] CERN. Cern – european organization for nuclear research. `https://home.cern/`.

[27] Lyndon Evans and Philip Bryant. LHC Machine. *Journal of Instrumentation*, 3(08):S08001, aug 2008.

[28] Cern's accelerator complex, 2022.

[29] The ALICE Collaboration. The ALICE experiment at the CERN Large Hadron Collider. *Journal of Instrumentation*, 3:S08002, 2008.

[30] The LHCb Collaboration. The LHCb detector at the LHC. *Journal of Instrumentation*, 3:S08005, 2008.

[31] The TOTEM Collaboration. The TOTEM experiment at the CERN Large Hadron Collider. *Journal of Instrumentation*, 3:S08007, 2008.

[32] The MoEDAL Collaboration. Technical Design Report of the MoEDAL Experiment, CERN-LHCC-2009-006. Technical report, CERN, 2009.

[33] The LHCf Collaboration. The LHCf detector at the CERN Large Hadron Collider. *Journal of Instrumentation*, 3:S08006, 2008.

[34] Jonathan L. Feng, Iftah Galon, Felix Kling, and Sebastian Trojanowski. ForwArd Search ExpeRiment at the LHC. *Phys. Rev. D*, 97:035001, Feb 2018.

[35] The CMS Collaboration. CMS Luminosity - Public Results.

[36] CMS Collaboration. Detector – CMS Experiment. `https://cms.cern/detector`, September 2025.

[37] The CMS Collaboration. Description and performance of track and primary-vertex reconstruction with the CMS tracker. *JINST*, 9:P10009. 80 p, Oct 2014.

[38] The Tracker Group of the CMS Collaboration. The CMS Phase-1 Pixel Detector Upgrade. Technical report, CERN, Geneva, 2020.

[39] The CMS Collaboration. Performance of Muon Reconstruction in the CMS High Level Trigger using pp Collision data at $\sqrt{s} = 13.6$ TeV in 2023, 2024.

[40] The CMS Collaboration. Description and performance of track and primary-vertex reconstruction with the cms tracker. *Journal of Instrumentation*, 9(10):P10009, oct 2014.

[41] The CMS Collaboration. *The CMS electromagnetic calorimeter project: Technical Design Report*. Technical design report. CMS. CERN, Geneva, 1997.

[42] The CMS Collaboration. Jet Energy Scale and Resolution Measurements Using Prompt Run3 Data Collected by CMS in the Last Months of 2022 at 13.6 TeV, CMS-DP-2023-045, 2023.

[43] P Adzicet and Others. Energy resolution of the barrel of the CMS Electromagnetic Calorimeter. *Journal of Instrumentation*, 2(04):P04004, apr 2007.

[44] The CMS Collaboration. *The CMS hadron calorimeter project: Technical Design Report*. CERN, 1997.

[45] The CMS Collaboration. Precise mapping of the magnetic field in the CMS barrel yoke using cosmic rays. *Journal of Instrumentation*, 5(03):T03021, mar 2010.

[46] The CMS Collaboration. *The CMS muon project: Technical Design Report*. Technical design report. CMS. CERN, Geneva, 1997.

[47] A Colaleo, A Safonov, A Sharma, and M Tytgat. CMS Technical Design Report for the Muon Endcap GEM Upgrade, 2015.

[48] The CMS Collaboration. Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13\,\mathrm{TeV}$. *JINST*, 13(06):P06015, 2018.

[49] The CMS Collaboration. Study of the effects of radiation on the CMS Drift Tubes Muon Detector for the HL-LHC. *Journal of Instrumentation*, 14(12):C12010, dec 2019.

[50] The CMS Collaboration. Performance of the CMS drift tube chambers with cosmic rays. *Journal of Instrumentation*, 5(03):T03015, mar 2010.

[51] Karol Bunkowski and Jan Krolikowski. Optimization, synchronization, calibration and diagnostic of the rpc pac muon trigger system for the cms detector, 01 2009.

[52] CMS Muon Collaboration. The upgrade of the CMS muon system for the High Luminosity LHC, 2024.

[53] C. Grandi. Cms distributed data analysis challenges. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 534(1):87–93, 2004.

[54] The CMS Collaboration. *CMS TriDAS project: Technical Design Report, Volume 1: The Trigger Systems*. Technical design report. CMS. CERN, 2000.

[55] 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. CERN, Geneva, 2002.

[56] The CMS Collaboration. Performance of the CMS Level-1 trigger in proton-proton collisions at $\sqrt{s} = 13$ TeV. *Journal of Instrumentation*, 15(10):P10017–P10017, oct 2020.

[57] The CMS Collaboration. Enriching the physics program of the cms experiment via data scouting and data parking. *Physics Reports*, 1115:678–772, 2025.

[58] The CMS Collaboration. *CMS computing: Technical Design Report*. Technical design report. CMS. CERN, Geneva, 2005. Submitted on 31 May 2005.

[59] The WLCG Collaboration. Worldwide LHC Computing Grid Tiers infrastructure.

[60] CMS Collaboration. Particle-flow reconstruction and global event description with the CMS detector. *Journal of Instrumentation*, 12(10):P10003–P10003, oct 2017.

[61] 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.

[62] F. Pantaleo. New Track Seeding Techniques for the CMS Experiment, 2017.

[63] P.C. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1):49–55, 1936.

[64] The CMS Collaboration. CMS tracking performance in 2024, CMS-DP-2025-026, 2025.

[65] The CMS Collaboration. Tracking performance using Tag and Probe with $Z \to \mu^+\mu^-$ in 2022 and 2023, CMS-DP-2024-054, 2024.

[66] Thomas Speer, Kirill Prokofiev, R Frühwirth, Wolfgang Waltenberger, and Pascal Vanlaer. Vertex Fitting in the CMS Tracker. Technical report, CERN, Geneva, 2006.

[67] K. Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE*, 86(11):2210–2239, 1998.

[68] R Frühwirth, Wolfgang Waltenberger, and Pascal Vanlaer. Adaptive Vertex Fitting. Technical report, CERN, Geneva, 2007.

[69] The CMS Collaboration. Identification of heavy-flavour jets with the cms detector in pp collisions at 13 tev. *Journal of Instrumentation*, 13(05):P05011, may 2018.

[70] CMS Collaboration. Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC. *Journal of Instrumentation*, 16(05):P05014, may 2021.

[71] CMS Collaboration. Reconstruction of signal amplitudes in the CMS electromagnetic calorimeter in the presence of overlapping proton-proton interactions. *Journal of Instrumentation*, 15(10):P10002–P10002, oct 2020.

[72] D. Valsecchi for the CMS Collaboration. Deep learning techniques for energy clustering in the cms ecal. *Journal of Physics: Conference Series*, 2438(1):012077, feb 2023.

[73] The CMS Collaboration. ECAL calibration performance in Run 3 with reprocessed data, CMS-DP-2024-022, 2024.

[74] The CMS Collaboration. Electron and photon reconstruction and identification performance at CMS in 2022 and 2023, CMS-DP-2024-052, 2024.

[75] W Adam, R Frühwirth, A Strandlie, and T Todorov. Reconstruction of electrons with the gaussian-sum filter in the CMS tracker at the LHC. *Journal of Physics G: Nuclear and Particle Physics*, 31(9):N9–N20, jul 2005.

[76] Muon ID and Isolation efficiencies with muons in proton-proton collisions at sqrt(s) = 13.6 TeV, CMS-DP-2024-067, 2024.

[77] M. Low D. Bertolini, P. Harris and N. Tran. Pileup per particle identification. *Journal of High Energy Physics*, 2014, set 2014.

[78] The CMS Collaboration. Pileup-per-particle identification: optimisation for Run 2 Legacy and beyond, CMS-DP-2021-001, 2021.

[79] Matteo Cacciari, Gavin P Salam, and Gregory Soyez. The anti-k_T jet clustering algorithm. *Journal of High Energy Physics*, 2008(04):063–063, apr 2008.

[80] Matteo Cacciari, Gavin P Salam, and Gregory Soyez. Fastjet user manual. *The European Physical Journal C*, 3 2012.

[81] A. J. Larkosk *et al.* Soft drop. *Journal of High Energy Physics*, 2014, may 2014.

[82] The CMS Collaboration. Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV. *Journal of Instrumentation*, 12(02):P02014, feb 2017.

[83] The CMS Collaboration. Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s}$=13 TeV using the CMS detector. *Journal of Instrumentation*, 14(07):P07004, jul 2019.

[84] The CMS Collaboration. DeepMET: Improving missing transverse momentum estimation with a deep neural network, CMS-PAS-JME-24-001. Technical report, CERN, Geneva, 2025.

[85] The CMS Collaboration. Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV. *Journal of Instrumentation*, 13(05):P05011, may 2018.

[86] Heavy Flavor Averaging Group Collaboration. Averages of b-hadron, c-hadron, and τ-lepton properties as of 2021. *Phys. Rev. D*, 107:052008, Mar 2023.

[87] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[88] G. Corso, H. Stark, S. Jegelka, and et al. Graph neural networks. *Nature Reviews Methods Primers*, 4:17, 2014.

[89] E. Bols *et al.* Jet flavour classification using DeepJet. *Journal of Instrumentation*, 15(12):P12012, dec 2020.

[90] Huilin Qu and Loukas Gouskos. Jet tagging via particle clouds. *Phys. Rev. D*, 101:056019, Mar 2020.

[91] Huilin Qu, Congqiao Li, and Sitian Qian. Particle transformer for jet tagging. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18281–18292. PMLR, 17–23 Jul 2022.

[92] Yue Wang *et al.* Dynamic Graph CNN for Learning on Point Clouds, 2019.

[93] Chase Shimmin *et al.* Decorrelated jet substructure tagging using adversarial neural networks. *Phys. Rev. D*, 96:074034, Oct 2017.

[94] The CMS Collaboration. Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques. *Journal of Instrumentation*, 15(06):P06005, jun 2020.

[95] The CMS Collaboration. Performance of the mass-decorrelated DeepDoubleX classifier for double-b and double-c large-radius jets with the CMS detector, CMS-DP-2022-041, 2022.

[96] The CMS Collaboration. Performance of heavy-flavour jet identification in Lorentz-boosted topologies in proton-proton collisions at $\sqrt{s} = 13$ TeV. *Journal of Instrumentation*, 20(11):P11006, nov 2025.

[97] The CMS Collaboration. Performance of boosted $b\bar{b}$ jet tagging at $\sqrt{s} = 13.6$ TeV with Run 3 CMS data, CMS-DP-2024-055, 2024.

[98] P. Grafström and W. Kozanecki. Luminosity determination at proton colliders. *Progress in Particle and Nuclear Physics*, 81:97–148, 2015.

[99] The CMS Collaboration. Luminosity measurement in proton-proton collisions at 13.6 TeV in 2022 at CMS. Technical report, CERN, Geneva, 2024.

[100] J. Alwall and Others. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *Journal of High Energy Physics*, 2014(7), jul 2014.

[101] T. Sjöstrand *et al.* An introduction to PYTHIA 8.2. *Computer Physics Communications*, 191:159–177, 2015.

[102] The CMS Collaboration. Extraction and validation of a new set of CMS pythia8 tunes from underlying-event measurements. *The European Physical Journal C*, 80(04), 01 2020.

[103] NNPDF Collaboration. Parton distributions from high-precision collider data. *The European Physical Journal C*, 77(663), 10 2017.

[104] J. Alwall *et al.* Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions. *The European Physical Journal C*, 53:473–500, feb 2008.

[105] J. Allison and Others. GEANT4 developments and applications. *IEEE Transactions on Nuclear Science*, 53(1):270–278, 2006.

[106] D. Krohn J. Thaler and L.T. Wang. Jet trimming. *Journal of High Energy Physics*, 2010:84, feb 2010.

[107] D. J. Hand. Statistical concepts: A second course, fourth edition by richard g. lomax, debbie l. hahs-vaughn. *International Statistical Review*, 80(3):491–491, 2012.

[108] The CMS collaboration. Determination of jet energy calibration and transverse momentum resolution in cms. *Journal of Instrumentation*, 6(11):P11002, nov 2011.

[109] The CMS Collaboration. The CMS Statistical Analysis and Combination Tool: Combine. *Computing and Software for Big Science*, 8:19, Nov 2024.

[110] S. S. Wilks. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *Annals Math. Statist.*, 9(1):60–62, 1938.

[111] The CMS Collaboration. The cms trigger system. *Journal of Instrumentation*, 12(01):P01020, jan 2017.

[112] H. Zhu et al. weaver-core: Core utilities for the weaver hep machine learning framework, 2023. GitHub repository.

[113] R. Brunand and F. Rademakers. ROOT – An Object Oriented Data Analysis Framework. *Nuclear Instruments and Methods in Physics Research Section A*, 389:81–86, 1997.

[114] J. Rodrigues and J. Alsing. Awkward array: Manipulating nested data in python. *Journal of Open Source Software*, 5(54):2301, 2020.

[115] Oren Ben-Kiki, Clark Evans, and Ingy d'ot Net. YAML Ain't Markup Language (YAML) Version 1.2.2. `https://yaml.org/spec/1.2.2/`, 2021.

[116] A. Paszke *et. al.* Pytorch: An imperative style, high-performance deep learning library, 2019.

[117] ONNX Contributors. Onnx: Open neural network exchange. `https://github.com/onnx/onnx`, 2017.

[118] P. F. Monni *et. al.* MiNNLO$_{\mathrm{PS}}$: optimizing $2 \rightarrow 1$ hadronic processes. *The European Physical Journal C*, 80:1075, 2020.

[119] T. Neumann. NLO Higgs+jet production at large transverse momenta including top quark mass effects. *Journal of Physics Communications*, 2(9):095017, sep 2018.

[120] ReCaS-Bari. Recas - bari computer center for science (recas-bari). `https://www.recas-bari.it/index.php/it/`.

[121] PyTorch Developers. torch.nn.crossentropyloss. `https://docs.pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html`, 2025.

[122] The CMS Collaboration. Search for a boosted Higgs boson decaying to bottom quark pairs in association with a hadronically decaying W or Z boson with the CMS detector using proton-proton collisions at $\sqrt{s} = 13$ TeV, CMS-PAS-HIG-24-017 , 2025.